Work Sample Testing

Militza Callinan and Ivan T. Robertson*

The general evidence supporting the use of work sample tests (good criterion-related validity, positive applicant reaction and job preview capability) is reviewed, limitations are also noted. The extent to which work sample tests may be used to limit adverse impact in selection is reviewed and the extent to which work sample tests may be used to complement other selection instruments, in particular, cognitive ability tests, is discussed. Work sample tests are based on attempts to embody key tasks within the job in a selection instrument. Five distinctive (though not mutually exclusive) types of work sample tests are identified and validity evidence is reviewed. Key dimensions (bandwidth, fidelity, task specificity, necessary experience, type of tasks and mode of presentation and response) are identified and used to compare different types of work sample tests.

Introduction

Work sample testing does not describe a single method or procedure. Rather, it is an approach or rationale for the assessment of individuals’ current or likely future job performance. A range of assessment measures, sometimes very different from each other in basic ways, have been categorised as work sample tests. The most literal definition, however, would be a hands-on performance test in which a job applicant or employee is required to actually perform a job-related task under the same conditions as those required on the job. Very simply, using this definition as a starting point, measures that are classed under the heading of work samples can be organised according to the degree by which they are removed from the two features of actual hands-on performance and a real work setting.

Most often, work samples have been used as predictor measures for the purpose of personnel selection. Although they are also used as criterion measures to validate training outcomes, performance ratings and predictor measures used in selection. Military applications have provided a proportion of the most recent research in this area (Campbell, Ford, Rumsey, Pulakos, Borman, Felker, De Vera and Riegelhaupt 1990; Carey 1991; Hedge and Teachout 1992).

Positive Aspects of Work Sample Tests

As predictors, when compared with other selection methods, some types of work sample tests have demonstrated higher predictive validity than even general mental ability. Schmidt and Hunter (1998) reviewed all the meta-analytic evidence available for the validity of 19 selection procedures in predicting job performance. The highest reported validity for an individual method was for work sample tests at .54 (corrected for error in the criterion measure). The next best single predictor was general mental ability with a validity of .51 (corrected for range restriction and measurement error in criterion).

At the same time as delivering higher validity, the tests seem also to have substantially less adverse impact against minority groups. Adverse impact is evidenced in a disproportionate rejection rate of one sub-group by comparison with the rest of the individuals being assessed for selection. Whereas cognitive ability measures routinely show one standard deviation difference in scores, in favour of White test-takers (Schmitt, Clause and Pulakos 1996), evidence suggests this is substantially reduced, if not eliminated with hands-on performance tests.

Schmitt et al. (1996) examined available data for the adverse impact associated with different selection methods. For work sample tests, across 37 instances and a total sample size of 15,738, the average effect was −.38 standard deviations (range .16 to −1.07) when African-American and White applicants performance scores were compared. A comparison of Hispanic-American and White candidates work sample performance (20 instances with a total sample size of 7,848) indicated that there was, on average, no difference between these groups.

Two further things add to the case for using such tests: a positive applicant reaction and the fact that they provide a preview of the job itself. Applicants see work sample tests as job-related and hence fair (Steiner and Gilliland 1996). These perceptions may be related to the incidence of legal challenges. Terpstra, Mohamed and Kethley (1999) found in their analysis of federal court cases...
that challenges to work sample tests were underrepresented when the level of usage by selectors was taken into account. Furthermore, their use was successfully defended in six of the seven cases recorded. Four dimensions have been found to form the basis for applicants’ fairness reactions to a selection process – perceived job relatedness, propriety of questions and interpersonal treatment (Gilliland 1993). It can be seen that three of these four dimensions are potentially present in a work sample procedure. Applicants’ perceptions of fairness have been shown to be related to organizationally important outcomes such as employee recommendations, organizational commitment and work performance (Gilliland 1995).

Because work sample tests reflect aspects of the actual job, they function as a realistic job preview for candidates as well as a selection tool. In Downs, Farr and Colbeck’s (1978) research, performance on a work sample was related to self-selection by candidates. All those who completed the test were invited to start work regardless of their score. The actual starting rates were related to performance on the test with progressively fewer people accepting the job as their rated performance scores on the work sample got lower. It seems that applicants made decisions about their own suitability for the job from the testing experience.

Limitations of Work Sample Tests

Despite possessing such positive characteristics as a selection method, however, work sample testing does not challenge the status of general mental ability as the most expedient and valid predictor of performance across jobs and people. In fact, Schmidt and Hunter (1998) argue that all other selection methods should be viewed as supplements to cognitive ability measures. This is because, despite their relatively high validity, the use of work sample tests is limited in important ways.

In many instances, the tests are not suitable for assessing applicants without job experience. Development and administration can also be complex and costly. By definition, a work sample needs to be individually developed and validated for each job or group of jobs. The test needs to be carried out either in the actual workplace or in a specially constructed simulated context. Individual administration is also likely to be required which is time-consuming and possibly cumbersome to arrange. Meeting these requirements may be precluded by logistical, financial and safety considerations.

