The changing face of testing

Dave Bartram, *winner of the Society's Award for Distinguished Contributions to Professional Psychology, on the role for psychologists in a new web-based era.*

PSYCHOLOGICAL testing probably touches more people more often than any other application of psychology. We are tested from cradle to grave; as we progress through the educational system; when we are well and when we are failing to cope with life; when we want to get a job, at work and when we need help to know what to do when we leave work; to help us spend our leisure time well and to help us excel at sport.

As psychologists, we concern ourselves with trying to ensure that psychological tests are used in all these areas in ways that are equitable, so that we create accurate and valid descriptions of people and make fair discriminations between them. The concept of equity in assessment is a complex one, because it involves competing demands. When assessing someone for a job we need to ensure that we assess *all* those attributes that are relevant for the job and none that are not. The way we assess needs to be accurate, valid and free from irrelevant biases. In addition, the methods of assessment need to be acceptable – not only to be fair but also to be seen to be fair and reasonable by all the parties involved (see Gilliland, 1993). Finally, they must be practical, in terms of cost, ease of use, user training requirements, and so on.

In practice, ensuring equity always involves compromise. People find short assessments more acceptable than long ones, but we know that, other things being equal, length is related to reliability and hence to validity. So if we want to make tests shorter we have to make them more efficient in measurement terms in order to enhance equity.

We know people want tests to be easy to use and inexpensive, and not to require long training courses. But test design, development and production are costly time-consuming processes; creating easy-to-use output from tests is both difficult and expensive. So how do we satisfy the consumer demand for assessments that cover all the attributes we need to assess and that are acceptable and practical, and also cover the psychometric needs of reliability, validity and freedom from bias?

These tensions have always been present, but they have become even more apparent with the advent of internet-based testing. This has not only revolutionised the ways tests are used in the work and organisational field, but has also impacted on the design and production of tests – creating a need for greater efficiency and industrialisation.

How the internet changed testing

We can look on 1995 as the time the internet started to become part of the fabric of many people's everyday lives. Since then, the range of applications and volume of use have mushroomed. It took radio about 38 years to develop a worldwide audience of 50 million people; for television, this time shrank to 13 years. For the web, it took a mere four years. In the past decade the web audience has grown tenfold to over 500 million people.

An obvious impact of the internet is that tests and documents can be downloaded. This means that the internet can be used as a complete commercial solution for test publishers. There is no longer any need for printing and production, warehousing and postal delivery services. More significant for testing, however, is the shift in locus of control provided by the internet from the 'client-side' to 'server-side'. For paper-and-pencil testing, publishers have had to provide users with test items, scoring keys and interpretation algorithms. As these are 'public', the danger of compromise and security breaches is high. Test users can (and do) pass these materials on to people who are not authorised to use them. All the test data also resides with the user. The process of developing norms, checking the performance of test items and carrying out validation studies is dependent upon costly procedures for recovering data from users. For the internet that situation is reversed. The data and the intellectual property reside on the publisher's server. The user has access only to those parts of the process that they need.

While the internet poses lots of potential benefits for publishers, test users and test takers, it also poses problems. We can view these in the context of the different modes of test administration on the internet, which form the basis for the 2005 computer-based testing guidelines developed by the International Test Commission (ITC: see weblinks).

Open mode These are conditions where there is no means of identifying the test taker and there is no human supervision. Examples of this include tests that can be accessed openly on the internet without any requirement for test-taker registration. One of the problems that has arisen is the growth of 'testing sites' that offer all sorts of free or cheap tests with little or no information available about their technical properties. While many are offered for use for 'amusement' or 'fun', others are offered for use in real decision making (such as job selection). This year the BPS is launching a test registration scheme to help the unwary test taker know when they are taking a 'real' test and when they are not.

Controlled mode This is similar to the open mode in that no human supervision of the test session is assumed. However, the

WEBLINKS

International Test Commission: *www.intestcom.org*

BPS Psychological Testing Centre:

www.psychtesting.org.uk

test is only made available to known test takers. For the internet this is controlled through the requirement for the test taker to be provided with a username and password. This is the most widely used mode for the delivery of the major personality inventories and other self-report measures (for example, the OPQ32, 16PF, HPI, MBTI, and so on). However, it has raised a classic equity balance issue – whether asking people to complete inventories without supervision (which people find more acceptable and practical) impacts on the reliability or validity of the data. Fortunately, research suggests that use of controlled mode for administration does not have any adverse effect on the properties of this type of instrument (e.g. Bartram & Brown, 2004).

