PERSONNEL PSYCHOLOGY
2007, 60, 271–301

A REVIEW OF RECENT DEVELOPMENTS IN INTEGRITY TEST RESEARCH

CHRISTOPHER M. BERRY*, PAUL R. SACKETT, SHELLY WIEMANN
Department of Psychology
University of Minnesota

A sizable body of new literature on integrity tests has appeared since the last review of this literature by Sackett and Wanek (1996). Understanding of the constructs underlying integrity tests continues to grow, aided by new work at the item level. Validation work against a growing variety of criteria continues to be carried out. Work on documenting fakability and coachability continues, as do efforts to increase resistance to faking. New test types continue to be developed. Examination of subgroup differences continues, both at the test and facet level. Research addressing applicant reactions and cross-cultural issues is also reviewed.

This paper is the fifth in a series of reviews of the integrity testing literature (Sackett, Burris, & Callahan, 1989; Sackett & Decker, 1979; Sackett & Harris, 1984; Sackett & Wanek, 1996). As with earlier reviews, the goals are to give the reader a comprehensive but readable summary of developments in this area of research and practice, and to influence future research by identifying key gaps in the literature. This review includes various published and unpublished work between 1995 and 2006, with the goal of identifying relevant work since the last review. We conducted electronic searches of the literature, examined Society for Industrial and Organizational Psychology (SIOP) conference programs, and corresponded with integrity researchers.

We continue to use the term "integrity testing" to refer to the commercially marketed instruments that have been the focus of the previous reviews. This review includes the two categories of instruments that Sackett et al. (1989) labeled "overt" and "personality-oriented" tests. Overt integrity tests commonly consist of two sections. The first is a measure of theft attitudes and includes questions pertaining to beliefs about the frequency and extent of theft, punitiveness toward theft, ruminations about theft, perceived ease of theft, endorsement of common rationalizations for theft, and assessments of one's own honesty. The second involves requests

We thank Vanessa Tobares for her work in locating and summarizing articles.

Correspondence and requests for reprints should be addressed to Christopher M. Berry, Department of Psychology, Wayne State University, 5057 Woodward Ave., 7th Floor, Detroit, MI 48202; berry@wayne.edu.

*Christopher M. Berry is now at wayne State University.

for admissions of theft and other wrongdoing. Commonly used tests of this type include the Personnel Selection Inventory (PSI), the Reid Report, and the Stanton Survey.

Personality-oriented measures are closely linked to normal-range personality devices, such as the California Psychological Inventory. They are generally considerably broader in focus than overt tests and are not explicitly aimed at theft. They include items dealing with dependability, Conscientiousness, social conformity, thrill seeking, trouble with authority, and hostility. Commonly used tests of this sort are the Personnel Reaction Blank, the PDI Employment Inventory (PDI-EI), and Reliability Scale of the Hogan Personality Series.

Integrity testing began as an attempt to detect dishonesty in job applicants without having to use polygraph tests. Although no longer viewed as surrogates for polygraphs, the focus typically remains on the prediction of counterproductive work behaviors (CWB). Though integrity tests may be designed to predict different specific CWBs, they have generally been found to predict most CWBs approximately equally well. This is not surprising given recent advances in the conceptualization of the CWB domain demonstrating that individual CWBs are related to each other (e.g., engaging in one CWB increases the likelihood that other CWBs will also be engaged in). For instance, Bennett and Robinson (2000) conceptualized the CWB domain as consisting of two dimensions: interpersonal and organizational deviance, each of which contains various interrelated behaviors. Sackett and DeVore (2002) suggested a hierarchical model with a general CWB factor at the top, several group factors (such as interpersonal and organizational deviance) below the general factor, and specific CWB domains such as theft below these group factors. A recent meta-analysis by Berry, Ones, and Sackett (2007) substantiated the hierarchical nature of CWB. Thus, it should not be surprising that integrity tests predict most CWBs.

Although integrity tests are generally designed to predict CWB, they have also been found to predict job performance (Ones, Viswesvaran, & Schmidt, 1993). In fact, Schmidt and Hunter (1998) identified integrity tests as the personnel selection method with the greatest incremental validity in predicting job performance over cognitive ability. This relationship between integrity and performance should not be surprising, given that CWBs are related to other performance behaviors such as organizational citizenship behaviors (Dalal, 2005; Sackett, Berry, Wiemann, & Laczo, 2006), and that supervisors' overall performance ratings reflect judgments regarding CWB (Rotundo & Sackett, 2002).

As background, we refer the reader to the last review in this series (Sackett & Wanek, 1996). Table 1 in the present review also offers brief thumbnail sketches of the major integrity testing research findings as of the Sackett and Wanek (1996) review. We have organized the present review of

TABLE 1

Brief Overview of Major Research Findings as of the Sackett and Wanek (1996) Review

Topic	Major research findings as of Sackett and Wanek (1996)
Criterion-related validity	Ones et al. (1993) meta-analyzed 665 validity studies. *Prediction of CWBs other than theft*: Overt tests predict .39 (.27 uncorrected), personality based (PB) predict .29 (.20 uncorrected). Credibility interval is lower for overt tests, so no clear basis for preferring one type of test over the other. *Prediction of theft*: Overt and PB tests predict theft .13 (.09 uncorrected). This estimate is artificially reduced because of the low base rate of theft. When corrected for the low base rate, validity is .33. *Prediction of job performance*: Overt and PB tests predict job performance .41 (.23 uncorrected).
Relationships among integrity tests	Integrity tests cannot be viewed as interchangeable, and thus meta-analytic findings do not generalize to anything with an "integrity test" label. Ones et al. (1993) found that the mean correlation (1) among overt tests is .45 (.32 uncorrected), (2) among PB tests is .70 (.43 uncorrected), and (3) between overt and PB tests is .39 (.25 uncorrected).
Relationships with personality variables	Integrity tests correlate substantially with Conscientiousness, Agreeableness, and Emotional Stability. Strongest correlation is with Conscientiousness. Partialling Conscientiousness out of integrity has only a small effect on integrity test validity, but partialling integrity out of Conscientiousness reduces criterion-related validity of Conscientiousness to near zero (Murphy & Lee, 1994; Ones, 1993).
Relationships with cognitive ability	Integrity tests are unrelated to cognitive ability (Ones et al., 1993).
Faking and coachability	Individuals can fake good when instructed to do so. One coaching study showed large effects on an overt test but not on a PB test (Alliger, Lilienfeld, & Mitchell, 1996).
Subgroup differences	Ones, Viswesvaran, and Schmidt (1996) meta-analysis found (1) negligible race differences, (2) women score between .11 and .27 standard score units higher, depending on the test.
Applicant reactions	Integrity tests generally do not produce strong negative reactions. In studies looking at reactions to a wide range of selection devices, integrity tests are in the middle of the pack relative to other devices. Findings are in conflict as to whether overt or personality-oriented tests produce more favorable reactions. Contextual factors (e.g., the explanation offered for the reason the firm is using the test) affect reactions to the tests.

new developments since Sackett and Wanek (1996) around seven themes. These will be listed briefly here; each item listed will then be the subject of a separate section of the paper: (a) What constructs do integrity tests measure? (b) Are there new insights into criterion-related validity? (c) Are there new insights into the fakability and coachability of integrity tests? (d) What new types of tests have emerged? (e) Are there new legal challenges to integrity test use? (f) Are there new insights into applicant reactions and their consequences? (g) What is the status of integrity test use outside the United States?

Construct Understanding

Links to Personality Variables

The major development in understanding the constructs underlying integrity tests in the period leading up to the Sackett and Wanek (1996) review was the finding that integrity tests were consistently correlated with three of the Big Five dimensions: Conscientiousness, Agreeableness, and Emotional Stability. As the focus on understanding the construct(s) underlying test scores increases among personnel selection researchers, a distinction is emerging between two types of personality traits: basic traits and compound traits (Hough & Schneider, 1996). According to these authors, basic traits are identified when the focus is on conceptual coherence, internal consistency, and temporal stability. We would characterize this as a "predictor-focused" approach. In contrast, there is a well-established tradition in selection research of focusing on a criterion of interest (e.g., CWB, customer service, sales effectiveness). In such a "criterion-focused" approach, items are retained on the basis of predictive relationships with the criterion, and the result may be a measure with low internal consistency, tapping multiple basic traits that may not all covary. Measures developed in such a fashion are labeled "compound traits"; Hough and Schneider identify integrity tests as an example of a measure of a compound trait. The key idea is that an empirically chosen combination of facets of basic traits (based on multiple studies, and thus not relying on chance features of single samples) designed to be maximally predictive of specific criteria in specific contexts should result in higher criterion-related validity than that of basic traits. The finding that integrity tests predict counterproductive behavior criteria better than Big Five measures, or composites of Big Five measures illustrates this argument.