Finally, although predictive validity is high at the point of selection, validity does seem to attenuate to a greater extent than other selection measures across time (Robertson and Kandola 1982). In Siegel and Bergman’s (1975) study, the predictive validity of a work sample was compared with standard pencil-and-paper tests for US Navy recruits admission into a machinist’s mate school. At 9 months, 5 of 7 performance criteria were predicted better by the work samples. This situation was reversed at the 18-month follow-up, however, with all but one of the criteria predicted by the standard tests but not at all by the work samples. The finding may be related to the specificity of skills that work sample tests sometimes measure. Continuous on-the-job learning and adaptation needed for successful performance over time may be more related to general underlying abilities. This issue is discussed later.

Nonetheless, there is a drive by selectors to develop predictors with equivalent or higher validity than general mental ability but without the associated level of adverse impact against minority groups. So far, for cognitive ability measures, there is no evidence for differential validity between groups. That is, differences in cognitive predictor scores at selection do relate to differences in job performance and so are not unfair according to the widely accepted Cleary (1968) definition.

According to Cleary (1968), an assessment measure is fair towards members of a particular group for selection purposes if their scores predict job performance equally as well as the scores of members of the majority group. However, looking more broadly, there are sound reasons to strive for workforce diversity and hence the development of selection procedures that do not preclude this outcome. The business benefits for organisations that achieve employee diversity have been well articulated by Kandola and Fullerton (1994) but a pure sense of social responsibility can still be argued to be a sufficient driver of new approaches to selection.

Cognitive ability is the best general, catch-all predictor but it represents only one small part of the domain of job performance. There is still a substantial portion of variance in performance between people left unexplained. More recently, it has been argued that more is not necessarily better for all jobs (Landy 2000). A certain amount of general intelligence is required in a job and once this is satisfied, other factors become more important for success. These might include personality characteristics, practical knowledge, self-discipline and social competence. Measures of these other factors may be less likely to show adverse impact against social and ethnic groups (Schmitt et al. 1996).

By expanding the range of factors measured at selection from cognitive ability alone, the outcome is two-fold. More of the predictor space is covered, which may bring higher predictive...
What do Work Sample Tests Measure?

The extent to which cognitive ability underlies individuals’ performance on work sample tests will vary according to the type of measure. Overall, however, Schmidt and Hunter (1998) report that work samples show incremental predictive validity of .12 over general mental ability. This represents a 24% increase in validity by using a work sample in addition to a cognitive ability measure for selection purposes. The evidence demonstrates that although correlated with ability (a correlation of .38 is reported by Schmidt and Hunter [1998]) work samples tap performance-related factors unrelated to general intelligence and so do usefully expand the amount of predictor domain covered.

It is perhaps unsurprising that work experience is also correlated with work sample performance. Quinones, Ford and Teachout (1995) investigated the relationship between work experience and job performance. Experience was most highly correlated with ‘hard’ work sample measures of job performance, compared to ‘soft’ measures such as ratings, with a correlation of .39 after correcting for sampling error and criterion unreliability. The main impact of experience on performance may be indirect, being delivered through the acquisition of job knowledge. For example, the impact of job experience on job knowledge has been found to be stronger than its impact on work sample performance (Schmidt, Hunter and Outerbridge 1986).

Although experience did also have a direct effect on performance in the same study, suggesting that an individual’s experience also affects their work sample performance above and beyond the task-specific knowledge they have accrued as a result of being on the job. One possible explanation for this finding is that to accomplish a task or set of tasks an individual will employ particular techniques or work methods, in addition to the task-specific skills required, in order to get the job done. These work methods represent ‘realistic work habits used to solve specific problems’ and may be a further source of differential performance between people (Smith 1991).

A concept related to work methods is that of ‘tacit knowledge’ (Sternberg 1996). Tacit knowledge is practical knowledge about what to do and when in order to succeed in an endeavour, it is not formally taught and often not verbalised. These work habits and/or practical knowledge may be related to actual on the job performance. Sternberg (1996) reports correlations of .3 to .4 between tacit knowledge measures and job performance. Individual assessments that focus on specific skills or abilities outside the context of a realistic job task do not allow these aspects of performance to be measured.

That work sample tests allow individual’s practical work habits to be assessed is one of a number of explanations suggested for the high validity of the measures. It is also thought that they allow the interaction of individual skills and aptitudes to occur naturally, leading to higher validities than when component aptitudes are measured independently and combined by the selector to reach a decision (Smith 1991). As already discussed, work samples may simply capture more of the predictor space than other measures because of their inherently multi-faceted nature (Schmitt et al. 1996).

A motivational explanation has also been put forward for the superior predictiveness of work samples. Job applicants who want the job on offer may be motivated to engage in the test and perform well because the face validity of the task leads to the belief that this is a fair opportunity to show the selector what they can do (Smith 1991; Robertson and Kandola 1982). As such, the measures will highlight the most motivated applicants for the job. It is also for this reason that in the motivated context of selection, work samples can be said to represent tests of maximal, rather than typical performance.

