How Useful are Work Samples in Validational Studies?

Douglas N. Jackson*, William G. Harris, Michael C. Ashton, Julie M. McCarthy and Paul F. Tremblay

Some job tasks do not lend themselves to formal on-the-job assessment because they do not occur with sufficient regularity to permit the standardized measurement required in validational research. The preparation of incident reports by police and security officers is such a job task. The production of accurate, literate incident reports is important because these reports are often required in legal proceedings. Their standardized evaluation on the job is not practical because incidents occur at unpredictable intervals with highly variable content. Given the limitations of on-the-job performance criteria, we developed a standardized work sample by preparing sets of non-verbal drawings depicting incidents, each of which required a written descriptive report by security officers. A total of 187 security officers completed a cognitive and personality test battery and criterion incident reports based on the standardized materials. Reports supported the usefulness of the standardized work sample, as well as the validity of the test battery – 100% of participants in the upper quartile of the test score distribution produced satisfactory reports, while only 17% of those in the lowest quartile produced satisfactory incident reports. A number of advantages of structured work samples as criterion measures are noted, including their greater standardization, the elimination of range restriction problems by administering them to all job candidates, the opportunity to obtain expert evaluations of work samples at remote sites, and their face and content validity resulting in acceptability to job candidates and to decision makers.

Dependable on-the-job performance criterion data are extremely difficult, if not impossible, to obtain for many jobs. This is particularly true of police and security officers who are often called upon to respond to unusual, diverse, and non-standardized situations, under conditions that are not readily measurable or observable by supervisors. As a consequence, it is difficult to validate selection tests for police and security officers. However, because selection tests are widely used for this job family, and are the subject of a high level of interest, controversy, and, in some jurisdictions, legal challenge, it is essential that appropriate validation take place. We propose to illustrate how a work sample can be used as a defensible intermediate criterion that can serve as a basis for validation.

An essential facet of job performance for a majority of police and security officer jobs is the ability to write detailed incident reports that describe the events encountered on the job. Incident reports are important because they serve as permanent records in investigations and court proceedings. For example, incident reports provide critical evidence in criminal cases for such offences as trespassing, theft, sabotage, and substance abuse. Well-constructed incident reports have credibility, while those that are incomplete or incoherent, or that contain grammatical or spelling errors, are an embarrassment and are more susceptible to legal challenge. To this end, a selection battery for police and security officers should contain predictors of their ability to write coherent incident reports (i.e. those prepared in the normal course of job duties). Unfortunately, validation research that uses real incident reports as criterion variables faces serious problems. Such field-produced incident reports would not be comparable across incidents, nor could they be readily evaluated for accuracy. Furthermore, officers are employed at many geographically diverse sites, and occasions requiring the preparation of incident reports on the job vary in their content and frequency.

A practical alternative to criterion measures based on real incident reports is the use of work samples. In the personnel selection domain, the use of work samples both as predictors of job performance and as an intermediate criterion for validating test batteries has had a long history (Hogan, Arneson and Petersons 1992; Newman, Howell and Cliff 1959). Probably their most extensive application has been in military jobs,

* Address for correspondence: Douglas N. Jackson, Department of Psychology, The University of Western Ontario, London, Ontario, N6A 5C2 Canada. e-mail: DJACKSON@JULIAN.UWO.CA

© Blackwell Publishers Ltd 2000, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

where work samples have been applied to evaluate training readiness and evaluation, and also have served as criterion measures for identifying the contributions of general and specific abilities (e.g. Ree, Carretta and Teachout 1995). When work samples are employed for selecting applicants already possessing the necessary training for the job, they yield the highest average predictive validity of any job selection procedure (Hunter and Hunter 1984, p. 91). In addition, a study by Newman *et al*. (1959) found inter-rater reliabilities of 0.90 or higher for work samples produced by applicants for dental positions.