So, integrity is a compound trait linked to Conscientiousness, Agreeableness, and Emotional Stability, but these three personality variables do not account for all of the variance in integrity and do not account for as much variance in CWB or job performance as does integrity (e.g., Murphy

& Lee, 1994; Ones, 1993). This leads to the question: "What is left in integrity other than these three Big Five traits?" Sackett and Wanek (1996) postulated that integrity tests have a greater emphasis on self-control than Big Five measures, though empirical research has yet to directly address this possibility. Becker (1998, 2005) has also offered suggestions as to what the construct of integrity may be comprised of, though Becker's theoretical position may be seen more as expanding the current definition of "integrity," rather than explaining what is left in the current construct of integrity beyond the Big Five.

Lee, Ashton, and de Vries (2005) and Marcus, Lee, and Ashton (2007) suggest that integrity tests may reflect a sixth personality dimension they have entitled "Honesty-Humility (H-H)" that is not adequately captured by the Big Five. Lee et al. (2005) define H-H "by such content as sincerity, fairness, lack of conceit, and lack of greed" (p. 182). In both Lee et al. (2005) and Marcus et al. (2007), the H-H scales had corrected correlations between .50 and .66 with integrity tests. Lee et al. (2005) demonstrated that multiple correlations between a six-factor model of personality including H-H (termed HEXACO) and workplace delinquency were .10 to .16 higher than the same multiple correlations using just the Big Five. Lee et al. also found that the HEXACO model was more correlated with the Employee Integrity Index (EII; an overt test) than was the Big Five (multiple correlations of .61 vs. .43). Further, Marcus et al. (2007) demonstrated that H-H accounted for more incremental variance over personality-based than overt integrity tests in predicting self-report CWB, implying that H-H may be reflected more in overt than personality-based integrity tests. Thus, there is some support for the idea that H-H may partially explain variance in integrity (especially overt integrity tests) beyond the Big Five.

Item-Level Analysis Across Tests

Although factor analyses of individual tests have been reported in earlier reviews, item-level analysis that combines data across multiple tests is a new development. Item-level analysis across multiple tests allows researchers to determine what factors are common and not common to the individual integrity tests contributing items. The first such study was reported by Hogan and Brinkmeyer (1997), who examined responses to the Hogan Reliability Scale (personality-based) and Reid Report (overt). All items from the Reliability Scale loaded on one factor, whereas the items on the Reid Report loaded on three other factors (punitive attitudes, admissions, and drug use). A second-level confirmatory factor analysis was conducted on the four factor scores; all loaded on a single factor, which the authors labeled Conscientiousness. This finding of a hierarchical structure

at the item level nicely complements the research by Ones (1993), who drew similar conclusions at the test-scale score level.

Wanek, Sackett, and Ones (2003) investigated the interrelationships between overt and personality-based integrity tests at the item level among a larger set of tests. A judgmental sort of 798 items from three overt tests (PSI, Reid Report, and Stanton Survey) and four personality-based tests (Employee Reliability Index, Personnel Reaction Blank, PDI-EI, and Inwald Personality Inventory) resulted in 23 distinct composites. Principal components analysis of these 23 indicated four components: antisocial behavior (e.g., theft admissions, association with delinquents), socialization (e.g., achievement orientation, locus of control [LoC]), positive outlook (e.g., viewing people as basically good and the world as basically safe), and orderliness/diligence. Although these four components underlie each of the seven integrity tests, individual tests differed in the strength of relationship with the four components, with whether tests were personality-based versus overt accounting for some of these differences.

Wanek et al. (2003) also computed correlations between the four integrity test components and Big Five scales. Results suggested that Conscientiousness and Emotional Stability cut across all four of the principal components and that Agreeableness correlated with the first three components, but less so with orderliness/diligence.

Therefore, combining the work of Wanek et al. (2003) and Hogan and Brinkmeyer (1997), it becomes apparent that integrity tests are multifaceted and that the construct they are measuring may be hierarchical in nature (i.e., an overall Conscientiousness factor, and Wanek et al.'s four components and 23 thematic composites as group and lower-order factors, respectively). Further, it is also apparent that the construct underlying integrity tests reflects a complex mix of all Big Five factors, with the strongest links being to Conscientiousness, Emotional Stability, and Agreeableness. What the item-level research has yet to directly address is what is left in integrity beyond the Big Five factors (e.g., H-H, self-control, etc.), making this an avenue for future research. In addition, Wanek et al. (2003) suggested another logical next step would be an examination of predictor-criterion evidence for the integrity composites identified. A study by Van Iddekinge, Taylor, and Eidson (2005) represents an early attempt to address one of these logical next steps. Van Iddekinge et al. reported predictor-criterion evidence for eight integrity facets they identified via judgmental sort of the PSI customer service scale (PSI-CS) items. Van Iddekinge et al.'s eight facets each map onto a subset of Wanek et al.'s (2003) 23 thematic composites. The eight integrity facets correlated between $-.16$ and $+.18$ with overall performance ratings, demonstrating heterogeneity.

Relationships With Cognitive Ability

A strong conclusion from earlier reviews was that the correlation between cognitive ability and integrity tests is essentially zero (Ones et al., 1993). This conclusion, though, has been based on overall scores on integrity tests. Given the developments outlined above, examination at the facet level is useful. Duehr, Sackett, and Ones (2003) investigated the relationships between cognitive ability and the 23 integrity facets identified by Wanek et al. (2003). Several personality-oriented integrity facets (e.g., Emotional Stability [$r = .16$], Extraversion [$r = .37$], LoC [$r = .30$], achievement [$r = .19$]) were positively correlated with cognitive ability, whereas honesty-oriented integrity facets (e.g., honesty attitudes [$r = -.33$], lack of theft thoughts/temptation [$r = -.22$]) were negatively related to cognitive ability. Thus, the near-zero correlation reported using overall integrity test scores is the result of combining facets with positive and facets with negative correlations with cognitive ability. Therefore, it would appear possible to produce more or less cognitively loaded tests by emphasizing different facets in constructing an overall scale.

Links to Situational Variables

Research on situational correlates of integrity tests has been sparse. One such study is Mumford, Connelly, Helton, Strange, and Osburn (2001), which related individual and situational variables from a biodata inventory to scores on the Reid Report and the PSI in a large undergraduate sample. A coherent pattern of findings emerged: For example, the strongest situational correlate of scores on both tests was exposure to a negative peer group. Mumford et al. argue that the fact that there are situational correlates of integrity test scores suggests that changing the situation an individual is in may result in a change in integrity test scores, though there are other plausible interpretations of such correlations (e.g., integrity causes association with negative peer groups). A study by Ryan, Schmit, Daum, Brutus, McCormick, and Brodke (1997) demonstrated an interaction between integrity and perceptions of the salience of situational influences. Students with lower integrity test scores viewed the situation as having less influence on their behavior than those with higher scores. Taken together, these two studies demonstrate that like virtually every individual difference construct in psychology, it is likely that both situational and dispositional influences play a part. Therefore, research taking an interactionist perspective may increase our understanding of the construct validity of integrity tests.

Conclusions

Our understanding of the constructs underlying integrity tests has increased considerably. New research suggests that integrity test scores may be affected by situational factors. Regarding dispositional factors, item-level research has reinforced the test-level conclusions that integrity reflects in part a complex mix of the Big Five. Further, a picture is emerging of integrity tests reflecting a hierarchical construct. Much more nebulous is what is left in integrity beyond the Big Five. New item-level research suggests that part of this answer may be "cognitive ability," depending on the specific facets measured in individual integrity tests. Other research suggests H-H might be a partial answer. Promising concepts outside of these personality taxonomies such as attitudes or situational variables may also exist. Answering the question "what is left in integrity beyond the Big Five?" is surely one of the most important unanswered questions regarding our understanding of the constructs integrity tests measure.

Validity

Criterion-Related Studies in Operational Settings

A number of new primary validity studies have been reported since the Sackett and Wanek (1996) review (Borofsky, 2000; Boye & Wasserman, 1996; Hein, Kramer, & Van Hein, 2003; Lanyon & Goodstein, 2004; Mastrangelo & Jolton, 2001; Nicol & Paunonen, 2001; Rosse, Miller, & Ringer, 1996). Findings were generally supportive, though this is not surprising given the wealth of predictive validity evidence demonstrated by cumulative meta-analytic investigations (Ones et al., 1993).