A number of attempts have been made to develop performance assessment measures that retain the desired characteristics of hands-on work sample tests but ease the practical barriers to their development and use. That is, to create measures with the same underlying principles as work samples that do not require the actual performance of the task in the real situation, often the source of practical difficulties. It is this approach that is most common in the more recently published literature on work sample testing.
In reality, what work samples measure and the reason that they work as they do is not actually well understood. It does seem to be that the tests work partly because they mediate the impact of job performance-related factors such as cognitive ability, work experience, job knowledge and motivation. Although the relative contribution of each factor will differ according to the particular type of work sample measure concerned. This complexity and an absence of conclusive research mean that distilling only the test features related to the desired outcomes into a more amenable package is not straightforward.

What is a Work Sample Test?

The rationale underlying work sample testing is that higher predictive validity is achieved when there is ‘point-to-point correspondence’ (Asher and Sciarino 1974) between predictor and criterion measures. Practically speaking, the content of the selection measure on which the hiring decision is made should mirror the content of the job in question by focusing on some or all of the behaviours required for successful overall performance.

This approach is in contrast to the use of cognitive ability and personality measures, for instance, that focus on predispositions to behave in certain ways. It represents the use of samples of behaviour rather than signs or indicators of behaviour. The work sample approach is predicated on people displaying behavioural consistency (Wernimont and Campbell 1968). That is, the best predictor of a person’s future performance is his or her past performance. It is thought that the inferential leap that an assessor must make using this strategy is much reduced. With predispositions, the assessor must decide, on the basis of knowledge about an individual’s latent potential, what is the probability that he or she will realize that potential in performance of specific activities, to the desired level, on the job in question. With evidence of an individual’s actual level of performance on a same or similar activity to that required in the job, the degree of inference needed is much reduced.

In reality, determining what is a sufficient degree of point-to-point correspondence between selection test and performance criteria is not straightforward for anything but the most direct hands-on performance test for a relatively uncomplicated job (Robertson and Kandola 1982). It is not always the case that correspondence between predictor and criterion is the same thing as the job-relatedness of the test.

Robertson and Kandola (1982) reviewed some cases in which work sample tests were used as both a performance predictor and performance criterion in the same study. In such cases the criterion measure is usually longer or more comprehensive than the selection measure. The authors suggested that high validities may be being sought by constructing the criterion measure to reflect the predictor measure and not with full consideration that the criterion reflect the job performance domain. In this case, point-to-point correspondence is created with the appearance of validity but performance on the criterion measure may not be strongly associated with actual job performance. The correlation between the two measures in such a case is an indicator of reliability, not validity.

Job-relatedness, or content validity, is a critical issue in relation to the construction of work sample tests. It is a prerequisite for point-to-point correspondence between a selection test and the job and therefore (theoretically at least) determines the level of criterion-related validity.

Applicant perceptions of fairness are based in part on the perceived relation of the selection procedure to the job for which they are applying. There is also evidence that race-linked bias is higher when job-irrelevant behaviours are being rated compared to job-relevant behaviours (Brugnoli, Campion and Basen 1979). Thus, job-relatedness may also be a determinant of reduced adverse impact. Care must be taken to ensure that face validity is not mistaken for real content validity. This argument has been levelled at some assessment centre exercises such as leaderless group discussions (Lowry 1997) and in-basket measures (Schippman, Prien and Katz 1990).

In addition, a work sample may be job related but may not be comprehensive enough in the range of tasks or behaviours it entails to fully reflect the actual job role. This lack of representativeness imposes an immediate limitation on the potential criterion-related validity of the measure. Assuming that it is not possible for a work sample test to include the whole job, some part of the job must be chosen for assessment. For a valid measure, the tasks chosen should be representative, as is a random sample of the whole domain. Or, the sample of tasks should have been determined to be those that are most critical to overall performance, as in purposive sampling (Green and Wigdor 1991; Campbell et al. 1990).

As a general principle, the interpretation of a performance measure depends on the degree to which it is representative of work required in the job (Green and Wigdor 1991). This notion is elaborated by the Military Committee for Performance Standards as the concept of domain-referencing of predictors. Using this approach, scores on a test are referenced against levels of job mastery, not against those of other test-takers. It then becomes possible to make predictions about an individual’s likely level of
future job performance on the basis of a test score, as opposed to relative competence compared to other job candidates. For example, a candidate who can perform 70% of the tasks on a selection test is likely to be able to perform 70% of the actual job. Well-constructed work sample tests should allow this level of inference and interpretability.

In short, work sample testing is a job performance assessment strategy based on the principle that to maximise validity and fairness a predictor or criterion measure should measure behaviours on tasks and activities within the job performance domain that are known or can be expected to be substantially related to actual overall performance in the job.

Scoring Procedures

A variety of scoring procedures has been adopted for work sample tests including global ratings, behaviour checklists, error scores and specific behaviour ratings (see, for example, Robertson and Downs 1989). There is no research available which provides a comparison of the advantages and disadvantages of different approaches. Although it is clear that, whatever the method of scoring is used, normal psychometric criteria of reliability, discrimination, fairness, administrative convenience (Smith and Robertson 1993) will need to be satisfied.

