Chapter 2

INTEGRITY TEST VALIDITY: CONCEPTS AND EVIDENCE

Chapter 2

INTEGRITY TEST VALIDITY: CONCEPTS AND EVIDENCE

Perhaps the first and most often asked question about integrity tests is whether they are valid. This question is not easily or intuitively answered. At its simplest, the question means "do the tests work?" -- or, "do people who do well on the tests actually tend to act 'honestly' more than those who do poorly?'

Beneath these rather obvious questions are layers of subtle problems that have challenged generations of psychologists and other social scientists: Is dishonesty a character trait? If so, is it permanent and does it manifest similarly in all workplace settings? Can written tests effectively and reliably expose the presence of the trait (if it exists) and/or an individual's propensity to commit certain behaviors of interest? Why probe attitudes, intentions, or feelings if evidence of past behavior is available and is considered a powerful predictor of future behavior?

Because the answers to these and related questions can influence decisions affecting many people, they raise a set of formidable public policy concerns (see also chapter 3). And even if one wished to concentrate on the purely empirical question -- how well do the tests do what they are purported to do? -- the research challenge is impressive. Gathering evidence to compare the behavior of individuals with different test scores, drawing statistically valid inferences (predictions) from those scores about individual propensities to act in certain ways, and establishing reasonable levels of confidence in those predictions require a mobilization of sophisticated analytical methods.

This chapter discusses these issues and reviews empirical research on the validation of integrity tests. Discussed first are general issues in validity: What is meant by validity and what are the important issues in test validation? Construct validity, content validity, predictive validity, test reliability, and the internal validity (research design) of studies designed to demonstrate test validity are described, as well as the relatively new concept of consequential validity. In the next section, studies designed to evaluate the construct and predictive validity of integrity tests are described and discussed. Particular attention is paid to issues of the quality of the research that has been conducted.

TEST VALIDITY: GENERAL ISSUES

Although intuitively appealing, the implied definition of validity in the opening sentences of this chapter is not, technically speaking, correct. For it is not a test, per se, which is valid or invalid; rather it is the set of inferences drawn from a test: "Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores.'" In common parlance it is customary to refer to a test's validity in either-or terms: either the test is valid or it is not. But measurement theorists now recognize that validity is a form of evaluation, of a number of issues, and that the result of the evaluative process is a sense of the relative strength or weakness of the inferences drawn from test scores. These varieties of evaluative information, which measurement scientists have attempted to group under various headings such as "content" or "construct" or "criterion-related" validity, come together in an "... argument [that] must link concepts, evidence, social and personal consequences, and values."[2] In a word, then, the best that can be said about any test is that attempts to validate it yield persuasive and acceptable inferences.[3]

Test theorists have identified several components of validation, and while". . . the 30-year old idea of three types of validity. . . is an idea whose time has gone. . . ,"[4] the ideas underlying "content," "construct," and "criterion-related" validity are still very much part of the psychometrician's arsenal.[5]

---

1. S. Messick, "The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement," Test Validity, H. Wainer and H. Braun, (eds.) (Hillsdale, NJ: Lawrence Erlbaum Associates, 1988), p. 33. See also American Psychological Association, Standards for Educational and Psychological Testing (Washington, DC: 1985).

2. L. Cronbach, "Five Perspectives on Validity Argument," in Wainer and Braun, op. cit., footnote 1, p. 4. Cronbach reminds test validators of the importance of what Messick calls "consequential" validity: "Tests that impinge on the rights and life chances of individuals are inherently disputable . . . the bottom line is that validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences." See also S. Messick, "Test Validation and the Ethics of Assessment," American Psychologist, vol. 35, 1980, pp. 1012-1027.

3. "It might also be pointed out that the use of any given test may have as many validities as there are inferences to be drawn from the scores. An integrity test may or may not have much validity for inferences about how generally honest a person is, and it may or may not have much validity for inferences about future counterproductive behavior on a specific job, but these are not interchangeable." Dr. Robert Guion, personal communication, August 1990.