We investigate in this study an alternative to an on-the-job performance measure as a criterion to validate a selection battery. A large security organization wished to strengthen its selection procedures by identifying job candidates capable of preparing incident reports that could serve as credible records in court proceedings. This study had three major aims. First, we sought to determine whether or not a job-relevant work sample for security officer incident reporting could be developed and scored reliably. Second, we wished to appraise the extent to which a cognitive test battery that was developed to be content valid for security officers would demonstrate concurrent validity in terms of a work sample. Third, we sought to evaluate whether or not a personality measure of dependability would demonstrate incremental validity over cognitive ability measures of the work sample criterion. Dependability was selected as a predictor based on the growing empirical evidence that so-called integrity (a subset of the dependability domain) is substantially linked to job performance (Ones, Viswesvaran and Schmidt 1993).

Participants

The sample completing the test battery consisted of 187 security officers, each with a minimum of six months' on-the-job experience and assigned to one of 24 sites of a large automobile manufacturer. None of the participants had been screened for employment using the measures employed in this study, or any other psychological tests. Accordingly, the sample did not suffer from restriction of range due to prior selection.

Materials

Cognitive ability measures

These measures were designed to support two goals: to capture cognitive skills essential for the preparation of incident reports, and to provide material with sufficient face validity to encourage user acceptability and legal defensibility. Two cognitive ability subtests were contained in the predictor battery. The first subtest, Situations, contained paragraphs describing various situations having relevance to the work of a security officer. Each situation was followed by a series of multiple choice questions requiring the respondent to identify from a set of alternatives the best response or course of action for that situation. For example, one situation concerned the proper course of action given certain company policies when an alarm had been activated. The second subtest, Language Usage, contained two parts, Spelling, a list of 54 words relevant to the work of a security officer, approximately half of which were misspelled, and Sentences, consisting of 12 items, each comprising a series of four sentences that might be used in an incident report, only one of which was grammatically correct. For Spelling, the task was to identify the correctly and incorrectly spelled words, while for Sentences the task was to identify the correct sentence. Extensive item analyses were undertaken separately for each subsection. Item statistics showed solid support for the test as a whole with only very minor adjustments necessary.

Dependability

A total of 70 questions were contained in the personality questionnaire. The items sampled three substantive facets of dependability, Responsibility, Low Risk Taking, and Integrity, each consisting of twenty items. The first two of these scales were adapted from the Jackson Personality Inventory – Revised (Jackson 1994), a standardized profile measure of personality that demonstrates exceptional psychometric properties. Scores from these scales have been shown by Ashton (1998) to be associated with a wide variety of counterproductive employee behavior, including, for example, theft, sabotage, and malingering among others. The third facet scale, Integrity, was developed and validated separately. These three facets yield a Dependability score.

Work-sample incident reports

With the aid of a professional artist, we devised two series of line drawings each depicting a problem that required the intervention of one or more security officers. The first series portrayed an injured employee who required first aid; the second set (Figure 1) showed a security officer discovering a fire and assisting an employee to safety.

The task of the respondent was to prepare for each of the two incidents an Incident Report in narrative form based on cues contained in the

© Blackwell Publishers Ltd 2000

Figure 1: Example of work sample stimulus materials.

drawings. Each report was rated independently by two judges each using two 7-point scales, one for the accuracy and completeness of content, and the other for style, including grammar, spelling, and sentence structure. After evaluating scoring for reliability, judges' ratings were aggregated.

Results

Reliability

Coefficient alpha estimates of reliability for the selection battery were: Situations, 0.90; Language Usage, 0.93; Spelling, 0.95; and Dependability, 0.90. The split-half reliability of the work sample obtained by correlating Incident 1 with Incident 2 was 0.88. The inter-rater reliability estimate was 0.87 (Table 1). These data were considered encouraging in light of the extensive literature in the education field indicating modest reliability for judging essay examinations (Coffman 1971).