The use to which the scores will be put might also help to determine which scoring methods are most suitable. For example, if detailed feedback for development and training is needed a global rating will be unhelpful, whereas in a selection context it may be sufficient.

Definitions of Work Sample Tests

The procedures categorized as work sample tests do not make up a homogenous group. The situations are more or less parallel and realistic, the tasks more or less relevant to a specific job and the applicant may be asked to verbalize how they would perform a task rather than provide an actual demonstration. Existing definitions vary in their exclusivity. Robertson and Kandola (1982) define work samples as involving ‘the applicant in performing a task or set of tasks which are thought (usually on the basis of job analysis) to have direct central relevance to the job in question’. They include job knowledge tests, normally pencil-and-paper measures, on the basis that the knowledge measured is of relevance to the job and they do not specify a particular context for the assessment.

In contrast, Smith (1991) states that work samples ‘measure job skills by requiring an individual to demonstrate competency in a situation parallel to that at work, under realistic and standardized conditions. Their primary purpose is to evaluate what one can do rather than what one knows’. This definition clearly excludes measures of job knowledge and even simulations that are not conducted in ‘realistic’ circumstances which might include some assessment centre exercises.

An inclusive stance will be taken here. All measures and procedures that have been associated with work sampling are discussed under the headings commonly employed by researchers and selectors. The headings are not discrete. Measures under separate headings may share some characteristics whilst those under the same headings may have differences. In an attempt to evaluate these relative similarities and differences some possible dimensions by which the individual examples can be distinguished are considered later.

Types of Work Sample Testing

Hands-on performance tests are the most straightforward and literal example of work sample tests. Psychomotor tests, by definition hands-on tests, had a median validity of .39 in the review by Robertson and Kandola (1982). Psychomotor tests developed for the US Marine Corps provide an illustrative example (Carey 1991). The Job Performance Measurement Project is a US joint service initiative to ‘examine the feasibility of measuring job performance and to link enlistment standards to job performance’ (Green and Wigdor 1991). Within this project hands-on performance tests (HOPTs) are ‘the benchmark measure of job performance’ (Carey 1991) for selection and diagnosis of training needs.

In this example, a rigorous process of job analysis was used to produce a list of critical tasks for each US Marine Corps duty area with a score sheet for each task broken down into the necessary steps for completion. Examples of the types of tasks assessed from different duty areas are ‘Prepare launcher for firing’, ‘Assemble and operate radio’ and ‘Treat sucking chest wound’ (Carey 1991). The HOPTs were trial tested and retired marines recruited as test administrators. Detailed training manuals were prepared and training conducted across all testing locations. Daily quality control checks were made on scoring accuracy and administrators rotated across locations to minimize error.

Of course, hands-on tests need not cover only psychomotor tasks. Work sample tests were developed as criteria to validate selection tests for telephone service representatives by Gael and Grant (1972). The employees had to deal with a number of typical telephone calls and the resulting administration such as record prepara-
tion and filing. In another study (Gael et al. 1975) telephone operators had to deal with a stream of incoming calls for one hour. Performance on each of the composite activities was assessed as being correct, incorrect or not performed at all, with a rating given for the overall effectiveness of each call.

One main point about these tests is that they can only be used to assess individuals who have experience on the same or a similar job. The military example of developing HOPTs described above, although admittedly on a large scale, also demonstrates the number of detailed steps and resources required for sound construction of a hands-on work sample measure. It is not surprising then, as mentioned already, that work in this area seeks to find alternatives to fully blown HOPTs. Indeed, Carey’s (1991) research was aimed at seeking suitable surrogate measures for the US Marine tests. As well as the expense of development and administration of HOPTs, in the armed services this is also to avoid the unnecessary use of resources such as ammunition, scarce equipment such as aircraft and danger to personnel and equipment.

Carey (1991) assessed a number of possible surrogates including field proficiency and conduct ratings, training scores, supervisor ratings, video-simulated marksmanship and pencil-and-paper job knowledge tests. Only the job knowledge test was found to be equivalent to the HOPT in that the same decisions would be made using either test. This was only the case for selection purposes, however, not for diagnosis of training needs. In this case then, there may be an argument that measuring actual performance is simply not necessary to achieve good validity.

Trainability Tests

Job learning or trainability tests are work samples that include a structured and controlled period of learning prior to the actual task performance. Unlike standard work sample measures, such tests allow the assessment of the potential of candidates with no experience or prior training to benefit from training once hired. Robertson and Downs (1989) carried out a meta-analytic review of the predictive validity of trainability tests. The results are not straightforward due to differences in the predictor and criterion measures used in the individual studies and also by the follow-up period employed. The highest validity of .48 was found when the number of errors during test performance was the predictor and training performance was the criterion measure. When job performance was the criterion, validities were significantly lower at .20 to .24. Although the result is confounded by the fact that follow-up periods were longer when job performance was considered and as for standard work samples, the predictive validity of trainability tests did attenuate over time.