Relations With Counterproductive Behavior in Controlled Settings

A major methodological difficulty in examining relationships between integrity tests and CWBs is that many of the behaviors of interest are not readily observable. Studies using detected theft as a criterion, for example, are difficult to interpret, as it is unclear what proportion of theft is detected and whether detected theft is a random sample of all theft. In response to such difficulties, and illustrating the classic tradeoff between internal and external validity, a growing number of researchers are turning to a research strategy wherein integrity tests are administered to individuals who are put into controlled research settings where they are presented with opportunities to engage in behaviors viewed as counterproductive by the researchers, and which are observable or indirectly detectable by the researcher without the participant's awareness. Although the behaviors

studied are not actual on-the-job behaviors, the research strategy has the advantage that the behaviors of interest can be reliably detected. Thus, this emerges as a useful adjunct to other strategies for studying integrity test validity.

In one such study, Mikulay and Goffin (1998) examined relationships between the PDI-EI and a variety of measures in a laboratory setting. Participants were observed through a one-way mirror as they attempted to earn a cash prize based on performance in solving a jigsaw puzzle with a fixed time limit while looking in a mirror rather than directly at the puzzle. A composite of time spent looking directly at the puzzle and extra time spent on the task beyond the time limit served as a measure of "rule breaking"; the difference between self-reported number of pieces placed and actual pieces placed served as a measure of "fraud"; and the number of pieces of candy removed from a dish served as a measure of "pilferage." PDI-EI scores were related to rule breaking ($r = .40$) and pilferage ($r = .36$), but not to fraud ($r = .07$).

In a similar study, Nicol and Paunonen (2002) examined the relationship between two overt tests (a measure developed for the study and the Phase II profile) and a variety of measures, including the puzzle task used in the study above. The puzzle task was combined with measures of whether participants added or changed answers when scoring intelligence or psychomotor tests they had taken to form a measure of "cheating"; and a composite of three behaviors was labeled "stealing" (e.g., taking coffee without the requested payment, taking change from the coffee payment bowl). Both tests were correlated with stealing (r s of $-.31$ and $-.32$); the new measure was also correlated with cheating ($r = -.25$).

Not all studies using controlled settings have had as positive results. For instance, Horn, Nelson, and Brannick (2004) investigated the relationship between PSI scores and an unobtrusive measure of claiming credit for more research participation time than actually spent in a sample of 86 undergraduates. Claiming extra credit was uncorrelated ($r = -.04$) with PSI scores. As another example, Hollwitz (1998) administered integrity measures to 154 participants in a controlled setting. Each participant was left alone at a table to complete the measures, and a folder labeled "exam answer key" was also left on the table. Hollwitz found no relationship ($r = -.08$) between scores on the EII and whether the participant opened the folder.

Use of controlled settings to examine integrity–criterion relationships is growing. One issue with this strategy is the use of single-act criteria, which are notoriously unreliable. Some studies offer multiple opportunities for misbehavior and create composites as a route to more reliable measures. When a single-act criterion is used, a null finding is hard to interpret, as a variety of features, from low reliability to low base rate

may affect the findings. Another issue is that it is unclear how much these criterion measures really reflect CWBs (e.g., is taking candy from a dish similar to on-the-job theft?) or whether they reflect counterproductivity at all (e.g., when a bowl of candy is left on a table is there not an implicit invitation to take a piece?). Nonetheless, findings from the set of studies using this strategy add to the body of support for the relationship between integrity tests and a wide range of counterproductive behaviors.

Relations with Absenteeism

Ones, Viswesvaran, and Schmidt (2003) reported a meta-analysis of relationships between integrity tests and non-self-reported voluntary absenteeism. Based on 13 studies of personality-based tests ($N = 4{,}922$), and 9 studies of overt tests ($N = 8{,}508$), they reported uncorrected means of .23 and .06 (.33 and .09 corrected for criterion unreliability and range restriction) for personality-based and overt tests, respectively. Thus, the data to date indicate a considerable difference in the predictive validity of the two types of tests, with personality-based tests more useful in the prediction of voluntary absenteeism. The reasons for this disparity are not particularly clear, though, and the number of studies contributing to the Ones et al. meta-analysis was relatively small. Thus, strong conclusions are tempered at this time.

Relationships With Peer and Interviewer Reports

Caron (2003) compared test scores obtained via traditional self-report, via a friend describing the target person, and via interviewer ratings of integrity. Caron found a positive correlation between self- and peer-reported integrity ($r = .46$) and self-ratings and interview ratings of integrity ($r = .28$). If future research demonstrates that there is predictive and conceptual value in using peer or interview ratings of integrity, it may prove a useful supplement to reliance on self-reports or supervisor ratings of integrity, each of which have their own conceptual limitations.

Conclusions

The range of criteria for which relationships with integrity tests has been found continues to expand. New laboratory research is examining deviance criteria that are both observable and verifiable, though the actual relationships between many of these new criteria and CWB is questionable and requires further research. If practitioners are interested in reducing voluntary absenteeism, personality-based tests may be preferable to overt tests, though the exact reason for this is unclear and additional research

would be useful. In addition, initial evidence suggests that peer reports may serve as a useful supplement to more traditional sources of information regarding integrity. Finally, though not mentioned above, there is preliminary evidence suggesting a relationship between integrity tests and academic cheating (Lucas & Friedrich, 2005), though this research has relied heavily on self-report. In all, the criterion-related validity evidence for integrity tests remains strong and positive.

Faking and Coaching

Faking

There has been a considerable amount of research on whether personality-oriented versus overt tests are more fakable, though we do not believe the conflict has yet been resolved. Alliger and Dwight (2000) reported a meta-analytic comparison of the two types of tests. Comparing "respond as an applicant" and "beat the test" conditions, they report a mean effect size of .93 *SD*s for overt tests and .38 *SD*s for personality-based tests. At first glance, this would appear to offer a clear answer as to which type of test was more resistant to faking. However, we believe such a conclusion is premature.

One important issue is the various instructional sets used in faking research. Three instructional sets are generally used in faking research: (a) respond honestly, (b) respond as an applicant, and (c) fake good to beat the test (e.g., Ryan & Sackett, 1987). Comparing results between these three instructional sets, the honest versus respond as an applicant comparison is the one we would characterize as attempting to estimate the effects of faking in an operational environment (e.g., a "will do" estimate of the typical amount of faking); the honest versus fake good comparison is one we would characterize as a "can do" estimate of the maximum amount of faking. It is not clear what is estimated by the applicant versus fake good comparison used by Alliger and Dwight (2000).

In terms of the more useful "can do" comparisons, a subsequent study by Hurtz and Alliger (2002) compares an overt test (the EII) and two personality-based tests (the Personnel Reaction Blank and the PDI-EI) under respond honestly versus faking conditions, and produces much more similar findings for the two types of tests ($d = .78$ for overt and .68 for personality based). Thus, the findings may vary as a result of the instructional sets being compared.

A second issue in interpreting the Alliger and Dwight meta-analysis is that there are apparent errors in the computation of effect size values. For example, they set aside a study by Holden (1995) due to an extreme value of 2.57; we obtain a value of .98 from Holden. They obtained a value of

2.89 from Ryan and Sackett (1987); we obtain a value of .89 for the theft attitudes scale. The only way we can obtain a value that matches theirs is to sum the *d*-values for the attitude scale, the admissions scale, and a social desirability scale included in the study. Clearly, the social desirability scale is not relevant to estimating the fakability of the integrity scales. Thus, we urge caution in drawing conclusions from Alliger and Dwight (2000).

Another issue in interpreting comparisons of overt and personality-based tests involves the fact that overt tests commonly include attitudes and admissions sections, which are often scored separately. Faking studies differ in terms of how they treat these separate sections. Some studies combine the attitudes and admissions sections in producing an estimate of fakability (e.g., Brown & Cothern, 2002; Hurtz & Alliger, 2002), whereas others use only the attitudes section (e.g., McFarland & Ryan, 2000; Ryan & Sackett, 1987). We believe the most useful strategy would be to report findings separately for each section where possible and to note instances where such separation is not possible. Until this is done, the issue of the relative resistance to faking efforts of different types of tests remains unclear.