4. Cronbach, op. cit., footnote 2, p. 4.

The construct validity of an instrument is the extent to which one can be sure it represents the construct which it seeks to measure. "A test with good construct validity can be considered a substitute for actually observing a person displaying a skill or attitude in everyday life."[6] Content validity refers to the "representativeness" of the sample of questions on a test, i.e., the extent to which they cover the construct or constructs being measured. "High content validity means that the test 'maps onto' the collection of possible questions by sampling representatively from its various manifestations. . . ."[7]

Both of these aspects of test validity are internal criteria, i.e., they relate to the construction of the test. To determine whether a test measures what it claims to measure, it should also satisfy external criteria: for example, how well the test mimics scores on established and reputable tests that are used to measure similar constructs would be one indication of its ability to measure what it claims to measure. But that would not be sufficient. It is more important to". . . find out whether it correlates with other things implied by [what the test claims to measure] and whether it is uncorrelated with things irrelevant to that claim."[8]

When a test is intended for selection, the most compelling aspect of its validity is the extent to which test scores correlate with later behavior. "Predictive validity," therefore, occupies a central place in discussions of personnel testing in general and of integrity testing in particular. A variant on predictive validity is the so-called "concurrent validity" approach, in which predictors and behaviors are measured at the same time. "Typically, concurrent validity data were taken as evidence that a newly proposed test, or a brief version of an existing test, was measuring a given trait if it correlated strongly with another test already acknowledged to be a measure of that trait . . . concurrent validity was, and still is, held to be useful for predictive purposes if it could be demonstrated, or argued convincingly, that scores on the test would not change systematically during the period between the

---

5. Classification of the various approaches to validity has evolved. During the 1940s and 1950s predictive and concurrent validity were considered separate aspects (or types) of validity, and were later combined under the single heading of "criterion-related" validity. For discussion, see W. Angoff, "Validity: An Evolving Concept," in Wainer and Braun, op. cit., footnote 1, pp. 19-32.
6. Marlene E. Henerson, Lynn L. Morris, and Carol T. Fitz-Gibbon, How to Measure Attitudes (Beverly Hills, CA: Sage Publications, 1978), p. 135.
7. Ibid., p. 140.
8. Dr. Robert Guion, personal communication, August 1990.

time when the test *might* have been given as an actual predictor and the time when criterion data would normally become available."[9]

In addition to these aspects of validity, which pertain to the usefulness of a test as a decisionmaking aid, researchers have begun to incorporate the notion of "consequential" validity in their studies. As argued by one prominent measurement theoretician, ". . . judging whether a test does the job it is employed to do . . . requires evaluation of the intended and unintended social consequences of test interpretation and use."[10] Note, however, the link between consequential validity and other aspects of validity: if adverse consequences can be ascribed to some aspect of score distributions (such as ethnic differences), ". . . which would directly reflect on the functional worth of the selection testing, . . . [the question becomes] whether the adverse impact is attributable to construct-relevant or construct-irrelevant test variance or to criterion-related or criterion-unrelated test variance. . . ."

Another important feature of a test instrument is its so-called "reliability," which reflects ". . . the extent to which measurement results are free of unpredictable kinds of error.'"2 For instance, repeated administrations of a test to the same sample of subjects should yield similar scores. Note that while a valid measure is always reliable, the opposite is not necessarily true: reliability does not necessarily imply validity.[13] Underlying the concept of reliability is the notion of a "true score," i.e., the score that an individual would obtain on a test as a reflection of his or her propensities or abilities. However, when the test is administered, the score falls within some range around this "true" score, and measures of reliability are generally based on estimates of the variability in the observed score around the true score.[14]

---

9. Angoff, op. cit., footnote 5, p. 21.
10. Messick, in Wainer and Braun, op. cit., footnote 1, p. 39.
11. Ibid., p. 40.
12. Henerson et al, op. cit., footnote 6, p. 146.
13. A broken watch is very reliable -- it always tells the same time. But because it provides no information about the real time, it is not valid.
14. For discussion of the technical issues in measurement of reliability, see, e.g., A. Anastasi, Psychological Testing, 3rd ed. (New York, NY: Macmillan, 1968); or Leonard S. Feldt and Robert Brennan, "Reliability," Educational Measurement, R. Linn (ed.) (New York, NY: American Council on Education/Macmillan Publishing, 1989), pp. 105-146.

A related issue is the sensitivity of a test: a test should yield results that can identify differences between two individuals, but it should also not give wildly divergent scores for two fairly similar individuals. A test that is not sensitive to differences is not useful in discriminating between individuals; but an overly sensitive test can lose some of its reliability.

An important consideration in understanding all efforts at test validation is the quality of the research conducted. A valid study design contributes to the confidence that can be placed in a study's results. issues of the quality of a research design are generally known as internal validity. The level of internal validity is the extent to which the relationships detected in a study are not spurious, that is, due to factors not accounted for in the study. Among the factors that may undermine internal validity are: poor sample selection, the occurrence of events during the course of a study that affect the outcome variable in unanticipated ways, nonindependence of observations, and unintended effects on a research subject of being measured.[15] Appropriate use of statistics is another important aspect of study design.