Supervised mode For this mode a level of human supervision is assumed, whereby the identity of the test taker can be authenticated and test-taking conditions validated. This mode also provides a better level of control over dealing with unexpected problems. For internet testing, this mode is achieved by requiring the test administrator to log-in the candidate and to confirm that the testing was completed correctly at the end of the session.

Managed mode This is where a high level of human supervision is assumed and there is also control over the test-taking environment. For computer-based testing this is achieved through the use of dedicated testing centres. The location is specified, the physical security levels are definable and known, and the nature and technical specification of the computer equipment is under control. As a consequence, test materials can be securely downloaded and item types can be used which make particular demands on the user work-station (e.g. streaming video).

The use of cognitive tests in controlled mode (i.e. restricted but unsupervised) raises interesting questions. We would generally regard cognitive tests as needing to be restricted to supervised conditions of administration, in order to ensure that cheating was not possible and the security of the items was not compromised. Baron *et al*. (2001), however, describe the use of online ability-screening instruments that are presented in controlled mode as part of a job recruitment process. A different test is created for each test taker, thus making it difficult for them to cheat. The delivery software is also carefully constructed to make it impossible to interfere with timing or other aspects of the delivery and administration (any such attempts are detected and reported). The use of online screening tests like this provides organisations with considerable benefits in terms of the average cost per hired person, as it enables them to screen out a higher proportion of applicants before they are brought into an assessment centre. However, it is also important to understand that the screening test is part of a larger process, a process which is made explicit to the test taker. Anyone who passes the screen may be reassessed under supervised conditions later in order to detect any false positives, and applicants are asked to 'sign up' to an honesty contract.

I believe we will see controlled mode becoming more like supervised mode as current technologies become embedded in everyday web interactions. Remote identification and authentication can be carried out through thumb-print or retinal eye-pattern recognition. Remote supervision is possible through CCTV monitoring and the use of data forensics (Foster & Maynes, 2004) will enable us to exercise a high degree of remote control over various forms of cheating. Finally, test- and item-generation technologies will allow us to create tailored, individual tests for people that will make it difficult for test security to be compromised.

Art, craft or technology?

The internet, by creating the pressure to move tests out of traditional low-volume individual or group administration modes into the remote high-volume modes, has also forced us to rethink the role of psychologists in the testing process. For some, testing is an art in that it relies almost wholly on the skills and creativity of the expert user of a set of tools that, in themselves, are of little value. For most others it is a craft, where standard procedures can be learnt and skills employed. However, increasingly it is becoming 'de-skilled' as technology starts to put the 'artist' and the 'craftsperson' into technological boxes.

For the work and organisational field, the dominant role of the psychologist is changing from the person who is the consummate artist – uncovering the mysteries of each person's psyche – to the expert craftsperson, who can add value to the results of using standard instruments through his or her expertise as a psychologist. There is also an increasing demand for the specialist technologist who can package the expertise in such a way that lay users (e.g. line managers) can get value directly from the results of testing.

Thus, one of the major factors affecting the impact of a test is how the scores are used and reported. It is surprising, therefore, that score reporting has received such scant attention, in comparison with other aspects of testing. The ITC guidelines on test use (ITC, 2001) emphasise the importance of providing feedback, of reporting that is unbiased and that does not over-interpret the results of a test. The more recent ITC guidelines on computer-based testing also place an emphasis on the need for test developers, test distributors and test users to pay attention to the provision of feedback.

In the traditional occupational testing scenario in the UK, results (especially of personality and other self-report inventories) were often fed back to test takers in a face-to-face session that provided opportunities for the feedback provider to address issues, and for the recipient of the information to ask questions. While it is still recommended that complex score reporting (such as the feedback of the results of a multiscale personality inventory) should be supported as much as possible with personal contact, the growth of internet-based testing has tended to make feedback and reporting as 'remote' an event as the test administration. The more complex issue of reporting back to non-experts (test takers or line managers in the occupational field) on constructs like personality, intelligence or motivation needs further attention.

It is because test results tend to impact indirectly on outcomes, because their effects are mediated through reporting processes, assessment policies and their combination with other information, that the simple psychometric concept of validity has limited impact in applied settings. Far more attention needs to be paid to this issue in the future to understand some of the differences between the science and the practice of testing.