Validity

Table 2 contains the concurrent validity coefficients separately for each of the cognitive tests and the personality measure, as well as the multiple correlation for the entire battery. These values are presented in uncorrected form and with a correction for unreliability in the criterion, assuming a perfectly reliable criterion from the vantage point both of judges and incident report item sampling. The validities for the three components of the cognitive ability test range from 0.43 to 0.53. The Dependability measure also showed a degree of association with the work-sample criterion, but yielded a negligible beta weight in the regression formula on which the multiple correlation was based. The latter was only marginally higher than the aggregation of the standardized cognitive ability scores,

Table 1: Reliability of work sample judgments

	Interrater	Split half (Incident 1 and 2)
Incident 1	0.87	–
Incident 2	0.83	–
Total	0.87	0.88

Note: Reliability estimates have been corrected by the Spearman-Brown formula.

© Blackwell Publishers Ltd 2000

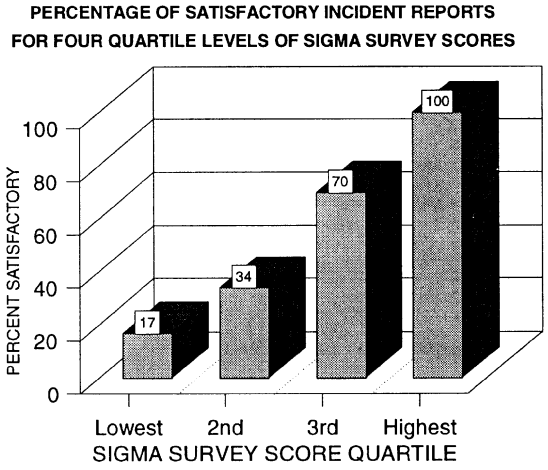

Figure 2: Percentage of satisfactory incident reports at four levels of predictor test battery scores.

indicating that the contribution of personality was largely due to its sharing variance with the cognitive ability measures.

A perusal of Figure 2, which contains the percentage of satisfactory reports produced by respondents in each of four quartiles of the predictor test score distribution, indicates that 100% of respondents scoring in the top quartile of the predictor test battery distribution produced satisfactory incident reports, while only 17% of those in the lowest quartile did so. If the groups are split at the median of the predictor test score distribution, 85% of the officers scoring in the upper half of the distribution produced satisfactory incident reports, while only 26% of respondents scoring in the lower half prepared satisfactory reports.

Discussion and Conclusions

As previously mentioned, the primary aim of this investigation was to examine whether or not a work sample could be developed and scored reliably. Results indicated that this can be accomplished – the work sample developed in the present study was both reliable and valid. The second purpose of this study concerned the validity of a cognitive battery with regard to an incident-report work sample. Analyses demonstrated that an easily scored cognitive battery shows substantial validity with regard to the incident report work sample. The use of such a standardized cognitive battery is more feasible and considerably less expensive for selection than would be a work sample measure that required expert judgement for scoring.

Finally, the third aim of the current study was to evaluate whether or not the Dependability measure demonstrated incremental validity over the cognitive battery. Although the Dependability measure was designed and validated as a personality scale to predict counterproductive behavior, it did show a significant zero-order correlation with the work sample criterion. This is useful information if a decision were to be made administratively to employ the Dependability scale without the cognitive component of the test battery. But when evaluated in terms of incremental validity, results indicate that Dependability adds little. This should not be regarded as evidence against the general validity of personality measures in job selection because this work sample criterion elicits mainly cognitive skills (cf. Johnson and Blinkhorn 1994; Jackson and Rothstein 1996). However, this finding does highlight the importance of evaluating the validity of personality measures in a context that also includes cognitive ability measures, if one's goal is to understand the basis for observed validities.

Table 2: Correlation of predictor tests with work sample criterion (N = 187)

Test	Validity	
	Uncorrected	Corrected
Situations	0.53	0.61
Spelling	0.43	0.49
Sentences	0.47	0.54
Total cognitive ability	0.57	0.65
Dependability	0.23	0.26
Multiple correlation (total cognitive + Dependability)	0.58	0.66

Note: Corrected validity coefficients are adjusted for unreliability in the criterion.