Finally, a critical consideration in determining the quality of any research is the quality and depth of the research report. Because science is a systematic process for creating and disseminating new knowledge, research reports should provide sufficient detail to enable independent scientists to evaluate the credibility of the reported results.

EMPIRICAL VALIDATION OF INTEGRITY TESTS

General Remarks

It should be clear from the foregoing discussion that validation of any test or treatment is a complex process, requiring a balance of subjective judgment and scientific evidence. The determination of construct validity often relies heavily on the opinion of experts who must define the theoretical constructs to be measured in order to identify the presence or absence of specific human traits. Content validation involves the assessment of how well test questions correspond to these

15. See, e.g., T. Cook and D. Campbell, Quasi-Experimentation: Design and Analysis Issues for Field Settings (Chicago, IL: Rand McNally, 1979); and L. Saxe and M. Fine, Social Experiments (Beverly Hills, CA: Sage Publications, 1981).

constructs. Criterion-related validation requires the implementation of the test and the subsequent determination of how the test compares with other measures of the same constructs (to assess "concurrent validity") or how well the test predicts behaviors or actions it is supposed to forecast ("predictive validity"). Concurrent and predictive validity studies require the identification of one or more "criteria," i.e., variables that serve as indicators of the types of behavior under study. Finally, all steps in the process should reflect generally accepted principles of valid research design and should be repotted in enough depth so that the research process is clear to readers.

For integrity testing, validity is especially problematic because integrity and honesty are extremely difficult constructs to define with sufficient precision to enable empirical measurement. On the one hand, the temptation to stick to easily defined acts of dishonesty, such as theft, is stymied by the relatively low frequency of detected theft and, therefore, its limited use as an external criterion. Extending the definition, however, to encompass a wider range of behaviors can result in greater ambiguity about the value of a test as a predictor of the kinds of dishonest acts of greatest interest to employers; "wayward impulse," for example, a construct included in one popular integrity test, maybe a meaningful psychological or characterological trait indicative of a propensity toward certain behaviors, but its usefulness as a predictor of an individual's future commitment of dishonest deeds is tenuous.

Other factors affecting the feasibility or accuracy of empirically validating integrity tests include: the multiple and often unobservable determinants of trends in aggregate measures of organizational productivity, which could confound time series studies of shrinkage; incentives for respondents to answer high-stakes tests strategically, rather than with complete candor, and the possibility that over time job applicants will learn how to answer the tests even more skillfully; and potential biases in criterion measures. Even the reviewers whose analyses end on a relatively optimistic note agree that research in this field faces formidable methodological problems. '6

16. For example, R. Michael O'Bannon, Linda A. Goldinger, and Gavin S. Appleby, Honesty and Integrity Testing: A Practical Guide (Atlanta, GA: Applied Information Resources, 1989): "Unlike much of the earlier research, studies are beginning to appear occasionally in the open literature after review by other professionals. . . . Honesty test publishers will need to become more supportive of independent efforts if a satisfactory body of research and knowledge is to evolve" (pp. 116-117). The 1989 review article by Sackett et al. (P. Sackett, L. Burris, and C. Callahan, "Integrity Testing for Personnel Selection: An Update," Personnel Psychology, vol. 42, 1989) is also cautiously sympathetic

Nevertheless, the amount of research on integrity test validity has increased considerably in recent years, and according to some reviewers the quality of this body of research has improved. For example, one group of reviewers notes that". . . there has been a substantial increase in the number of studies using an external criterion . . . and significant correlations with absence, turnover, behavioral indicators such as grievances and commendations, and supervisory ratings are being reported."[17] These authors were able to report on 24 studies using external, nonpolygraph criteria in their 1989 review (see below for a discussion of problems in studies using polygraph results as criteria) whereas in 1984, they found only 7 such studies.

Aside from methodological problems, a serious issue concerns the proprietary nature of the tests and the fact that ". . . nearly all research is being conducted by investigators associated with honesty test publishers."[18] While this does not necessarily impugn its quality, it does undermine its credibility. The reasons commonly cited for this state of affairs in integrity test research offer little consolation: the proprietary nature of scoring keys, the difficulty in gaining cooperation from some publishers, and the fact that it is not a traditional area for academic research". . . may help explain the lack of independent research, [but] without independent research there is no compelling response to the speculation that only successes are publicized."[19]

Method of OTA's Review

To conduct its review of the research literature on integrity testing, OTA reviewed the two most current reviews of the integrity testing literature,[20] as well as reviews of specific tests published in test review compendiums.[21] OTA also reviewed copies of tests provided by leading publishers, and

---

albeit more favorable in tone than the earlier work by P. Sackett and M. Harris, "Honesty Testing for Personnel Selection: A Review and Critique," Personnel Psycholoay, vol. 37, 1984 pp. 221-245.