A different but related type of assessment is the minicourse approach. The required training for some jobs is lengthy and expensive. Hence an organization will want to avoid investing in individuals who do not have the ability to complete the course to the required level and select those with the greatest aptitude for training who will go on to function effectively on the job. In such cases, a candidate’s performance on a sample of training content, analogous to a sample of job content in a work sample, is used to predict their likely success in the full-scale course.

This was the case in the examples described by Reilly and Israelski (1988) of minicourses they developed for AT&T where training for telecommunications jobs such as minicomputer maintenance lasted one year. In addition to the cost of training, the arrival of new technologies meant that the job roles in question were completely new and other selection methods based on job analyses could not be used. The seven minicourses described ranged in duration from 2 hours to 3 days for target jobs such as electronic switching systems technicians and facilities assigners. Each minicourse consisted of a number of modules followed by a test to measure learning of the material. Some criterion-related validity studies were carried out on these tests. The criteria examined included time to complete training, training performance and job performance for which the average validity was .62. In the one case where actual job performance was the criterion, a validity coefficient of .39 was reported. In addition to the good predictive validity of the tests, the authors also note broader benefits to the company. The few labour arbitration hearings that have been conducted in relation to the minicourses have upheld their use for decision-making. Also, the company observed direct savings related to decreased failure rates and shorter training times.

The same principles of relatedness and representativeness discussed for standard work samples need to be applied to trainability tests and minicourses to achieve good validity. Robertson and Downs (1989) suggest that the validity of trainability tests will depend on good sampling not only of the job content domain but also of the learning context that occurs on the job. This process is more straightforward for minicourses in that the context in which full-scale training takes place is likely to be clearly defined, unlike the informal learning situations that occur in the workplace.
Situational Tests

Situational tests are clear products of researchers' attempts to develop measures with work sample logic and desirable characteristics but without the practical difficulties. A standard work sample requires the testee to actually perform the task at hand whereas these tests require them to describe how they believe they would behave if he or she were faced with a given situation. Respondents are presented with scenarios, or situations akin to those that might be faced on the job that require some decision or action to be taken. Required responses might be verbal or written, or some combination of the two and are commonly compared to predetermined scoring keys developed from the responses of experienced and/or successful job role-holders or experts. Usually the format is a pencil-and-paper questionnaire. Although technology has had an impact on situational testing with video and computer-based applications being developed (Weekley and Jones 1997; Burke 2000).

Some researchers argue that situational tests are merely job knowledge tests (e.g. Schmidt 1994). This is because correlations between situational tests and cognitive ability, which is mediated by job knowledge (Wagner 1997; Schmidt, Hunter and Outerbridge 1986), are often large. In McDaniel, Finnegan, Morgeson, Campion and Braverman’s (1997) meta-analysis involving 95 studies, the tests were estimated to correlate .53 with general cognitive ability. The criterion-related validity of situational tests was estimated to be .56 in the same meta-analysis. Others argue, as with other types of work sample measures, that situational judgement measures are multi-faceted measures, tapping several abilities and skills (Chan and Schmitt 1997).

The ‘interview’ approach taken by Hedge and Teachout (1992) is an integration of standard work sample testing and situational testing. These researchers developed both hands-on work sample tests and interview-based performance tests for a subset of a representative set of tasks from jobs in the Air Force. In the interview approach the candidate is asked to describe in detail how they would perform a task and he or she is assessed in the same way as hands-on testing, according to whether each necessary step has been carried out, or described, satisfactorily. As with Carey’s search for surrogate measures, Hedge and Teachout (1992) found that the two types of measures were equivalent for selection purposes, in that they led to the same rank ordering of individuals. But they were not equivalent for training needs assessment because the task-level correlations between the hands-on and interview formats were highly variable within and between job specialties. The authors suggest that this kind of testing not only allows the ‘rigour of work sample measurement’ with the benefit of ‘measurement efficiency’ but affords sound assessment of tasks that are not suitable for hands-on testing.

The ‘low fidelity simulation’ developed by Motowidlo, Dunette and Carter (1990) for selecting telecommunications managers exemplifies paper-and-pencil situational testing. Applicants were presented with written descriptions of work situations designed to assess problem-solving and interpersonal behaviour and required to choose one response that they would most likely make and one they would least likely make from five alternatives. The test correlated .28 with supervisory ratings of performance and showed no significant differences between racial groups.

The results of Motowidlo et al. (1990) are more positive than those reported by other researchers. Stevens and Campion (1999), for instance, created a pencil-and-paper test of teamwork situations. The test did show good correlations with teamwork performance ratings of .32 and .37 with overall performance, but was discovered to be a ‘highly redundant measure of general mental abilities’. The correlation between the Teamwork Test and the aptitude tests used was estimated to be between .91 and .99.

