Hornke, L.F. & Kersting, M. (2005). Optimizing Quality in the Use of Web / Computer based Testing for Personnel Selection. In D. Bartram & R. K. Hambleton (Hrsg.), Computer-Based Testing and the Internet: Issues and Advances (p. 149-162). New York: Wiley.

CHAPTER 8

Optimizing Quality in the Use of Web-Based and Computer-Based Testing for Personnel Selection

Lutz F. Hornke and Martin Kersting
Aachen University of Technology, Germany

INTRODUCTION

This chapter traces the history and development of computer- or web-based testing. It is argued that for all kinds of testing the standards of quality and control can only be achieved by experimental approaches that guarantee objectivity, reliability and, above all, validity. A brief look at possible areas of application clearly indicates that there is a growing demand for computer- and web-based tests. Apart from traditional multiple-choice item formats, the most promising approach is seen in simulation and game test procedures. Very often they are cast in joblike scenarios so that managers and applicants find them particularly attractive for their real life perspective. However, these tests are problematic in terms of providing good psychometric measurements. Finally, we propose some ideas and criteria for quality assurance.

The introduction of computer-assisted testing (CAT) is directly connected to the prevalence of personal computers themselves. Thus, 1/1/1980, the earliest date put into the boot ROM of the first IBM PCs, may be taken as the starting point of CAT. Sands, Waters, and McBride (1997) tell this story very well. However, there had been earlier attempts, which were compiled by Suppes, Jerman, and Brian (1968), and by Suppes and Morningstar (1972) in their seminal report on computer-based training. And if one goes back even further, then Pressey's (1926) learning machine may be considered a very early attempt at 'mechanising' training and testing, although it was quite far removed from what we today understand as computerized testing.

Computer-Based Testing and the Internet: Issues and Advances.
Edited by D. Bartram and R. K. Hambleton.

Hence the advent of versatile PCs marks the beginning of the modern approaches to testing and training. PCs flourished as mass storage became available, so that nowadays they are to be found in every office and most homes. This increase of storage on hard drives and CDs and in other forms, also allows for the use of sounds and motion in training and testing. It is possible to envision a future when testing embedded in training and psychological testing will be done with PCs or with something like a general home multi-media control box, the successor of our present television set, which will be a versatile computer that, apart from radically transforming our home life, providing entertainment, tele-work and education, will also fulfil training and testing functions. Already the modern testing of today emulates a good television show and is a far cry from the electronic page turners of the early days.

In psychological measurement generally, and in computer-based testing in particular, it is considered to be crucial to keep all items concealed from participants. Only specimen items are released to participants so that they get an idea of what is going to be asked of them, what the testing session will look like, and what possible results would emerge. In many cases this information is so inadequate that it patently fails in preparing the participants for the real testing situation.

An entire literature stemming from psychological experimentation has been borrowed for computerised testing, but the fundamental emphasis is still on 'experimental control'. Testing should be organised in such a way that nothing but the underlying psychological trait or competency carries over into the results. Standardisation is another important factor besides 'experimental control'—seating arrangements, lighting, taped instructions, strict time control, and other conditions have to be the same for every participant. If any of the situational factors are considered to have a bearing on the final results, then all participants must be equally subjected to them in the same manner. Paper–pencil testing, individual testing of one person by one psychologist, and even computer-assisted testing follow the same principle of control and standardisation in order to ensure quality of results.

However, this sense of strict control becomes more difficult to implement with the notion of an entertainment box in any household where testing may also be delivered. Nevertheless, the internet makes this possible, and there is no way of stopping it. Rather, the challenge is how to assure quality of individual results and personnel decisions. Careers and pay will depend on it, as will self-esteem and social recognition. It seems that solutions at hand continue to follow the strict control paradigm and use technical devices to achieve this end, but the question remains as to how the professional development of internet-based testing programs can be taken forward.