17. Sackett et al., op. cit., footnote 16, p. 507.
18. O'Bannon et al., op. cit., footnote 16, p. 117.
19. Sackett et al., op. cit., footnote 16, p. 521.
20. Ibid.; and O'Bannon et al., op. cit., footnote 16.
21. J. Mitchell (cd.), The Ninth Mental Measurements Yearbook (Lincoln, NE: The Buros Institute of Mental Measurements, University of Nebraska-Lincoln, 1985); J. Conoley and J. Kramer (eds.), The Tenth Mental Measurements Yearbook (Lincoln, NE: The Buros Institute of Mental Measurements, University of Nebraska-Lincoln, 1989); and J. Keyser and R. Sweetland (eds.), Test Critiques (Kansas City, MO: Test Corporation of America, 1987). Note that these reviews are written by single individuals, and are not subject to outside review.

reviewed studies conducted by major test publishers. Many studies using counterproductivity as a criterion were supplied by publishers. These studies are not cited, however, in response to the test publishers' request that only studies published in journals be referenced. The studies provided were used to analyze the methodology used by test publishers to conduct such studies. OTA also conducted interviews with a number of experts on various aspects of testing. Some of these experts are intimately familiar with integrity testing; others specialize in related testing issues.

Concurrent Validation Research

One strategy of concurrent validation research is to compare test results with other accepted measures of a particular behavior.[22] There have been numerous attempts to use polygraph 'cores '" this context, some of which have yielded particularly high validity scores.[23] But reviewers have highlighted numerous problems with some studies of this sort, which

> . . . use only the theft attitudes section as the predictor, while others include . . . theft admissions; some use only admissions made during the polygraph as the criterion, while others use polygrapher judgment about the suitability of the candidate for employment; the time interval between the integrity test and the polygraph is often not specified; [it is not] always clear whether or not candidates expected that a polygraph exam would follow the integrity test [in which case individuals would perhaps decide not to conceal, on the integrity test, history of wrongdoing]; some studies preselect equal numbers of individuals passing and failing the polygraph for inclusion in the study, thus maximizing variance in the criterion and increasing the resulting correlation between test and criterion. .. .24

But perhaps the most obvious reason to be wary of concurrent validation studies using polygraph is that polygraph itself has never been demonstrated to be sufficiently valid when used in personnel selection.[25] In one of the two reviews of integrity test validity research, the authors excluded research that ". . . used polygrapher judgments as a criterion . . . because of controversy surrounding the reliability and validity of polygrapher ratings."[26]

---

22. In the physical sciences, for example, a new instrument designed to measure length would obviously need to be validated against previously accepted instrumentation (e.g., the standard meter, wavelength of light, etc.).
23. Product moment correlations were in the range of 0.29 to 0.86 in 14 studies reviewed by Sackett and Harris, op. cit., footnote 16, pp. 221-245.
24. Sackett et al., op. cit., footnote 16, p. 500.
25. See, for example, U.S. Congress, Office of Technology Assessment, Scientific Validity of Polygraph Testing: A Research Review and Evaluation, Technical Memorandum (Washington, DC: U.S. Government Printing Office, 1983), p. 100.
26. O'Bannon et al., op. cit., footnote 16, p. 70,

Although concurrent validation studies are not considered an adequate substitute for predictive validity,[27] these efforts show promise for measuring similarities between the constructs measured by integrity tests and those measured by other personality and cognitive tests.[28]

Validation Research Using "Contrasted-Groups" Method

The basic principle in this approach to construct validation is that". . . if the honesty test is indeed a good measure of integrity, large differences should be found [between the scores of two groups of people who are known a priori to differ in honesty]."[29] There have been less than a dozen such studies, most of which compare honesty test scores of convicted felons and job applicants. The results have generally shown statistically significant differences (as large as two standard deviations) between average test scores of the two groups.[30] Unfortunately, the underlying assumption that convicted felons have attitudes and lifestyles similar (in construct) to those of normal job applicants or employees "who pilfer small amounts of merchandise atwork" cannot be substantiated.

Admissions of Prior Wrongdoing

A common method of validating honesty tests is to compare a test's predictions based on attitudes to an individual's own confessions of wrongdoing, provided contemporaneously. In other words, for a given definition of dishonesty, admissions of prior acts are compared to how closely responses on the test would have been able to predict the propensity to commit those acts. These tests vary in their definitions of honesty; i.e., what kinds of acts to include in confessions, in the methods used to obtain admissions, and in the ways in which scores and admissions data are associated.