Regardless of the relative fakability of types or sections of integrity tests, when instructed to fake on an integrity test, it appears that respondents are able to do so. A generally unanswered question is whether job applicants actually do fake on integrity tests. Van Iddekinge, Raymark, Eidson, and Putka (2003) examined this issue by comparing mean scores on the PSI-CS of applicants for and incumbents of customer service manager positions. Applicants only scored .09 standard score units higher than incumbents, implying that the integrity test was resistant to faking, though Van Iddekinge et al. suggested other possible explanations. Further research in operational settings is definitely needed.

If applicants can or do fake, the next obvious question is what can be done about it? One possibility is the use of response latency to identify faking on computerized integrity tests. Holden (1995) administered 81 delinquency-related items drawn from the Hogan Reliability Scale and the Inwald Personality Inventory to students responding under honest versus fake good conditions, and found significant differences in response latencies. Against a 50% chance rate, 61% of participants could be correctly classified on the basis of response latency as to whether they were in the honest or the fake good condition. In a second sample of unemployed job seekers, the correct classification rate rose to 72%. Dwight and Alliger (1997a) conducted a similar study with students, substituting an overt integrity test (the EII), and adding a coaching condition. Against a 33% chance rate, they found that 59% of participants could be correctly classified on the basis of response latencies as to whether they were in the honest, fake good, or coached condition. Finally, though results were

mixed, Leonard (1996) found some support for the use of a response latency measure of faking in a within-subjects study. Thus, response latency appears to be an avenue meriting further investigation.

Another possibility under investigation for controlling faking is the use of forced-choice measures. Jackson, Wroblewski, and Ashton (2000) explored whether recasting an existing integrity measure into a forced-choice format would reduce fakability. Undergraduates in one sample took the test in its original format under "respond honestly" and then "respond as a job applicant" conditions. Undergraduates in a second sample took the same integrity test recast into forced-choice format under the same two response conditions. In the original format sample, scores in the applicant condition were .95 *SD*s higher than scores in the honest condition, and the correlation between the integrity scale and a self-report CWB criterion dropped from .48 in the honest condition to .18 in the applicant condition. In the forced-choice sample, the mean difference between response conditions was only .32 *SD*s, and the correlations with the CWB criterion were .41 and .36 in the honest and applicant conditions, respectively. Thus, response conditions did not affect correlations with the criterion in the forced-choice format, although the effect was substantial in the original format.

Jackson et al. acknowledge that the data come from a simulated applicant setting, and thus caution is needed. This caution is, we believe, an important one. We point to the U.S. Army's recent implementation of a forced-choice personality measure as an example of obtaining very different findings regarding the resistance to faking in operational versus research settings. A composite of multiple personality dimensions on the Assessment of Individual Motivation (AIM) was used. In research settings, it appeared resistant to faking; Young, McCloy, Waters, and White (2004) report a mean difference of .15 *SD*s between standard instruction and fake good conditions in a large sample of recruits. However, when the measure was put into operational use, mean scores rose by .85 *SD*s (Putka & McCloy, 2004) relative to research conditions. In addition, correlations with attrition at 3 months dropped from $-.12$ to $-.01$. Thus, although we find the Jackson et al. findings very interesting, it is clear that investigation under operational conditions is warranted before drawing strong conclusions about the prospects for reducing fakability.

A variety of additional issues related to faking have been addressed. First, the role of cognitive ability in faking has been investigated. In a within-subject study, Brown and Cothern (2002) found a significant correlation ($r = .22$) between faking success on the attitude items of the Abbreviated Reid Report and cognitive ability, but no relationship ($r = .02$) for the admissions items. In a between-subjects study, Alliger et al. (1996) found larger correlations between a cognitive ability measure and

both an overt test (EII) and a personality-based test (PRB) in fake good conditions (correlations ranging between .16 and .36) than in respond as an applicant conditions (correlations of .17 and .20).

Second, Ones and Viswesvaran (1998b) reported a value of .06 as the meta-analytic mean estimate of the correlation between social desirability measures and integrity test scores. Third, Alliger and Dwight (2001) found a negative correlation between item fakability and rated item invasiveness: Items rated as more invasive are less fakable.

Coaching

Hurtz and Alliger (2002) conducted a replication of an earlier study by Alliger et al. (1996) examining the coachability of overt and personality-based integrity tests. Participants completed an overt test (EII) and two personality-based tests (PRB and PDI-EI) under an honest or one of two coaching conditions. One group received coaching oriented toward improving scores on an overt test, the other received coaching oriented toward a personality-oriented test. Faking conditions were also included in the study to permit a determination of whether coaching produced an incremental effect above that of faking. All interventions increased scores over a respond honestly condition. However, neither of the coaching interventions produced an increment more than .10 *SD*s over faking in any integrity score. Thus, coaching effects are minimal for these particular coaching interventions. Although the study is an effective examination of the efficacy of available advice for how to beat the tests, it is unclear whether more effective coaching interventions could be designed.

Conclusions

It is clear that respondents' integrity test scores can be increased via either faking or coaching (though preliminary evidence suggests existing coaching interventions are no more effective than simply asking a respondent to fake). However, a number of more nuanced issues regarding faking and coaching are being addressed or need addressing. For instance, though respondents *can* fake, there is still not definitive evidence that applicants *do* fake. Thus, more research such as that of Van Iddekinge et al. (2005) examining faking in applicant samples is needed. In addition, though integrity tests in general seem fakable, research is beginning to address whether certain types of tests or test items are more fakable than others. Although there is a meta-analysis focused on the relative fakability of overt versus personality-based tests, we view this issue as unresolved and deserving of future research that pays closer attention to the types of instructional sets given to respondents. Regarding fakability of different items, there

is preliminary evidence that more invasive items are less fakable. This is interesting and we encourage more research related to the fakability of different types of items or sections of tests. In addition, though mean score differences are one way to examine faking at a group level, more nuanced means, such as response latency, for detecting faking at the individual level are being investigated. Much like social desirability scales, the construct validity of response latencies as measures of faking is questionable and deserves more research, as do most areas dealing with the fakability of integrity tests.

New Types of Tests

Conditional Reasoning

A number of new types of tests have been designed as alternatives to current integrity tests. The most systematic program of research into new approaches is the work of Lawrence James and colleagues (2005) using an approach they label "conditional reasoning." James' overall theoretical approach is based on the notion that people use various justification mechanisms to explain their behavior and that people with varying dispositional tendencies will employ differing justification mechanisms. The basic paradigm is to present what appear to be logical reasoning problems, in which respondents are asked to select the response that follows most logically from an initial statement. In fact, the alternatives reflect various justification mechanisms that James posits as typically selected by individuals with a given personality characteristic.

For instance, an illustrative conditional reasoning item describes the increase in the quality of American cars over the last 15 years, following a decline in market share to more reliable foreign cars. Respondents are asked to select the most likely explanation for this. Consider two possible responses: "15 years ago American carmakers knew less about building reliable cars than their foreign counterparts" and "prior to the introduction of high-quality foreign cars, American car makers purposely built cars to wear out so they could make a lot of money selling replacement parts." The first is a nonhostile response, the second a hostile one. Choosing the second would contribute to a high score on an aggression scale and to a prediction that the individual is more likely to engage in CWB. James has developed a set of six justification mechanisms for aggression and has written conditional reasoning items with responses reflecting these mechanisms.

A number of validity studies have been conducted. The measure itself has been in flux. Later studies converged on a 22-item scale, now referred to as "CRT-A." As evidence of criterion-related validity, James

et al. (2005) included a table summarizing 11 validity studies. Each of the studies produced validity estimates ranging from .32 to .64, with an average uncorrected validity estimate of $r = .44$. A meta-analysis by Berry, Sackett, and Tobares (2007) located a larger set of studies, with a total sample size roughly twice that of James et al. (2005). Excluding studies with low rates of CWB, Berry et al. found that conditional reasoning tests of aggression had mean uncorrected validities of .25 and .14 for the prediction of CWB and job performance, respectively. Thus, the additional studies located by Berry et al. produce lower validity estimates than the earlier James et al. estimate, although the mean validity estimate for prediction of CWB is Still relatively comparable to those reported by Ones et al. (1993) for traditional integrity tests (.27 for overt tests; .20 for personality-oriented tests).

