Controversies

Integrity Testing for Personnel Selection

Paul R. Sackett

Paper-and-pencil predictors of counterproductive work behavior have been in use for more than 50 years. However, the number of tests available and the extent of their use for personnel selection have grown rapidly over the past decade. Recent estimates suggest that several million such tests are administered in the United States each year. Although integrity is an issue across the full spectrum of jobs in the workplace, integrity testing is used predominantly for low-paying, entry-level jobs in settings where employees have access to money or merchandise (e.g., retailing, financial services). Many of the early integrity tests were initially billed as paper-and-pencil surrogates for the polygraph, and evidence of their validity typically consisted of correlations between test scores and polygrapher judgments. These tests arose outside mainstream psychological testing, and were generally sold not to psychologists but to loss-prevention or human resource personnel in organizations. Scoring keys were typically not available to the firms using the tests; raw responses had to be mailed or phoned to the publisher for scoring.

In the past decade, a number of trends have developed. First, there is increasing involvement by psychologists in this area of testing. Second, tests are being marketed quite differently. The link to the polygraph is either downplayed or eliminated completely, and the tests are billed as predicting not only employee theft, but a wide variety of counterproductive behaviors, including violation of work rules, fraudulent worker’s compensation claims, and absenteeism. Third, markedly different types of paper-and-pencil measures have emerged. Three basic types of tests can be identified:

- Overt integrity tests. Alternately labeled clear-purpose tests, these commonly consist of two sections. The first, a measure of theft attitudes, includes questions concerning beliefs about the frequency and extent of theft and other counterproductive behavior, punitiveness toward theft, ruminations about theft, perceived ease of theft, endorsement of common rationalizations for theft, and the test taker’s assessments of his or her own honesty. The second asks applicants to report the frequency and amount of their own involvement in theft and other illegal or counterproductive activity. These two sections may also be packaged with other scales intended to measure factors such as drug abuse and tendencies toward violence. Commonly used tests of this type include the London House Personnel Selection Inventory, the Reid Report, and the Stanton Survey.

- Personality-oriented measures. Alternately labeled disguised-purpose tests, these are typically developed by psychologists and are closely linked to normal-range personality assessment devices, such as the California Psychological Inventory. Personality-oriented measures are considerably broader in focus than overt integrity 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 Personnel Decisions, Inc. (PDI), Employment Inventory, and the Reliability Scale of the Hogan Personality Series.

Paul R. Sackett is Professor of Industrial Relations at the University of Minnesota. Address correspondence to Paul R. Sackett, Industrial Relations Center, University of Minnesota, 271 19th Ave. South, Minneapolis, MN 55455.
Clinical measures. Measures developed for other uses, such as the Minnesota Multiphasic Personality Inventory (MMPI), may be used by a clinical psychologist to form a diagnostic opinion about an applicant’s integrity.

The research base for integrity testing focuses virtually exclusively on the first two categories of tests.

THE RESEARCH BASE

Within the field of employment testing, fairly standard technologies for evaluating the validity of tests have been developed. On the surface, what is needed is straightforward: test a group of applicants, hire without regard to test scores, measure subsequent on-the-job behavior, and correlate test scores with these measures of job behavior. In the area of integrity testing, measuring subsequent on-the-job behavior has proven to be the Achilles’ heel of the validation process. It is generally acknowledged that much theft and other counterproductive behavior goes undetected (if detection were easy, there would be considerably less need for or interest in preemployment screening). Thus, if an individual fails a test but is not caught stealing on the job, a test critic, on the one hand, may label this a “miss” and use this case as evidence of test invalidity. A test advocate, on the other hand, may assert that the individual may in fact simply not have been caught.

As a result of this criterion problem, a variety of other research strategies have also been used to investigate integrity tests. These strategies include correlation with polygraph test results, correlation with anonymous admissions of theft, use of time series to examine aggregate rates of inventory shrinkage before and after the introduction of a testing program, and comparison of test performance by groups hypothesized to differ in integrity (e.g., convicted felons vs. job applicants). Recent work has made more extensive use of more accessible criterion measures, such as absenteeism. Although these measures are less fraught with interpretational difficulties than are measures of detected theft, questions arise as to whether the various criterion measures (e.g., absence and theft) can be interpreted as reflections of a common underlying criterion construct labeled “counterproductivity.” Very little is known about the interrelationships among the array of counterproductive behaviors that have been used as criteria in studies of integrity tests.