While it is believed that admissions provide more data than detected thefts, researchers recognize the inherent limitations to admissions data as criteria: incentives to withhold information, coupled with the bounds on precision of the definition of the acts to be included in admissions, make the admissions criteria very imperfect. A fundamental logical conundrum is that the admission of a past wrongdoing is itself an act of honesty.

---

27. See, for example, Robert Guion, Personnel Testing (New York, NY: McGraw Hill, 1965), p. 371.
28. Sackett et al., op. cit., footnote 16, p. 515.
29. O'Bannon et al., op. cit., footnote 16, p. 70.
30. Sackett et al., footnote 16, p. 512.

The basic conclusion of various reviews is that there is a positive relationship between honesty test scores and confessions, but that ". . . admissions studies are limited to demonstrating a relationship between two types of self-description. . . ."[31]

The use of admissions data as validity criteria also raises a conceptual puzzle. If these data are assumed to be reliable, i.e., if job applicants included in a validity study sample are assumed to confess prior wrongdoing with candor, then why would this assumption not extend to all job applicants? In a word, why not simply ask job applicants about their prior behavior, rather than use tests designed with (imperfect) surrogates for evidence of prior dishonesty?[32] On the other hand, if the answer is that job applicants will have incentives to conceal some information, or to exaggerate other information, then the question becomes whether that type of information can be admissible as criteria in a validation study.

Predictive Validation Using External Criteria

The most compelling line of research on integrity tests is based on the predictive-validity model, which addresses the following basic question: if an integrity test is used in the process of selecting job applicants in order to screen out individuals most likely to commit certain kinds of behavior, to what extent does the test actually predict the relevant behavior? Thus, most industrial psychologists would agree with the statement that ". . . when the objective is to forecast behavior on the basis of scores on a predictor measure, there is simply no substitute for [predictive validity]."=

There have been two basic approaches to validation research using external criteria in which the unit of analysis is the individual: studies using detected theft as the criterion and studies using other external criteria, such as absenteeism, turnover, and supervisors ratings. The trade-off in the

31. Ibid. These reviewers add that ". . . high correlations are found when correlating the attitude and admission sections of various tests; lower correlations are found when single-item measures (admission of arrests, admission of being fired from a previous job) rather than many composites across many illegal activities are used" (p. 508).

32. O'Bannon et al. (op. cit., footnote 16) raise the same question.

33. W. Cascio, Applied Psychology in Personnel Management (Reston, VA: Reston Publishing Co., 1982), p. 150.

value of these studies can be summarized thus: the former address a principal concern, namely theft at the workplace, but are hindered by the difficulty in detecting theft; the latter are more feasible to conduct, but raise concerns about appropriate measures of outcome criteria. A third approach, in which the unit of analysis is the organization, is discussed below under "Time Series Designs." These studies can use either theft or counterproductivity as external criteria.

Theft Studies

A point frequently raised in this report is that workplace theft is a particularly difficult behavior to use as a criterion -- for evaluating any instrument -- if the assumption that a large fraction of workplace theft goes undetected is true.[34] This problem continues to undermine the credibility of predictive validity studies. Because few researchers believe that detected theft is an accurate measure of true theft, the correlations from their studies are probably inaccurate. To clarify this point, suppose that it is known with certainty that some thieves are caught and some are not. Then the correlation found to exist between test score (predictor) and detected theft (criterion) would be lower than the true correlation, as long as those thieves not detected are assumed to score the same as those who are detected. If, however, detection and test performance are not independent, e.g., if the high scorers are the thieves who are best at evading detection, then the observed correlation could be lower, higher, or the same as the true correlation.

in addition to the basic problem of undetected theft, which may not be able to be remedied by improvements in research design and reporting, independent reviewers -- including OTA -- have identified other design flaws in the available studies attempting to use theft as a criterion. For example, there are problems in criterion definition. In one study, mishandling of cash is equated with stealing, when some of the employees so identified may have been careless rather than dishonest.[35] Another study of Salvation Army bellringers had a similar problem; it did not adequately establish that the monetary differences among volunteers' collections resulted from theft, as the researchers concluded; the volunteers could have been in more or less generous locations.[36]

---

34. This assumption does not necessarily mean that there is a very high rate of theft, but rather than whatever the true rate of theft is, much of it is difficult to detect. The question of detection, then, can be distinguished from the question of incidence.
35. O'Bannon et al., op. cit., footnote 16.
36. Alternatively, the volunteers could have spent less time at their posts, an indicator of

In some studies it is difficult to interpret either methods or results for one or more reasons: several scales developed by the same company are used to screen employees, thus preventing an unequivocal assessment of the honesty scale; numbers in subgroups are not reported; test results for these not terminated for theft are not reported; and statistical tests of significance are not presented.[37]

OTA identified five predictive validity studies in which the criterion measure was either detected theft or a reasonably close proxy. The characteristics of these studies, chosen because their research design involved predictive validity, are summarized in table 7. Two of these studies involved applicants for jobs in the grocery industry; two of the studies involved department stores; and one study was of a national convenience store chain. All the studies were conducted by the publisher of the integrity test analyzed in the studies.