Weekley and Jones (1999) carried out two large validation studies of situational judgment tests for retail and hotel staff. Test scores had a validity of only .19 for predicting performance ratings and, although adverse impact against non-White groups was lower than that for cognitive ability it was still significant. In fact, the same authors (1997) also developed video-based situational tests and warned that they are not low adverse impact measures with differences between White and non-White candidates of half a standard deviation in favour of the former.

Video and computer-based situational testing offers the potential for creating more realistic simulations of work contexts than afforded by written tests, but with more practical convenience than live simulations. Video clips of social interaction sequences, for example, are more akin to the way such interactions are experienced on the job than written descriptions. Computer-based assessments also provide new opportunities such as the investigation of real-time decision-making where this is critical to performance. Burke (2000) has designed a computer-based assessment tool for air traffic controllers, for instance.

The availability of alternative modes of presentation and response for performance testing also potentially allows for the non-job related cognitive load of assessments to be reduced. This has implications for improved
content validity of tests and also for fairness and adverse impact. As discussed previously, it is the general cognitive ability component of measures that delivers the greatest sub-group differences. Where there is little demand for evaluations of written descriptions on the job, pencil-and-paper situational tests may be introducing a confounding factor by using this mode of presentation (Schmidt et al. 1996). Saying that, Weekley and Jones’s (1997) research on video-based tests did not indicate that the medium of presentation alone impacts on race effects of tests.

**Job Knowledge Tests**

Job knowledge tests are most usually pencil-and-paper measures of the amount of information the candidate has about the job and have been shown to be good predictors of work performance. Robertson and Kandola (1982) reported a median validity of .4 for predicting job performance for ‘job-related information’ tests. More recently, Schmidt and Hunter (1998) calculated the predictive validity of the tests at .48. Job knowledge is perhaps the greatest departure from the notion of work sample testing in that ‘knowing’ is clearly not the same as ‘doing’. However, it is only one step removed from the situational tests described above.

Maybe the biggest difficulty in accepting job knowledge under the heading of the work sampling approach is that individuals’ scores on the measures are often highly related to scores of cognitive ability. It may be through job knowledge that cognitive ability has its impact on job performance. (Schmidt and Hunter 1998; Weekley and Jones 1999). On the other hand, Carey (1991) did find that the job knowledge test for US Marines was a suitable surrogate for HOPTs in personnel selection. In this case, however, the written tests were developed to parallel hands-on content to a high degree, likely not to be the case for many other job knowledge measures.

**Assessment Centre Exercises**

Job-related simulations have been described as the ‘hallmark’ of assessment centres (ACs) (Howard 1997). The overall predictive validity of ACs is reported at .37 by Schmidt and Hunter (1998). The validity associated with the individual components, or exercises that make up a particular AC are, however, rarely reported. Particular types of work sample exercises are commonly associated with the AC method, namely in-baskets, leaderless group discussions and role-based group discussions.

There have been some evaluations of these individual work sample exercises. Robertson and Kandola (1982) examined group discussions as a separate category of work samples and reported a median validity of .34 for predicting job performance. Schippman, Prien and Katz (1990) examined the psychometric properties of in-basket measures which are used in 95% of ACs (Thornton and Byham 1982). The measures are simulations ‘using the materials which are typically found in the in-basket of a target (typically managerial) position . . . letters, memos, records, or other items which initiate a performance demand requirement and structure the response’ (Schippman et al. 1990). The authors did find many significant correlations between the measures and work-related criteria. However, there was a large degree of variation in the nature of the measures and the procedures used in individual studies making any overall judgment of the validity of in-basket measures difficult.

In fact, it was the ‘paucity’ of information about methods, procedures, reliability and validity data that was highlighted as the most outstanding finding of the research given the widespread confidence in the efficacy of these exercises. Schippman et al. were also unconvinced about the content validity of the majority of in-basket measures suggesting that the concept was being confused with face validity. In their view, statements of content validity were simply not supported by data and the measures ‘fall seriously short of the mark’. Those measures constructed for particular jobs had higher validities than generic off-the-shelf measures although the use of the latter are far more common. The reason for an apparent disregard for job-relatedness in simulations used in ACs may, though, be driven by differences in the underlying rationale for their use in this context and that for ‘stand alone’ work sample measures.

The objectives adopted for using work sample tests within ACs and the way in which the performance information is used in the scoring process are sources of distinction from other applications of work sample testing. In fact, there are obvious conflicts between the underlying theoretical basis for using work samples and that guiding the practice of ACs. First, ACs are frequently aimed at predicting future potential to perform at a certain level (i.e. managerial) rather than performance on a specific job. As a result, the exercises are chosen to involve generalised activities reflecting the core part of most jobs in that class or at that level. Immediately, the point-to-point correspondence issue becomes somewhat problematic. One of the founders of the AC method, William Byham, has suggested, in fact, that assessors should avoid making the tasks too similar to the job because raters get distracted by the detail of the task instead of
assessing the underlying competencies of interest (Mayes 1997).