In addition, there is a program of research on fakability of the CRT-A. We direct the interested reader to LeBreton, Barksdale, Robin, and James (2007). Also of interest is LeBreton's (2002) variant on the conditional reasoning approach called the "Differential Framing Test (DFT)," in which respondents are presented with what appears to be a synonyms test. For example, two options for the stimulus word "critique" are "criticize" (an aggressive response) and "evaluate" (a nonaggressive response). LeBreton cross-validated empirical keys to predict conduct violations in an academic setting, finding that cross-validities were in the .30–.50 range in two samples. Internal consistency and test–retest estimates were generally acceptable and correlations with the CRT-A were low. In all, LeBreton's (2002) DFT shows promise. We do caution, though, that early validity studies for the CRT-A also suggested criterion-related validities similar to those exhibited thus far by the DFT, but later validity studies tended to find much lower validity for the CRT-A. Thus, although promising, more validity evidence is needed for the DFT.

New Test Formats

Although the work of James and Becker seeks to measure "integrity" from new theoretical perspectives, other work seeks to create prototypical integrity tests in new formats such as biodata, interviews, voice-response, and forced-choice response options. Beginning with biodata, Solomonson (2000) developed a set of construct-oriented biodata scales as an alternative to integrity tests. In a large undergraduate sample, Solomonson reported correlations of .71 and .50 with the EII (overt) and Personnel Reaction Blank (personality oriented), respectively. Solomonson also reported moderate correlations (.34–.48) with measures of Conscientiousness, Agreeableness, and Emotional Stability. Manley, Dunn, Beech, Benavidez, and Mobbs (2006) developed two biodata scales: one

designed to measure Conscientiousness and one designed to measure LoC. When administered to an undergraduate sample along with established measures of Conscientiousness and LoC, both biodata scales demonstrated adequate convergent and discriminant validity, and correlated .40–.42 with an inbox measure of "ethical decision making." Thus, combining the results of Solomonson (2000) and Manley et al. (2006), the use of biodata as an alternative format for integrity tests appears promising.

However, we note that the distinction between biodata and personality items is often blurred, particularly when biodata is not restricted to reports of past behavior. In addition, many integrity tests include admissions items, which might be classified as biodata. In the present case, Solomonson's biodata scale correlates roughly as highly with integrity tests as integrity tests do with each other. In sum, biodata as an alternative to integrity tests is worthy of further exploration, though the degree to which biodata represents a distinctive alternative to existing integrity tests is not yet clear.

Other research has attempted to design interviews as alternatives to paper-and-pencil integrity tests. Hollwitz (1998) developed two different structured interviews, both designed to capture the seven factors underlying one specific written integrity test: the EII. Two 10-item interviews were developed, one where interviewees were asked to describe past behaviors and one where interviewees were asked to describe what they would do in a particular situation. In a student sample, the past behavior and situational interviews correlated .49 (.65 corrected) and .60 (.79 corrected) with the EII, respectively. Fairness perceptions were essentially the same for the test and the two interviews. In order to address criterion-related validity, a single-act behavioral criterion was included in the study: a folder labeled "exam answer key" was left prominently displayed in the room where each participant took the EII, with procedure in place to permit the researcher to determine whether the participant opened the folder. This "snooping" correlated −.08 with the EII and −.02 with the situational interview, but correlated significantly (−.32) with the behavioral interview. In sum, the study shows that information similar to that obtained via a written integrity test can be obtained via interview, at least in a research setting. As interviews are more cost and labor intensive than integrity tests, practitioners would have to weigh these costs against any benefits of an integrity interview.

Integrity tests have also been recast using voice-response technology. Jones, Brasher, and Huff (2002) described a revision of the PSI, now labeled the Applicant Potential Inventory, with the goal of moving from paper-and-pencil to voice response format, where candidates listen to test items by phone and respond using the phone keypad. The use of such a format might be attractive when reading ability is an issue in an applicant pool or when organizations wish to decentralize the screening process.

Jones et al. reported a variety of studies documenting the reliability and criterion-related validity of the revised test. They also reported extensive operational data on subgroup differences using the voice response format. As in prior research, race and gender differences were generally minimal. In short, the study documents an effective transition from a paper-and-pencil format to a voice response format.

Finally, Jackson et al. (2000) recast the Dependability scale from the Employee Selection Questionnaire (Jackson, in press), a personality-oriented integrity test, into forced-choice format. (A second integrity measure, labeled "Giotto," also uses the forced-choice format, but no data are reported on the issue [Rust, 1999]). Jackson et al.'s forced-choice research is described in detail in the previous section on faking. At this point we will simply mention that well-established integrity tests such as the PDI-EI have sections using a forced-choice response format, so it is not clear how new an idea a forced-choice integrity test really is.

Conclusions

A great deal of research has gone into developing new types of integrity tests. Some of these new tests (i.e., conditional reasoning tests) seek to measure "integrity" from new theoretical perspectives. Other new tests (i.e., biodata, interviews, forced-choice measures, voice-response measures) seek to create prototypical integrity tests using new formats. Each of the new types of tests has advantages and disadvantages. For instance, although conditional reasoning tests are an innovative and exciting new development, it appears that initial reports of criterion-related validity may have been overly optimistic. Although some new formats such as interviews and voice-response formats have met with initial success, the value added beyond traditional integrity tests (especially given the labor and costs involved in the development of these new formats) may be an important issue. Although some formats such as biodata show promise in expanding the way in which we think of integrity tests, it is unclear whether these formats can really be called something new. Even with these concerns in mind, though, we voice optimism and encouragement for future research investigating new types of integrity tests.

Legal Developments in Integrity Testing

Legal Threats to Integrity Test Use

There have been no significant changes to the legal climate surrounding the use of integrity tests since the Sackett and Wanek (1996) review. There

has been no new legislation and no court decisions involving integrity tests since the prior review.

There have been a number of law review articles examining and commenting on integrity testing (Befort, 1997; Buford, 1995; Faust, 1997; Stabile, 2002; Vetter, 1999). One recurring issue in these articles addresses whether administering an integrity test to an applicant before hire could be construed as a pre-job-offer medical examination, which is illegal according to the Americans with Disabilities Act (ADA; Befort, 1997; Stabile, 2002; Vetter, 1999). This concern has been dismissed by legal writers because, according to the Equal Employment Opportunity Commission (EEOC) guidelines, a personality test can only be considered a pre-job-offer medical examination if the test was designed and used to identify a mental disorder in the test taker (Befort, 1997; Stabile, 2002; Vetter, 1999). The EEOC guidelines explicitly offer integrity tests as an example of a pre employment test that is not designed or used to point out a mental disorder in the test taker, and thus pre-offer administration is permissible (Befort, 1997). However, of note is an emerging body of literature suggesting that integrity tests could be used to signal a mental disorder. For instance, Iacono and Patrick (1997), Connelly, Lilienfeld, and Schmeelk (2006), and Murray (2002) each reported research linking scores on various overt and personality-based integrity tests to scores on psychopathy measures.

Does this emerging evidence signal legal risk to employers using integrity tests? Our answer is "no," at least for those using integrity tests in the manner intended by their developers. The EEOC Guidelines focus on whether an instrument is designed or used to identify a mental disorder. Integrity tests were not designed for this purpose nor are they used for this purpose: They are used to predict subsequent on-the-job productive and counterproductive behavior. They also do not match any of the features listed by the EEOC as indicating that a procedure may be a medical examination (i.e., is the measure administered or interpreted by a health care professional; is the measure administered in a health care setting; is the measure physiological, or physically invasive?) Thus, unless used in a very different manner from that intended by their developers, we do not foresee integrity tests being reconsidered as medical examinations under the ADA.

Subgroup Differences

Given adverse impact concerns, subgroup differences on integrity tests could signal legal risk. Ones and Viswesvaran (1998a) investigated four large data sets from job applicants who took one of three different overt integrity tests (Reid Report, Stanton Survey, or PSI). Gender data were available for 680,675 job applicants. Women scored .16 *SD* higher than

males on integrity tests. Age data were available for 78,220 job applicants. Ones and Viswesvaran reported that persons over 40 score .08 *SD* higher than persons under 40. Race differences were negligible. Comparing each group to Whites, Blacks ($d = -.04$), Hispanics ($d = .05$), Asians ($d = .04$), and American Indians ($d = .08$) did not exhibit significantly different mean scores. We note that comparable data for personality-based tests are a need for future research.

Van Iddekinge et al. (2005) suggested that subgroup differences may exist at the facet level of integrity tests. Using a sample of 152 customer service managers, Van Iddekinge et al. (2005) examined subgroup differences on eight facets they identified in the PSI-CS, finding some sizable racial/ethnic, gender, and age differences on various facets. Such a finding is interesting, and may warrant future research, but given that selection decisions are generally made using overall test scores instead of facet scores, the legal risk of facet subgroup differences is not necessarily high.