The industrialisation of test production and delivery

A 'test factory' is a set of systems that brings together a range of technologies and levels of automation associated with test design, test development and manufacturing, warehousing and inventory control, test assembly and, finally, delivery of the test to the customer. Item development and banking requires a database, authoring and prototyping tools and interfaces, and some means of managing content through inventory control and management. Test assembly and construction requires item selection and quality-control procedures and procedures for dealing with the final rendering of the test – whether for online delivery or for offline printing as a paper test. The factory analogy is a useful one as it highlights the way in which we are beginning to apply industrial procedures to what has been a craft process.

The first generation of tests were pre-industrial. In the pre-industrial world, product manufacture was a distributed craft activity. Guilds and other organisations brought produce together from manufacturers and found markets for it. Until relatively recently the development of tests has followed the craft industry approach (and still does in many areas). A test author crafts a new product and then seeks a publisher to sell it.

In the second generation of test production, we see the application of industrialisation procedures to tests. The first stage of industrialisation was to bring these resources together into factories and to 'de-skill' the craftsmen. We are seeing that happen now as the larger test producers automate the procedures needed for item design and creation, develop item banks and item warehouses, and automate the quality-control procedures needed to ensure that the finished products meet some specified requirements and psychometric standards. However, this first stage of industrialisation still follows the traditional manufacturing model in most cases. It is a 'push' model that works on the basis of identifying a market need, creating a product that should meet that need and then pushing that out into the market. Tests are designed and developed, admittedly with far greater efficiency and control over quality, and then made available for delivery, being put into store either physically or virtually until required.

'All these changes are making it more and more important for us to be vigilant'

The third generation of test production involves the application of new industrialisation procedures to 'just-in-time testing'. Here the factory process is a 'pull' one rather than a 'push' one. The requirements are identified from the market, these requirements are converted into a technical specification and fed into the factory as an 'order' and the factory then delivers a test directly back into the market.

This move from craft industry to post-industrial manufacturing process is a major change in testing. Not surprisingly, it is in the licensing and certification field and the high-volume educational testing areas that we see these procedures most highly developed. The same industrialisation process is taking place in the occupational testing field now. SHL, for example, delivered over two million online test administrations worldwide last year, mostly in controlled mode, and the numbers are growing rapidly. In the past two or three years we have radically redesigned our test production procedures, developing industrialised just-in-time procedures like those discussed above.

The need for vigilance

All these changes are making it more and more important for us to be vigilant in the need for standards and guidelines to help people ensure that various competing demands on equity in assessment do not get out of balance. While we should not let demands for practicality result in tests becoming unreliable and invalid, we also have to recognise the need to address the changing nature of the environment within which testing is used, and to rethink how to keep technical standards in balance with user-group demands and pressures.

The various ITC guidelines will help in this process. Through the work of the BPS Steering Committee on Test Standards and the EFPA Standing Committee on Tests and Testing, we will continue to ensure that psychology remains centre-stage in defining these guidelines and standards. This becomes increasingly important as web-based technologies makes test construction tools and delivery mechanisms more accessible to all. Psychology has to show where and how it can add value.

■ *Professor Dave Bartram is Research Director for SHL Group. E-mail: Dave.Bartram@shlgroup.com.*

DISCUSS AND DEBATE

If test choice, administration and interpretation are all automated, and if administration is managed remotely, how meaningful is the concept of 'responsible test user'?

Do people cheat in unsupervised web-based cognitive testing; and if so, does it have an impact on assessment outcomes?

How do we convince consumers of test results that the technical quality of a test matters more than the superficial appearance of it or the 'glossiness' of a report?

What are the implications of the developments discussed here for the current BPS Level A and Level B standards for accrediting test users?

Have your say on these or other issues this article raises. E-mail Letters on psychologist@bps.org.uk or contribute to our forum via www.thepsychologist.org.uk.

References

Baron, H., Bartram, D. & Miles, A. (2001, April). *Using online testing to reduce time-to-hire.* Paper presented at the Society for Industrial and Organizational Psychology Conference, San Diego, CA.

Bartram, D. & Brown, A. (2004). Online testing: Mode of administration and the stability of OPQ32i scores. *International Journal of Selection and Assessment, 12*, 278–284.

Foster, D. & Maynes, D. (2004, February). Detecting test security problems using item response times and patterns. In D. Foster & R. Hambleton (Chairs) *Solve testing security problems using technology and statistics.* Symposium conducted at the Association of Test Publishers Conference.

Gilliland, S.W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694–734.

International Test Commission. (2001). International Guidelines for Test Use. *International Journal of Testing, 1*, 93–114.