Table 8 presents the raw frequency counts as reported in the respective studies. The top row in these tables gives the number of employees not caught committing theft, and the bottom row gives the number detected; these figures are cross-tabulated by test performance as marked in the studies. Note that because some theft undoubtedly is not detected, the bottom row in each table potentially underestimates the true amount of theft. To illustrate the meaning of these tables, consider Study # 2: a total of 3,790 employees were given the test and hired regardless of their test performance. Subsequent investigations by management revealed that 91 employees had committed some type of theft. Among these 91, 75 had failed the integrity test and 16 had passed. Among the 3,699 for whom the investigation did not reveal any theft, 2,145 had failed the test and 1,554 passed. Thus, 75 of those taking the test (2 percent of the total 3,790) are known to have been characterized correctly by the test, and 16 are known to have been characterized incorrectly. But what about the rest? If those 3,699 not detected as thieves are assumed to be honest, then 2,145 (58 percent) were misclassified; if a substantial number of them were indeed thieves, the observed correlation between the test and the outcome measure could be higher, lower, or equal to the actual correlation.

A central concern for public policy is the potential for classification errors, especially of honest

---

counterproductive behavior, though not outright theft (0' Bannon et al., op. cit., footnote 16).

37. See O'Bannon et al., op. cit., footnote 16; and Sackett et al., op. cit., footnote 16.

Table 7- Predictive Validity Studies of Overt Integrity Tests Using Detected Theft or Close Proxy as Criterion

Study	Sample size	Criterion	Test performance* Number passed (percent of total)	Number failed (percent of total)	Number of persons detected committing theft or other dishonest act (percent of total)
1	479	"Thefts detected by admissions and/or signed statements of employees."	241 (50%)	238 (50%)	(3.5%)
2	3,790	'Terminated for reasons of dishonesty."	1,570 (41.4)	2,220 (58.6)	91 (2.4)
3	527	"Discharged for theft or some related offense."	173 (32.8)	354 (67.2)	33 (6.3)
4	61	"Caught stealing cash/ merchandise or disciplined for mishandling company cash/ merchandise."	50 (82.0)	(I8)	6 (9.8)
5	801	"Caught stealing."	472 (58.9)	329 (41.1)	(2.6)

* "Passed" or "failed" in these studies reflect cut scores defined for research purposes. These cut scores mayor may not be the cut scores used by any given employer.

SOURCE: Office of Technology Assessment.

Table 8

Forecasting Efficiency of Integrity Tests (2x2 Contingency Tables for Validation Studies Using Detected Theft or Close Proxy for Criterion)

Study 1

	Failed test	Passed test	Total
Not detected	222	240	462
Detected	16	1	17
TOTAL	238	241	479

Study 2

	Failed test	Passed test	Total
Not detected	2,145	1,554	3,699
Detected	75	16	91
TOTAL	2,220	1,570	3,790

Study 3

	Failed test	Passed.test	Total
Not detected	326	168	494
Detected	28	5	33
TOTAL	354	173	527

Study 4

	Failed test	Passed test	Total
Not detected	8	47	55
Detected	3	3	6
TOTAL	11	50	61

Study 5

	Failed test	Passed test	Total
Not detected	318	462	780
Detected	11	10	21
TOTAL	329	472	801

SOURCE: Office of Technology Assessment.

persons incorrectly identified as dishonest. Table 9 shows that the overall level of misclassification in these studies ranged from 18 percent (in a study with small sample size) to over 60 percent. From less than 1 percent to 6 percent of those passing the tests (that is, identified by the tests as honest) were later found to have stolen from their employers, meaning that upwards of 94 percent of those identified by the tests as thieves were correctly identified.[38] Such reported results are no doubt compelling to employers. But of concern to potential employees, the data in the fourth column of the table suggests why the predictive validity research, even if it is found to be valid, provokes public controversy: of those classified as dishonest on the basis of an integrity test, the proportion who are not detected committing theft ranges from 73 to 97 percent. These data are useful to illustrate the divergence between possible consequences that is at the core of the public policy dilemma.