Conclusions

Although there has been no new case law regarding integrity tests, there is emerging evidence that integrity tests could be used to diagnose mental disorders. We conclude, though, that this is not a significant legal risk as long as integrity tests are only used in their intended fashion. The only other research being done that has legal ramifications is that which addresses subgroup differences. Meta-analytic research suggests subgroup differences are negligible on overt integrity tests, although no research addresses whether findings for overt tests can be extended to personality-based tests. This last is an area in need of further research.

Applicant Reactions

Because integrity tests delve into touchy subject matter (e.g., determining whether an applicant has enough "integrity"), applicant reactions to integrity tests continue to be studied. As a whole, such studies ask applicants to report their reactions (e.g., perceptions of fairness, face validity, fakability, etc.) to integrity tests. For instance, a meta-analysis by Hausknecht, Day, and Thomas (2004) compared mean ratings of the favorability of nine personnel selection procedures (e.g., interviews, work samples, cognitive ability tests, etc.) to favorability ratings of integrity tests, and found that integrity tests were rated lower than all methods except graphology. Although this suggests that integrity tests are viewed relatively negatively, it should be mentioned that in all of the studies meta-analyzed by Hausknecht et al., respondents did not actually experience the selection procedures they were rating but instead were simply presented

with written descriptions of each procedure. This, combined with evidence that respondents do not generally react especially negatively to integrity tests when they actually take an integrity test (see Sackett & Wanek, 1996; Whitney, Diaz, Mineghino, & Powers, 1999), calls into question how generalizable are Hausknecht et al.'s negative findings for integrity tests.

A selection method not included in Hausknecht et al. was drug testing. Two papers by Rosse and colleagues (Rosse et al., 1996; Rosse, Ringer, & Miller, 1996) compared reactions to integrity testing and drug testing. These authors suggested that integrity tests should be perceived more positively by applicants than urinalysis because they are less invasive and do not presume guilt. In the first study, college students were placed into a testing condition (no test, overt integrity, personality-based integrity, urinalysis) and their attitudes toward the firm, behavioral intentions to apply, and intentions to accept a job offer were subsequently measured. Results indicated that across a number of different reactions measures, reactions were most positive in the no-test condition, followed by overt and urinalysis (which were both generally significantly lower than no test, but not significantly different from each other), and then personality-based test (which was generally significantly lower than overt and urinalysis). In a related study, Rosse et al. (1996) found that persons classified as "drug users" equally disliked urinalysis, overt, and personality-based integrity tests. Drug use was negatively correlated with applicant reactions (in other words, higher drug use was related to more negative attitudes, ranging from $r = -.08$ to $-.48$). Thus, in the two Rosse and colleagues studies, integrity tests were viewed as negatively as urinalysis (and sometimes viewed more negatively).

The above studies mostly focused on comparing reactions to integrity testing to reactions to other selection methods. Another interesting question is whether types of integrity tests are reacted to differently. In a study of the relative reactions of participants to overt versus personality-based tests, Whitney et al. (1999) asked a sample of primarily female college students to complete reaction measures (e.g., their estimation of face and predictive validity) after completing the Phase II Profile and the PRB. Later, participants were given their integrity test scores, publisher test norms, and a statement telling the participants whether they had "passed"; and then participants were asked to complete a distributive justice survey. Overt tests were perceived as more face valid and predictive valid than personality-based integrity tests, supporting similar findings by Rosse and colleagues. These validity perceptions were unrelated to test performance, but those who "passed" integrity tests viewed them as more distributively fair.

Dwight and Alliger (1997b) noted that previous research on applicant reactions has focused primarily on reactions to test *types* (e.g., cognitive

ability, integrity) as opposed to test *items*. Therefore, a small undergraduate sample was given a bank teller job description and asked to rate EII items in terms of their invasion of privacy, ease of faking, and job relatedness. Items were sorted by subject matter experts into question types such as admissions (e.g., "I stole X dollars from my employer last month"), "lie questions" measuring the extent to which test takers respond in a socially desirable manner, and protecting others who have engaged in deviant behaviors. Items of the admit-behavior type were perceived as most invasive and easiest to fake, but were considered relatively job related. Lie questions were perceived as least job related and hardest to fake. "Protection of others" items were considered most job related.

The previous two studies demonstrated that test and item content may moderate applicant reactions to integrity tests. Polson and Mullins (manuscript under review) suggest that the response format may also moderate applicant reactions. Polson and Mullins noted that narrow response formats may make test takers feel unable to perform to their full potential. Thus, Polson and Mullins compared two response formats (true–false vs. a five-point Likert scale) for an integrity test they created by examining test manuals of various existing integrity tests. Half of their sample of 86 undergraduates was given the true–false version of the integrity tests, and the other half was given the five-point scale version. Participants who received the five-point scale version rated the integrity test as fairer ($d = 1.22$) and more face valid ($d = .58$), and felt they had performed better ($d = 1.03$). Thus, it appears that five-point response formats may be viewed more positively than true–false formats, though research with other common response formats and with established integrity tests is needed.

Conclusions

Applicant reactions are a popular area of inquiry for industrial-organizational psychologists (e.g., a special issue of *International Journal of Selection and Assessment* on the topic in 2004). Research seems to suggest that respondents do not have especially positive reactions to integrity tests, though the reactions appear to be moderated by such factors as the type of integrity test, the specific items in the integrity test, and the response format used. However, a variety of key questions remain unanswered. For instance, do negative applicant reactions have measurable long-term effects on incumbents and organizations? Are there links between negative reactions and deviant behavior in organizations? We encourage research in this area to take a long-term and outcome-focused approach.

Cross-Cultural and Language Translation Issues

Cross-Cultural Data on Prevalence of Integrity Test Use

Ryan, McFarland, Baron, and Page (1999) reported data on how often various personnel selection methods were used in 959 organizations in 20 countries for managerial selection. Across the 20 countries, integrity test use is quite limited (an average rating of 1.33, where 1 = *never* and 2 = *rarely*). This, and all other findings on the prevalence of integrity testing in the Ryan et al. study, should be interpreted with caution though because their prevalence questions were referring to how often integrity tests were used to select managers, even though integrity tests are more commonly used for lower-level, entry-level, or retail/sales jobs. Nonetheless, there are interesting findings from Ryan et al.'s data. The first is that there were differences in prevalence rates between countries. For instance some countries (e.g., Belgium, Canada, Greece, among others) reported above-average use of integrity tests, whereas others (e.g., United States, France, Germany, among others) reported very low rates of integrity testing.

Further, we computed the correlation between each country's reported power distance and the prevalence of integrity testing, and found $r = .48$. Power distance was defined as the extent to which the less powerful people in a society accept the fact that power is distributed unequally. Countries higher in power distance tend to accept and expect hierarchical decision making. Organizations in countries high in power distance may not be as worried about applicant reactions because their applicants are more likely to accept decisions from the organization about what constitutes an appropriate selection procedure.

In addition, countries rated as high on uncertainty avoidance (the extent to which members of a culture feel threatened by uncertain or unknown situations) tend to also have a higher prevalence of integrity testing ($r = .40$). Ryan et al. cite Stohl (1993), who notes that organizations in cultures high in uncertainty avoidance should engage in more structuring activities, such as standardization of practice, whereas those low in uncertainty avoidance would be more tolerant of spontaneity in practice. So, one explanation for this correlation is countries that are high in uncertainty avoidance are willing to do more (e.g., administer integrity tests) to reduce uncertainty in the hiring process.

Translation and Generalization of U.S. Measures to Other Cultures

There has been research on translating integrity tests into other languages to determine if the tests still exhibit the same properties.

Fortmann, Leslie, and Cunningham (2002) translated the Abbreviated Reid Integrity Inventory into appropriate languages for Argentina, Mexico, and South Africa. Using these translated versions, Fortmann et al. found no major differences between these three countries and the United States in terms of scale means and standard deviations, criterion-related validities, or CWB admissions. There were also no gender differences across or within countries. Marcus et al. (2007) translated the Inventar Berufsbezogener Einstellungen und Selbsteinschätzungen (Marcus & Schuler, 2004), a German test with both overt and personality-based components, into English and translated the CPI-Cp (an English-language personality-based test) into German, and then administered the tests to Canadian and German samples, finding the tests to be essentially equivalent across samples. Posthuma and Maertz (2003) reported attempts to translate integrity items into Spanish and administer them to Mexican samples (Posthuma & Maertz, 2003), but due to small sample sizes, the results are inconclusive. Overall, initial attempts to translate or create integrity tests for different cultures have met with some success.