This leads to the second, related point, the subject of performance ratings on the exercises. Byham is referring to the fact that within ACs, candidates are rated according to whether they elicit behaviours indicating basic competencies or dimensions, like decision-making, leadership and communication, believed to underlie successful performance across all exercises. A traditional work sample approach is quite contrary to this. The basic rationale is to obtain ratings of behavioural performance through observation and to avoid inferences about the links between underlying abilities and performance-related behaviour. It may be, though, the holistic appraisal of interacting skills, abilities and experience on a complex task, rather than the combination by the selector of independently measured individual dimensions, that actually leads to the high validity found in work sample tests (Smith 1991). If that is the case, the AC method is paradoxically employing individual exercises because of their superiority as predictors and stripping them of the component by which that validity is delivered.

This discussion touches on one of the most contentious current issues in AC research. Assessor’s ratings consistently cluster according to exercises rather than dimensions as intended (Robertson, Grattan and Sharpley 1987; Schneider and Schmidt 1992; Howard 1997), suggesting that dimensions lack discriminant validity. This is used as evidence to argue that AC exercises are job samples and there is no reason to ignore overall task performance in favour of less valid inferences about underlying abilities. The necessary information is already available. A full treatment of the value of ACs themselves is beyond the scope of this article and the argument is stated simply. The main point in relation to work sample testing is that in the AC context, simulations are used differently and contrary to the theoretical rationale. The exercises are ‘forums’ (Lowry 1997) for the expression of particular behaviours sought by the assessor rather than work samples in the literal sense.

One final point is that prevailing advice about maximizing the construct validity of ACs suggests reducing the range of different types of behaviours elicited by exercises to those relevant to dimensions of interest in order to facilitate rating (Lieverens 1999). Again, this may be contrary to the notion of allowing a natural interaction of behaviours to occur within activities that would occur on the job and may reduce the degree to which the job domain can be adequately sampled. The construction of exercises in this instance is led by the integrity of the AC process rather than by the goal of realistically recreating the job situation.

**Some Dimensions for Comparing Work Sample Measures**

Looking at the range of measures that are based on the work sample rationale, the large variability between them is obvious. In order to allow the measures to be considered relative to each other, it is useful to consider some dimensions on which they differ. The dimensions also highlight some of the decisions that selectors must make in constructing and using a specific test.

**Bandwidth**

Bandwidth is the degree to which the entire job performance domain is represented by the tasks that make up the measure. The importance of representativeness for validity has already been discussed. Systematic job analysis procedures form the basis for deriving any sample of the job, whether the method is random or purposive. In the case of AC exercises, the job domain is sampled through the combination of individual exercises. In other cases, such as minicourses, the measure is intended to capture all the key aspects of the whole job or, in this case training course. Practical factors such as the assessment time allowed per candidate will, of course, enter into the decision about how much to measure. The key point, however, is that what is measured should be job-related and even if limited be representative of the most critical aspects of the job.

**Fidelity**

The measures described differ in the extent to which the assessment task and context mirrors those actually present on the job. Hands-on performance tests requiring the assessees to use equipment in the workplace on which the job will be carried out would be at one end of the continuum. A written description of a work situation administered in a ‘classroom’ situation which requires a written response would be at the other end of the continuum. At various points in between these two extremes are simulations with varying degrees of realism. According to the principle of point-to-point correspondence, the more realistically the measure represents the job role, the higher the potential predictive validity.

Motowidlo et al. (1990) argue that there is currently little understanding about whether the extra validity gained from high fidelity simulations justifies the time and resources needed to construct them. It is not known at what distance from the actual job in terms of the context, presentation of information and response mode of the measure that validity is substantially reduced. Low fidelity measures have shown
good relationships to job performance. Although the other benefits of work samples such as applicant reactions, realistic job previews and lower adverse impact may not necessarily transfer to lower fidelity methods. Research is currently not much help to selectors in providing data or methods by which the investment-validity trade-off can be calculated taking into account all of these factors. Hoffman and Thornton (1997) demonstrate that established methods of calculating the practical utility of selection methods are inadequate when adverse impact is a critical consideration. When cut scores were set to eliminate adverse impact, the practical utility of an AC was higher than a cognitive ability measure despite being more costly. The authors argue that organizations may need to balance more than cost and selection ratio when choosing a selection system.

Specificity of Tasks

This dimension is likely to be related to fidelity, in that the more contextualised the task, the more situation- or job-specific the skills required of the candidate. Hands-on performance tests target specific behaviours relevant to a particular job. Assessment centre exercises are more likely to target behaviours that generalise between jobs at the same level, as with managers.

It has been noted that work sample validities seem to attenuate over time, whereas basic abilities such as cognitive functioning are more durable predictors of long-term job performance (Robertson and Kandola 1982). Where job roles are unlikely to change and short-term performance is key, selecting for highly specific skills and abilities makes sense. The issue for the selecting organization may be in deciding on the degree of flexibility required within the workforce. If employees may be required to move between job roles and/or adapt to new requirements in their current roles through ongoing learning, assessing for basic underlying abilities may be a better long-term option.