Counterproductivity-Based Studies

In contrast with the limited amount of research relying on detected thefts for criterion measures, there have been many studies using a variety of counterproductivity-based outcomes, including supervisory data, terminations, and absenteeism. One of the two principal reviews reported on the results of a number of these studies,[39] although they did not evaluate in depth each study's design and conduct.

Measures of counterproductivity used as outcome variables vary considerably. Some measures are specific and discrete (e. g., absenteeism, terminations) and some consist of composites. Some measures are counts from employee records and some are supervisors' ratings. Objective measures of counterproductive behavior include tardiness, absenteeism, accidents, number of worker compensation claims, voluntary turnover, terminations for theft or gross misconduct, and damage to property. Indicators of "productivity," such as mean number of days employed, are also used. Similarly, supervisors' ratings are made of overall performance or misconduct, *or* of more specific

38. It is important to note that the studies used different definitions and measures of theft, and are methodologically flawed.

39. Sackett et al., op. cit., footnote 16. There were no such studies in 1984 when Sackett and Harris conducted their first review. O'Bannon and his colleagues explicitly excluded most studies using counterproductivity as a criterion. There was only one predictive study reviewed by O'Bannon et al. (op. cit., footnote 16) that used terminations as a criterion, and that study focused primarily on terminations for theft.

Table 9- Classification and Misclassification in Five Predictive Validity Studies Using Detected Theft or Close Proxy as Criteria*

Study	Correct classifications: Of those passing test, not detected	Correct classifications: Of those failing test. % detected	Percent of total sample misclassified	Misclassifications: Of those failing test, % not detected	Misclassifications: Of those passing test, % detected
1 & 2	99.6	6.7	46.6	93.3	0.4
	99.0	3.4	57.0	96.7	1.0
3	97.1	7.9	62.8	92.1	2.9
4	94.0	27.3	18.0	72.7	6.0
5	97.9	3.3	40.9	96.7	2.1

* "Passing" and "failing" in these studies reflect cut scores defined for research purposes. These cut scores mayor may not be the cut scores used by any given employer.

SOURCE: Office of Technology Assessment.

measures such as absenteeism and tardiness. This variety of criteria reflects the attempts of researchers to generate useful information. It does, however, make an overall judgment about predictive validity difficult.

Research results from these studies are reported in primarily two ways: (1) in terms of correlation coefficients that serve as a measure of association between integrity test scores and one or more indicators of counterproductive behavior, usually scored continuously; and (2) in terms of proportions of the honest and dishonest individuals who are correctly and/or incorrectly identified by the tests.

As for the theft studies, OTA reviewed a number of counterproductivity studies in order to evaluate their methodology; and as with the theft studies, issues arose with respect to both study design and criterion measurement.

For example, in one study, 169 hotel industry applicants were tested and hired regardless of test scores. The criterion measure was termination. This study, although flawed, suggested somewhat better results, from the point of view of misclassification, than those shown in table 9. First, with respect to persons who "failed" the test: the study showed that among these 53 applicants (31 percent of the total sample), 16 (30 percent) remained employed. Second, among the 116 who "passed," and were therefore presumed honest, 49 (or 42 percent) were eventually terminated. It should be noted, however, that just as detected theft probably underestimates the true amount of theft in the studies reported in table 8, the termination variable in this study probably overestimates dishonesty: there is substantial ambiguity over the causes of termination.[40]

Despite (or because of) flaws in methodology and reporting, the predictive correlational studies reported by Sackett and his colleagues found a range of generally low, but statistically significant, associations between a range of integrity test scores and a wide range of counter-productive measures. Correlation coefficients ranged from 0.16 to 0.62; only one study reported a correlation coefficient greater than 0.35.[41]

40. This is a good example of the trade-off between "criterion variance" and "method variance." See Sackett et al., op. cit., footnote 16, p. 507.

41. From 3 to 38 percent of the variance in counterproductive behaviors would be predicted (explained) by the test scores in a multiple regression model.

It is not possible to ascertain from the studies reporting only correlation coefficients the proportions of honest and dishonest individuals correctly and incorrectly classified. In three studies providing the relevant data, misclassification of dishonest individuals ranges from 17 to 29 percent; in two of these studies, 22 and 29 percent of honest individuals were misclassified. Another study found that the mean number of days employed was significantly higher among those passing the test (95 versus 87 days in the year of the study) .42

Time-Series Designs

Studies that focus on the reduction of organization-level inventory losses and counter-productivity have been termed by some "time-series designs."[43] Almost all of the studies included in the two published independent reviews reported reductions in shrinkage, overall levels of terminations, *or* counterproductive behavior after introduction of the tests.[44] However, flaws in the research designs made it difficult to determine the sources of the change. The most prominent of the flaws was the failure to use appropriate control groups, thus leaving open the possibility that other factors (e.g., seasonal fluctuations in shrinkage; changes in management; perceived changes in company tolerance of theft) accounted for the observed improvements.[45]