Conclusions

Cross-cultural research regarding integrity testing has been very sparse and is an area requiring further research. A small amount of research has documented differences between countries in the prevalence of integrity testing and that these differences are related to power distance and uncertainty avoidance at the country level. These results must be interpreted cautiously for the present purposes because the focus of Ryan et al. (1999) was on managerial selection. A small amount of research has also begun documenting efforts to either translate existing integrity tests or develop new integrity tests for use in different cultures with some success. It remains unclear whether it is more effective to translate existing integrity tests versus developing new ones.

General Discussion

The increment in our knowledge of and insight into integrity testing since the Sackett and Wanek (1996) review is substantial. Some of the most important work related to advancing our knowledge of integrity tests deals with the constructs underlying integrity tests. Because integrity tests are not all interchangeable (Ones, 1993), a particularly useful recent approach has been to examine integrity tests at the item level by pooling items across multiple integrity tests (e.g., Wanek et al., 2003). Such a technique allows researchers to determine what factors are common and not common to the individual integrity tests contributing items. This item-level research has

identified multiple levels of factors underlying many integrity tests and suggested a hierarchical structure to the construct(s) underlying integrity tests. The item-level research has both substantiated the construct validity work done at the test level (e.g., links between integrity and the Big Five) and challenged the work done at the test level (e.g., links between integrity and cognitive ability). Of particular interest would be work (either at the item, scale, or test level) addressing variance in integrity tests beyond the Big Five. Although establishing the link between the Big Five and integrity tests was an extremely important discovery, it is time for integrity researchers to more vigorously look beyond the Big Five to fill in the rest of the construct validity picture.

Another area of research that has blossomed since the Sackett and Wanek (1996) review is the development of new types of integrity tests. Some of these new types of tests seek to measure "integrity" from new theoretical perspectives, whereas others seek to create prototypical integrity tests in new formats. This work is innovative and deserves further attention. There are at least two main concerns that these lines of research will need to address. First is the degree to which they truly measure the construct of "integrity" better than integrity tests. That is, when one creates a measure from a new theoretical perspective, when does that theoretical perspective diverge sufficiently enough that the new test no longer measures "integrity?" This may be less of a concern if new measures prove to predict criteria such as CWB as well or better than existing integrity tests (something that would require a large cumulative literature), regardless of the constructs the new tests measure. Either way, this is an issue that new tests need to address. Second, new types of tests need to make clear their value added beyond existing integrity tests. We note that integrity tests have a large cumulative literature that has established their high criterion-related validity, substantial incremental validity, and low relative cost of administration. New types of tests will need to make very clear why they are viable alternatives to existing integrity tests.

In terms of incremental value, the contribution of Ones et al.'s (2003) meta-analysis of the integrity–absenteeism relationship is clearer, and acts as an example of the type of work needed in the integrity domain. The Ones et al. (1993) meta-analysis of integrity test validity broke down studies into two categories: theft criteria and broad criteria, defined as any CWB other than theft, including violence, tardiness, and absenteeism. Ones et al. (1993) produced a mean validity estimate for broad criteria of .39 for overt tests and .29 for personality-based tests. Thus, without Ones et al.'s (2003) new separate meta-analysis focusing on absenteeism, one would have hypothesized based on the earlier meta-analysis that overt tests would be a better predictor of absenteeism than personality-based tests. But Ones et al.'s (2003) more specific meta-analysis concluded that

personality-based tests predict absenteeism better. Research such as Ones et al. (2003) advances our knowledge of criterion-related validity by using specific, interpretable criteria and by providing a more fine-grained analysis than previous work. We encourage more such work.

Finally, in terms of faking, the ultimate questions our research should address is whether actual job applicants fake on integrity tests, what effects this has on selection decision accuracy (instead of simply documenting mean score changes), and if applicants fake, what can organizations do about it? In a perfect world this entails the use of job applicant samples in faking research. When applicant samples cannot be used, laboratory studies must be very careful about what instructional sets most closely mimic the applicant setting. In terms of applicant reactions, the ultimate questions our research should address are whether job applicants' reactions to integrity tests have any behavioral or long-term implications. Are these reactions transient or lasting? Do these negative reactions have any behavioral implications (e.g., CWB, decreased motivation, etc.)?

Conclusion

A quite sizable body of new literature on integrity tests has appeared since the last review of this literature in 1996. New test types continue to be developed. Validation work against a growing variety of criteria continues to be carried out. Understanding of the constructs underlying integrity tests continues to grow, aided by new work at the item, rather than the scale, level. Work on documenting fakability and coachability continues, as do efforts to increase resistance to faking by changing test formats and efforts to find new ways of detecting faking (e.g., response latency). Examination of subgroup differences continues, with work both at the test and facet level. Interest in integrity testing remains high; it is our hope that this review helps clarify what is and is not known about integrity testing.

REFERENCES

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

Alliger GM, Dwight SA. (2001). Invade or evade? The tradeoff between privacy invasion and item fakability. *Applied HRM Research, 6*, 95–104.

Alliger GM, Lilienfeld SO, Mitchell KE. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7*, 32–39.

Becker TE. (1998). Integrity in organizations: Beyond honesty and Conscientiousness. *Academy of Management Review, 23*, 154–161.

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

Befort SF. (1997). Pre-employment screening and investigation: Navigating between a rock and a hard place. *Hofstra Labor Law Journal, 14*, 365–422.

Bennett RJ, Robinson SL. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85*, 349–360.

Berry CM, Ones DS, Sackett PR. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis *Journal of Applied Psychology, 92*, 410–424.

Berry CM, Sackett PR, Tobares V. (2007, April). *A meta-analysis of conditional reasoning tests of aggression.* Poster to be presented at the 22nd Annual Conference of the Society for Industrial and Organizational Psychology, New York.

Borofsky GL. (2000). Predicting involuntary dismissal for unauthorized absence, lateness, and poor performance in the selection of unskilled and semiskilled British contract factory operatives: The contribution of the Employee Reliability Inventory. *Psychological Reports, 87*, 95–104.

Boye MW, Wasserman AR. (1996). Predicting counterproductivity among drug store applicants. *Journal of Business and Psychology, 10*, 337–349.

Brown RD, Cothern CM. (2002). Individual differences in faking integrity tests. *Psychological Reports, 91*, 691–702.

Byford KU. (1995). The quest for the honest worker: A proposal for regulation of integrity testing. *SMU Law Review, 1*, 329–374.

Caron SJ. (2003). *Personality characteristics related to counterproductive behaviors in the workplace.* Unpublished doctoral dissertation, California State University, Fullerton, CA.

Connelly BS, Lilienfeld SO, Schmeelk K. (2006). Integrity tests and morality: Association with ego development, moral reasoning, and psychopathic personality. *International Journal of Selection and Assessment, 14*, 82–86.

Dalal RS. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*, 1241–1255.

Duehr EE, Sackett PR, Ones DS. (2003, April). *An examination of facet-level relationships between integrity and cognitive ability.* Poster presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Dwight SA, Alliger GM. (1997a). *Using response latencies to identify overt integrity test dissimulation.* Presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Dwight SA, Alliger GM. (1997b). Reactions to overt integrity test items. *Educational and Psychological Measurement, 57*, 937–948.

Faust QC. (1997). Integrity tests: Do they have any integrity? *Cornell Journal of Law and Public Policy, 6*, 211–232.

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.

Hausknecht JP, Day DV, Thomas SC. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. PERSONNEL PSYCHOLOGY, *57*, 639–683.

Hein MB, Kramer JJ, Van Hein JL. (2003, April). *Validity of the Reid Report for selection of corrections.* Poster presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Hogan J, Brinkmeyer K. (1997). Bridging the gap between overt and personality-based integrity tests. PERSONNEL PSYCHOLOGY, *50*, 587–599.

Holden RR. (1995). Response latency detection of fakers on personnel tests. *Canadian Journal of Behavioural Science, 27*, 343–355

Hollwitz JC. (1998). *Investigations of a structured interview for pre-employment integrity screening*. Unpublished doctoral dissertation, University of Nebraska, Lincoln, Nebraska.

Horn J, Nelson CE, Brannick MT. (2004). Integrity, conscientiousness, and honesty. *Psychological Reports, 95*, 27–38.

Hough LM, Schneider RJ. (1996). Personality traits, taxonomies, and applications in organizations. In Murphy KR (Ed.), *Individual differences and behavior in organizations* (pp. 31-88). San Francisco: Jossey-Bass.