Of course, that may mean sacrificing some validity at the point of selection, assuming that it is the point-to-point correspondence of work sample tests that brings better prediction in the shorter term. Targeting less specific behaviours for assessment does not mean that work sample measures cannot be used, however. Motowidlo et al. (1990) constructed their low fidelity simulation around the two areas of competence that were common to job roles across a number of functional areas, problem-solving and interpersonal skills, which they determined with job analysis. The arguments discussed previously that have been levelled at work samples used in the context of assessment centres, is that some of the exercises do not represent aspects of any job for which the candidate is being selected, rather than simply the generality of the skills measured.

Necessary Experience

Hands-on performance and job knowledge tests clearly require the candidate to be experienced, or at least trained for the job in question before undergoing the assessment. Trainability tests, on the other hand, assume the test-taker has no experience. In the case of situational tests, the position is not as straightforward. The experience requirement is more ambiguous hence the suitability of a measure for an applicant population of mixed experience may not be easily determined.

Evidence for the impact of experience on work sample performance generally was discussed earlier in this paper. More specifically, in relation to experience-ambiguous measures, Weekley and Jones (1999) report a correlation of .23 between scores on a situational judgment measure and a broad experience scale. In the case of tacit knowledge measures it has been found that college students with no experience can get ‘respectable scores’ (Sternberg 1996). Experience is a notoriously difficult concept to measure and findings are not easily compared due to variation in the measures used between studies (Tesluk and Jacobs 1998). For selectors, the findings do suggest that the effect of prior experience on performance in these measures cannot easily be quantified.

Type of Tasks

An obvious difference between the measures described is that between the types of tasks and behaviours they are designed to measure, such as psychomotor, verbal and social tasks. The broad subject of the measure is clearly driven by the job in question. At a more detailed level, however, the assessment process may involve a range of sub-tasks which may introduce an unintended influence on performance. A hands-on work sample test for a psychomotor task administered with written instructions has a verbal, cognitive component because performance depends on the candidate’s ability to read and remember the task instructions. Research is only just beginning to shed light on these issues, maybe as a result of new mass technologies, such as video and computers as discussed above. The provision of alternatives to conventional methods may have spotlighted previously non-conscious decision-making about the structure of assessments.

Mode of Presentation and Response

Following on from the points made above, there are decision alternatives in constructing work
sample measures regarding the mode of delivery of information and the way in which the candidate is required to respond. Delivery and response can be behavioural, verbal or written. Behavioural assessments may have the benefit of reducing race effects (Schmidt et al. 1996). Goldstein, Yusko, Braverman, Smith and Chung (1998) found that the cognitive load associated with different AC exercises was related to racial sub-group differences in scores. When the cognitive component was removed statistically the sub-group differences disappeared but some incremental validity remained. The authors argue that cognitive load confounds scores for other performance dimensions of interest such as interpersonal skills and suggest alternative modes of delivery such as video-based measures.

Choices on these issues may not, therefore, be entirely driven by practical considerations such as cost and time to administer. The appropriateness and correspondence between the modes of delivery and response and the skills and abilities being measured need to be determined.

Conclusion

Work sample testing is best seen as a rationale for the assessment of current or likely future work performance rather than a single method. This approach seeks to achieve point-to-point correspondence between the content of the assessment measure and that of the job performance domain. It is also based on the assumption that an individual’s past behaviour is the best predictor of his or her future behaviour.

The types of measures encompassed by this general approach range from hands-on performance tests of job-related tasks to written tests of job-related knowledge. Differences and similarities between individual work sample inspired measures can be considered in relation to a number of dimensions, including fidelity, bandwidth and the specificity of the tasks. Hands-on performance tests have shown higher criterion-related validity and lower adverse impact against minority groups than any alternative selection procedures. However, work samples are subject to a number of limitations, mainly practical and financial, that mean they are not as generally useful as cognitive ability measures, which are equally valid but far more expedient.

Still, selectors are seeking new high validity methods of assessing performance with the convenience of cognitive ability measures but the low adverse impact and acceptability to candidates of work samples. The underlying principles of work sample testing have been used to create a number of hybrid measures, although the desired characteristics have not always been transferred. This is probably because although there are a number of theories, there is no detailed understanding of why work samples produce their associated outcomes. Consequently, reproducing the features of work sample tests related to desired outcomes and discarding those that are not cannot be a very precise process.

Selectors are faced with a number of fundamental decisions when considering the development and use of a work sample measure in comparison to alternative methods of assessment. These decisions will include the acceptable degree of trade-off between the investment of resources and criterion-related predictive validity, as well that between the desired levels of validity and adverse impact against minority groups. The extent to which selection is focused on short-term performance levels and specific job roles, rather than longer-term performance and achieving individual flexibility between job roles is another issue to be decided.

The work sample approach is, however, a theoretically sound and justifiable basis for developing selection measures. Measures based on this rationale have the potential to contribute to the achievement of some of the current challenges of employee selection, such as legal defensibility and workplace diversity.

References


Selection and Assessment, 7, 26–34.