In one study,[46] the greatest reduction in shrinkage occurred in the first 2 months after a switch from polygraph to integrity testing screening. The reviewers note, however, that unless there was extraordinarily high turnover, use of the integrity test for selection could not have been the reason for this sudden reduction.[47]

---

42. It may be important to note that Sackett et al. (op. cit., footnote 16) reported both correlation coefficients and dichotomous results for only one study; therefore there is almost no overlap between these types of studies, and results of the studies reporting both types of predictive error may not be generalizable to the studies reporting a single correlation coefficient.

43. O'Bannon et al., op. cit., footnote 16; and Sackett et al., op. cit., footnote 16.

44. Because of differences in measurements used by the various studies, it is not possible to report a meaningful range of results. For example, one study reported a correlation of 0.68 between scores on tests taken by convenience store managers and average monthly store shortage reduction figures. Another reported that 80 percent of all terminations for theft occurred in the control group stores. A third reported a reduction in the termination ratio; a fourth reported both average monthly reductions in terminations for theft and average monthly total voluntary reductions.

45. According to O'Bannon et al. (op. cit., footnote 16), the one study that did use two control groups found that differences in shrinkage among the stores involved in the study were not statistically significant (reported in O'Bannon et al.).

46. Ibid., pp. 88-89.

47. Most employees -- the same ones who were with the company during the baseline measures -- would still be with the company. See ibid., pp. 88-89.

In addition, the following problems were observed in one or more of these studies:

- inappropriate measurement of shrinkage, including shrinkage and cost-savings estimates not based clearly on the study organizations themselves, but on industry averages;[48]
- use of other predictive scales in addition to honesty scales, thus making it difficult to disentangle the effects of the honest scales;[49] and
- concurrent use of polygraph testing for screening a subset of employees.[50]

Reviewers are skeptical about the available time-series studies for these and other reasons, but they believe the results of these studies are grounds for guarded optimism about continued research. While noting that a problem with these studies is the unreliability of the criterion measure (". . . in at least some of the studies it is evident that error is present in the measure of shrinkage. . ."), one reviewer concludes that while ". . . this group of studies cannot be considered unequivocal in demonstrating the validity of honesty tests . . . they do begin to establish a foundation of evidence which may become more convincing as additional studies accumulate."[51]

General Remarks

Industrial and organizational psychologists recognize the difficulty in surmounting methodological barriers to the "ideal" predictive validity study. For example, ". . . the most useful study would be one in which no other selection screening is done, providing a 'pure' examination of the honesty test."[52] The appeal of this model is tempered, however, by the test publishers' claim that their tests are not intended to be the sole (or even the primary) selection criterion.[53] Thus, the truly ideal study would be one in which the various selection procedures continue to be used in combination, but

48. Ibid.
49. Ibid., p. 92.
50. Sackett et al., op. cit., footnote 16.
51. O'Bannon et al., op. cit., footnote 16, p. 92.
52. O'Bannon et al., op. cit., footnote 16, p. 79.
53. See Association of Personnel Test Publishers, Model Guidelines for Preemdoyment Integrity Testing Progams, 1st ed. (Washington, DC: 1990).

which accounts explicitly for the independent effects of the honesty test and for the interaction effects between the test and the other screening procedures. This type of study would not be easy to carry out.

With respect to counterproductivity-based studies using supevisory ratings, in particular, ". . . for a fair assessment to be made, test scores should not be known within the company while the data is being collected . . . [so that the scores cannot] influence the outcome by biasing the opinions of managers toward some employees."[54] In other words, human resource professionals and industrial psychologists recognize a common feature of experiments in the physical and social sciences, i.e., the "double-blind" model. Few of the reported studies indicate whether test scores intended for use in reaching hiring decisions are kept secret from individuals assessing employee performance, and if they were, how it was handled.

Methodological constraints notwithstanding, prominent academic and industrial psychologists, have reviewed the results of the available predictive validity studies. Although these reviews have been conducted by individuals who are generally sympathetic with the objectives of psychological and personnel testing, their findings are couched in cautious tones and their principal conclusion is that better research is very much needed:

> **The most clear cut finding from reviewing predictive validity studies is an observation on the state of this body of research. . . . The field of honesty testing has a great need for producing additional high quality studies in this area.[55]**

54. O'Bannon et al., op cit., footnote 16, p. 79.

55. Ibid., p. 85.