Our study highlights several advantages in employing work samples for validating employment tests, particularly work samples that yield a concrete result that can be evaluated independently. First, work samples can be administered under standardized conditions, involving identical stimuli and providing results that can be judged on a common scale. Second, work samples can be administered to all job candidates in the course of a validation study. This permits an estimation of validity over the entire range of the relevant population, rather than only on the selected job applicants, where restriction of range is inevitable when the test battery is used to make a selection decision. Third, because work samples can be judged apart from and independently of other information about the person, the well-known political and social biases (Longenecker, Sims and Gioia 1987), that are encountered in supervisory ratings of employees are avoided. Fourth, because results from the work sample can be

© Blackwell Publishers Ltd 2000

evaluated at some distance in time and place by expert judges, it is possible to evaluate the reliability of ratings of the performance, reliability that is likely to be higher than that obtained from independent field supervisory ratings. This permits the design of work samples to vary in the number of samples and judges so as to obtain the requisite level of reliability for obtaining stable estimates of validity (Borman and Hallam 1991). Finally, carefully selected work samples possess the content validity that makes them acceptable to job candidates (Robertson and Kanddola 1982), and convincing to decision makers.

For all of the above reasons, we consider it advantageous, where possible, to employ work samples in validational research. However, a work sample is not necessarily a sample of all important facets of the job. Obviously, in the case of police and security officers, job analyses would reveal other important facets. In the present study, participants are required to respond to a variety of emergency situations, including fire emergencies, as well as to render first aid. The validities demonstrated for the test battery with regard to the writing work sample might generalize to the ability to benefit from training programmes in fire fighting and first aid, although such a conclusion awaits additional validational data.

The results of this study demonstrate clearly that it is possible to develop reliable work samples of incident report preparation for security officers, that these work samples can be evaluated reliably by judges, and that a test battery designed to assess job-relevant problem-solving ability and language usage is a valid predictor of this important component of job performance. These findings invite the further development of work samples as an alternative to criterion measures based on supervisory ratings in the validation of test batteries.

References

Ashton, M.C. (1998) Personality and job performance: the importance of narrow traits. *Journal of Organizational Behavior*, **19**, 289–303.

Borman, W.C. and Hallam, G.L. (1991) Observation accuracy for assessors of work-sample performance: consistency across task and individual differences correlates. *Journal of Applied Psychology*, **76**, 11–18.

Coffman, W.E. (1971) Essay examinations. In R.L. Thorndike (ed.), *Educational Measurement* 2nd edn. Washington, DC: American Council on Education, 271–302.

Hogan, J.C., Arneson, S. and Petersons, A.V. (1992) Validation of physical ability tests for high-pressure cleaning occupations. *Journal of Business and Psychology*, **7**, 119–35.

Hunter, J.E. and Hunter, R. (1984) Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96**, 72–98.

Jackson, D.N. (1994) *Jackson Personality Inventory-Revised*. Port Huron, MI: Sigma Assessment Systems.

Jackson, D.N. and Rothstein, M.E. (1996) The circumnavigation of personality. *International Journal of Selection and Assessment*, **4**, 159–63.

Johnson, C. and Blinkhorn, S. (1994) Desperate measures. *The Psychologist: Bulletin of the British Psychological Association*, **4**, 167–70.

Longenecker, C.O., Sims, H.P. and Gioia, D.A. (1987) Behind the mask: the politics of employee appraisal. *Academy of Management Executive*, **1**, 183–93.

Newman, S.H., Howell, M.A. and Cliff, N. (1959) The analysis and prediction of a practical examination in dentistry. *Educational and Psychological Measurement*, **19**, 557–68.

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

Ree, M.J. Carretta, T.R. and Teachout, M.S. (1995) Role of ability and prior knowledge in complex training performance. *Journal of Applied Psychology*, **80**, 721–30.

Robertson, I.T. and Kanddola, R.S. (1982) Work sample tests: validity, adverse impact and applicant reaction. *Journal of Occupational Psychology*, **55**, 171–83.

© Blackwell Publishers Ltd 2000