A recent meta-analysis of integrity test validity studies located 665 validity coefficients. This quantitative analysis makes clear that differences in research design have a large impact on outcomes of validity studies: Markedly higher validities were found when self-reports of wrongdoing were used as criteria than when external (i.e., non-self-report) criteria were used. Among studies using external criteria, markedly lower validities were found in studies using detected theft as the criterion, in comparison with studies using other external measures (e.g., absenteeism). Typically, higher validities were found in concurrent studies (i.e., studies administering tests to job incumbents) than in predictive studies (i.e., studies administering tests to job applicants).

These design issues confound attempts to compare validities for different types of tests. The modal validity study for overt tests used a concurrent validity strategy and a self-report criterion; the modal study for personality-oriented tests used a predictive validity strategy and an external criterion. The most compelling studies are those using a predictive validity strategy and external criteria. For overt tests, 7 such studies using theft criteria produced a mean validity of .39, and 10 studies using nontheft criteria produced a mean validity of .39. For personality-based tests, 62 such studies, all using nontheft criteria, produced a mean validity of .29. (These correlations have been corrected for range restriction [i.e., lower test variance in the study sample than in the applicant pool] and criterion unreliability, as uncorrected correlations are downwardly biased. The uncorrected correlations are .09, .27, and .20, respectively.)

A recent trend in integrity test validation research is the use of supervisory ratings of overall job performance, rather than measures of counterproductivity, as criteria. In these studies, using a predictive strategy, the mean validity has been .41 (.25 prior to correction for range restriction and criterion unreliability), and overt and personality-oriented measures have produced comparable results. One interpretation of these findings is that integrity tests are measures of the broad construct of general conscientiousness, a dimension emerging consistently in work on the structure of personality. Research on measures of conscientiousness has shown this dimension to be predictive of job performance across the occupational spectrum.

Thus, a large body of validity evidence consistently shows scores on integrity tests to be positively related to both a range of counterproductive behaviors and supervisory ratings of overall performance. However, virtually all the research has been done by test publishers, leading skeptics to question whether only successes are publicized.

LEGAL STATUS OF INTEGRITY TESTS

At the federal level, there is currently no regulation of the use of integrity testing. There has been extensive documentation that women, racial minority groups, and older
workers do not systematically perform more poorly than other groups on these tests, and thus there have been no successful challenges to integrity tests under federal antidiscrimination laws.

At the state level, within the past few years, there has been considerable activity aimed at restricting integrity testing. Massachusetts prohibits written examinations for the purpose of rendering a diagnostic opinion regarding the honesty of an individual. Rhode Island law has similar language, though it allows written tests to be used if they are not the primary basis for an employment decision. This may signal a trend: With the support of labor unions and employee rights groups, legislation has been introduced in other states.

**WHY ARE THESE TESTS SO CONTROVERSIAL?**

Integrity tests have proven controversial for a variety of reasons. First, when the notion of measuring honesty surfaces, many psychologists recall the conclusion drawn by Hartshorne and May in their classic early work on the topic, namely, that there is no such thing as a trait called honesty. Less well recalled is their basis for that conclusion, namely, that they failed to support their hypothesis that honest people would be honest in all situations and dishonest people dishonest in all situations. As more recent commentators have noted, perfect consistency of behavior is an unrealistic standard; reanalyses of the data do show a large common factor underlying measures of honesty across situations.

Second, both the view of these tests as polygraph surrogates and the marketing of early tests by polygraphers rather than psychological test publishers have tarnished the image of integrity tests among psychologists. Although the vast majority of test validity research is now independent of the polygraph, the level of professionalism exhibited by test publishers remains an area of concern. The quality of supporting research and the accuracy of marketing claims vary widely.

Third, some people argue that the concepts of honesty and integrity hold a special place in the panoply of human characteristics. Honesty is indeed the single characteristic we value most in others. It is also one of the few characteristics on which virtually no one rates himself or herself as below average. Given the sensitive nature of the construct, some commentators, such as the U.S. Congress Office of Technology Assessment (OTA), have asserted that measures of this construct should be held to a higher standard than measures of other constructs. Misclassification ("being wrongly labeled a thief") is seen as a more serious problem than with other types of tests. A number of aspects of the misclassification issue merit attention.

One important point is that misclassification rates for integrity tests are difficult to assess—in fact, I do not believe that there is a factual basis for making such an assessment at present. Consider the data in Table 1, from a study that influenced the conclusions of OTA’s critical 1990 report. Note that only 17 individuals were caught stealing. OTA’s analysis of these data was based on the implicit assumption that all thieves were caught. Only if this assumption is made can one make statements about the accuracy of the test. OTA focused on the 238 individuals failing the test, noted that 222 of them were not caught stealing, and concluded that 93.3% of individuals who fail the test are misclassified. It did not give serious attention to the possibility that some unknown proportion of the 222 individuals who failed the test and were not caught stealing may either have stolen but not been caught or have engaged in some other counterproductive behavior that would justify a “do not hire” recommendation based on the test score.