Hurtz GM, Alliger GM. (2002). Influence of coaching on integrity test performance and unlikely virtue scale scores. *Human Performance, 15*, 255–273.

Iacono WG, Patrick CJ. (1997). Polygraphy and integrity testing. In Richard R (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 252–281). New York: Guilford.

Jackson DN, (in press), Employee Selection Questionndirc Manual. Port Huron, MI: Sigma Assessment System.

Jackson DN, Wroblewski VR, Ashton MC. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371–388.

James LR, McIntyre MD, Glisson CA, Green PD, Patton TW, LeBreton JM, et al. (2005). A conditional reasoning measure for aggression. *Organizational Research Methods, 8*, 69–99.

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

Lanyon RI, Goodstein LD. (2004). Validity and reliability of a pre-employment screening test: The Counterproductive Behavior Index (CBI). *Journal of Business and Psychology, 18*, 533–553.

LeBreton JM. (2002). *Use of differential framing to measure implicit social cognitions associated with aggression*. Unpublished doctoral dissertation, University of Tennessee, Knoxville, Tennessee.

LeBreton JM, Barksdale CD, Robin JD, James LR. (2007). Measurement issues associated with conditional reasoning tests: Indirect measurement and test faking. *Journal of Applied Psychology 92*, 1–16.

Lee K, Ashton MC, de Vries RE. (2005). Predicting workplace delinquency and integrity with the HEXACO and five-factor models of personality structure. *Human Performance, 18*, 179–197.

Leonard JA. (1996). *Response latency as an alternative measure of performance on honesty tests*. Unpublished doctoral dissertation, University of South Florida, Tampa, Florida.

Lucas GM, Friedrich J. (2005). Individual differences in workplace deviance and integrity as predictors of academic dishonesty. *Ethics and Behavior, 15*, 15–35.

Manley GC, Dunn KM, Beech M, Benavidez J, Mobbs T. (2006, May). *Developing personality-based biodata integrity measures*. Poster presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Marcus B, Lee K, Ashton MC. (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. (2004). Antecedents of counterproductive behavior at work: A general perspective. *Journal of Applied Psychology, 89*, 647–660.

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

McFarland LA, Ryan AM. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812–821.

Mikulay SM, Goffin RD. (1998). Measuring and predicting counterproductivity in the laboratory using integrity and personality testing. *Educational and Psychological Measurement, 58*, 768–790.

Mumford MD, Connelly MS, Helton WB, Strange JM, Osburn HK. (2001). On the construct validity of integrity tests: Individual and situational factors as predictors of test performance. *International Journal of Selection and Assessment, 9*, 240–257.

Murphy KR, Lee SL. (1994). Does Conscientiousness explain the relationship between integrity and job performance? *International Journal of Selection and Assessment, 2*, 226–233.

Murray LD. (2002). *MMPI-2 scores as a predictor of outcomes on the Phase II Profile integrity inventory*. Unpublished doctoral dissertation, University of Rhode Island, Kingston, Rhode Island.

Nicol AM, Paunonen SV. (2001). Validity evidence for the different item styles of overt honesty measures. *Journal of Business and Psychology, 16*, 431–445.

Nicol AM, Paunonen SV. (2002). Overt honesty measures predicting admissions: An index of validity or reliability. *Psychological Reports, 90*, 105–115.

Ones DS. (1993). *The construct validity of integrity tests*. Unpublished doctoral dissertation, University of Iowa, Iowa city, Iowa.

Ones DS, Viswesvaran C. (1998a). 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 DS, Viswesvaran C. (1998b). The effects of social desirability and faking on personality add integrity assessment for personnel selection. *Human Performance, 11*, 245–269.

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

Ones DS, Viswesvaran C, Schmidt FL. (1996, April). *Group differences on overt integrity tests and related personality variables: Implications for adverse impact and test construction*. Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, California.

Ones DS, Viswesvaran C, Schmidt FL. (2003). Personality and absenteeism: A meta-analysis of integrity tests. *European Journal of Personality, 17*, S19–S38.

Polson J, Mullins ME. *Impact of integrity test response format on applicant perceptions of fairness, face validity, and performance*. Manuscript under review.

Posthuma RA, Maertz CP. (2003). Relationships between integrity-related variables, work performance, and trustworthiness in English and Spanish. *International Journal of Selection and Assessment, 11*, 102–105.

Putka DJ, McCloy RA. (2004). Preliminary AIM validation based on GED Plus program data. In Knapp DJ, Heggestad ED, Young MC (Eds.), *Understanding and Improving the assessment of individual motivation (AIM) in the army's GED Plus Program*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rosse JG, Miller JL, Ringer RC. (1996). The deterrent value of drug and integrity testing. *Journal of Business and Psychology, 10*, 477–485.

Rosse JG, Ringer RC, Miller JL. (1996). Personality and drug testing: An exploration of the perceived fairness of alternatives to urinalysis. *Journal of Business and Psychology, 4*, 459–475.

Rotundo M, Sackett PR. (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.

Rust J. (1999). The validity of the Giotto integrity test. *Personality and Individual Differences, 27*, 755–768.

Ryan AM, McFarland L, Baron H, Page R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. PERSONNEL PSYCHOLOGY, *52*, 359–391.

Ryan AM, Sackett PR. (1987). Pre-employment honesty testing: Fakability, reactions of test takers, and company image. *Journal of Business and Psychology, 1*, 248–256.

Ryan AM, Schmit MJ, Daum DL, Brutus S, McCormick SA, Brodke MH. (1997). Workplace integrity: Differences in perceptions of behaviors and situational factors. *Journal of Business and Psychology, 12*, 67–83.

Sackett PR, Berry CM, Wiemann SA, Laczo RM. (2006). Citizenship and counterproductive work behavior: Clarifying relationships between the two domains. *Human Performance, 19*, 441–464.

Sackett PR, Burris LR, Callahan C. (1989). Integrity testing for personnel selection: An update. PERSONNEL PSYCHOLOGY, *42*, 491–529.

Sackett PR, Decker PJ. (1979). Detection of deception in the employment context: A review and critique. PERSONNEL PSYCHOLOGY, *32*, 487–506.

Sackett PR, DeVore CJ. (2002). Counterproductive behaviors at work. In Anderson N, Ones DS, Sinangil HK, Viswesvaran V (Eds.), *Handbook of industrial, work, and organizational psychology* (Vol. 1, pp. 145–164). London: Sage.

Sackett PR, Harris MM. (1984). Honesty testing for personnel selection: A review and critique. PERSONNEL PSYCHOLOGY, *37*, 221–245.

Sackett PR, Wanek JE. (1996). New developments in the use of measures of honesty, integrity, Conscientiousness, dependability, trustworthiness, and reliability for personnel selection. PERSONNEL PSYCHOLOGY, *49*, 787–829.

Schmidt FL, Hunter JE. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Solomonson AL. (2000, April). *Relationships between the Big Five, integrity, and construct-oriented biodata.* Presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Stabile SJ. (2002). The use of personality tests as a hiring tool: Is the benefit worth the cost? *University of Pennsylvania Journal of Labor and Employment Law, 4*, 279–314.

Stohl C. (1993). European managers' interpretations of participation: A semantic network analysis. *Human Communication Research, 20*, 97–117.

Van Iddekinge CH, Raymark PH, Eidson CE, Putka DJ. (2003, April). *Applicant–incumbent differences on personality, integrity, and customer service measures.* Poster presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Van Iddekinge CH, Taylor MA, Eidson CE. (2005). Broad versus narrow facets of integrity: Predictive validity and subgroup differences. *Human Performance, 18*, 151–177.

Vetter GR. (1999). Is a personality test a pre-job-offer medical examination under the ADA? *Northwestern University Law Review, 93*, 597–640.

Wanek JE, Sackett PR, Ones DS. (2003). Towards an understanding of integrity test similarities and differences: An item-level analysis of seven tests. PERSONNEL PSYCHOLOGY, *56*, 873–894.

Whitney DJ, Diaz J, Mineghino ME, Powers K. (1999). Perceptions of overt and personality-based integrity tests. *International Journal of Selection and Assessment, 7*, 35–45.

Young MC, McCloy RA, Waters BK, White LA. (2004). An overview of AIM and the preliminary efforts to support its operational use. In Knapp DJ, Heggestad ED, Young MC (Eds.), *Understanding and improving the assessment of individual motivation (AIM) in the Army's GED Plus Program*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.