A second important aspect of the misclassification issue is that misclassification rates are interpretable only in comparison with alternatives. A test that misclassifies, say, 25% of test takers may prove dismal, or it may prove a great improvement over available alternatives.

A good example of the failure to apply this form of analysis to the misclassification problem can be found in comments made by Senator Edward Kennedy during debate about the Employee Polygraph Protection Act. Kennedy noted that even if one accepts the claims of polygraph proponents that polygraph examinations have an accuracy rate of 95%, if 1 million preemployment polygraph exams are administered annually, a 5% error rate means that 50,000 innocent job applicants are misclassified. Kennedy argued that any device that misclassifies 50,000 people per year should be banned. This argument reduces all personnel selection to the absurd, in that all selection devices, from tests to interviews, are certainly less than perfectly accurate, and thus should be banned by Kennedy’s standard. Yet the alternative—random selection or first-come, first-

<table>
<thead>
<tr>
<th>Table 1. Integrity test results and detected theft</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theft category</strong></td>
</tr>
<tr>
<td>No detected theft</td>
</tr>
<tr>
<td>Detected theft</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
served—results in higher rates of misclassification than any selection device with nonzero validity.

A third issue concerns assertions about the stigma of misclassification. In contrast to an American Psychological Association report\(^7\) arguing that integrity tests should be held to the same standards as other tests—standards outlined in the Standards for Educational and Psychological Testing\(^8\)—OTA asserted that integrity tests should be held to a higher standard because of the stigma of being mislabeled as dishonest. If integrity tests become widespread, individuals mistakenly labeled dishonest could be permanently barred from the work force. In addition, labeling may create a self-fulfilling prophecy: If organizations already view an individual as dishonest, the individual may conclude that there are no longer sufficient incentives to refrain from dishonest behavior.

I do not find the OTA position persuasive. First, job applicants rarely receive feedback about test scores. Typically, rejected applicants are turned away with something like "We regret to inform you that we've filled the position with someone who provides a better fit to our present needs"; without communication of test performance to applicants, the issues of stigma and self-fulfilling prophecy become moot.

Second, the speculation about tests becoming a permanent bar to employment seems to me unrealistic. Personnel selection research documents clearly that job performance is multidimensional, and that a variety of individual difference constructs are related to various dimensions of job performance. As organizations strive for competitive advantage, they differentially value aspects of job performance. Consider retail sales, a setting in which use of integrity tests is common. One firm may value speed and accuracy in completing customer transactions, and thus may screen job applicants for cognitive ability. Another firm may value customer service, and thus may screen applicants for a constellation of personality factors labeled "service orientation." A third firm may value controlling costs, and may screen applicants using an integrity test, hoping to reduce inventory shrinkage. The opportunity to be selective is a scarce resource; that is, with a limited applicant pool, an employer may find it difficult to be highly selective in several areas of job performance. For example, if the ratio of job applicants to job openings is 2 to 1, a firm cannot simultaneously demand that applicants score in the top 50% in cognitive ability, service orientation, and integrity. Screening in the top 50% in any one of these areas produces as many applicants passing the test as there are jobs, precluding any further screening unless the employer is willing to let jobs go unfilled.

**CONCLUSION**

The past decade has seen great changes in the field of integrity testing. The number of commercially available tests has increased. The involvement of mainstream psychology in this area has increased. Linkages to the broader field of personality assessment are being pursued, leading to interesting hypotheses emerging about one personality factor—conscientiousness—being the common link among the various tests. The research base is expanding rapidly: A 1984 review located 45 measures of effect size,\(^1\) but a recent meta-analysis located 665 measures of effect size.\(^2\) Controversy remains about the broader consequences of integrity test usage—a controversy that has been fueled by misinterpretation of what existing validity research can tell us about misclassification rates and by a tendency of critics to overlook a basic axiom of personnel selection, namely, that the standard of comparison for any selection system is not perfect accuracy, but the degree of predictive accuracy achieved by available alternatives.

**Notes**

1. For reviews of the integrity-testing literature, see P.R. Sackett and M.M. Harris, Honesty testing for personnel selection: A review and critique, Personnel Psychology, 37, 221–245 (1984); and P.R. Sackett, L.R. Burris, and C. Callahan, Integrity testing for personnel selection: An update, Personnel Psychology, 42, 491–529 (1989). A broader treatment of workplace honesty can be found in K.R. Murphy, Honesty in the Workplace (Brooks-Cole, Pacific Grove, CA, 1993).


This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.