COMPUTER- AND INTERNET-BASED TESTING

It is impossible to predict what form CAT will take in the future. New programming tools, new test designs, and new ways of measuring complex

behaviours are constantly being devised. So most of the features of tests are in flux. The sections below aim simply to provide an overview of contemporary testing programmes. Computer- and web-based tests were advocated for the following reasons. They:

- make scoring and norm-based interpretation easier and assure objectivity;
- use psychometric features to estimate a person's ability efficiently;
- allow us to design items that make use of certain *multimedia* features such as simulation and *dynamic* presentation formats in order to *better* diagnose problem solving abilities;
- provide scores in order to obtain immediately a:
  - number right score, formula score, maximum likelihood score (1PL, 2PL, 3PL)
  - process scoring (as in the learning test approach)
  - person fit index (IRT based), response time index, plausibility check
  - norm or criterion referenced score
  - feedback of results (personalised, immediate, quality based on item content, behaviour oriented).

ABCNews (17 December 2000) states on its web-page that 'The biggest advantages of computer-adaptive testing are time-related. The test is offered on demand at many testing centers around the country. Students see their scores immediately. Adaptive tests are usually shorter than paper-and-pencil tests. The GRE, for example, has gone down from three hours to two hours, 15 minutes. Proponents of computerised tests say they more accurately reflect the taker's ability. The computerised GRE now includes a writing assessment. The Test of English as a Foreign Language now includes an audio component'.

This list of advantages of CATs is further supplemented by the accounts given in other chapters in this book. Some of what flows over the internet at the moment is innovative not so much in terms of constructs or psychometrics but more in terms of design features of new item types. Here, one finds not only the good old multiple choice items but also items enriched with sounds, motion pictures, and audio–video sequences. Laboratories announce that olfactory as well as tactile oriented items may be the fad of tomorrow. Some 400 smells can be produced electronically today and could be used in assessing firemen, health inspectors, perfume designers and many more. So quite new means of 'controlling' testing are being developed.

The more recent psychometric developments in item response theory (IRT) will have to deal with issues of equivalence of scales, parallel measures, repeated measurement, and psychometric properties. Furthermore, the design of a testing system has to address the problems of test faking, checking the identity of participants, personal data and item security, and response latencies due to participants or due to operating systems or hardware.

Administering tests over the web requires taking account of environmental and situational control issues, public versus in-house provision of test systems, and computer system requirements and idiosyncrasies such as different

handling of shapes, colour, sound, motion, streaming formats for motion pictures, display settings, browser configuration, and stability of communication on/off line. Undeniably, reliability, validity, appeal, and user friendliness and, for most companies and managers, costs, are important. Again, this is by no means a complete list.

TESTING (FOR) THE HOMO LUDENS

Problem Solving Scenarios

While the vast majority of all current CATs and web-based tests use traditional page-turning items, it seems that the more innovative item sets built on content and psychometric criteria as in Computer Adaptive Tests and problem solving scenarios will be very well suited for the *Homo ludens* of the future. Therefore the latter test formats are given some attention here. Problem solving scenarios explore aspects such as the finer cognitive processes involved in problem solving, e.g. typical stages of problem solving, general or specific strategies, typical errors and differences in the problem solving skills of experts and novices. Research also focuses on the relation between problem solving skills and personality characteristics on the one hand, as well as the relation between problem solving ability and intelligence or knowledge, on the other. It is claimed that newly coined concepts such as 'heuristic competencies', 'operative intelligence', 'system based thinking', and 'net based decision making' can be measured through these scenarios.

According to Dörner, Kreuzig, Reither, and Stäudel (1983), complex problems can be described and simulated as systems of interconnected variables. These problems have the following characteristics:

- *Complexity.* Numerous aspects of a situation have to be taken into account simultaneously by participants.
- *Interconnectivity.* The various aspects of a situation are not independent and cannot, therefore, be independently influenced. Interconnectivity also includes the important role of feedback loops and side-effects.
- *Dynamics.* Changes in the system conditions also occur without intervention from the problem solver.
- *Intransparency.* A situation is labeled intransparent when only a part of the relevant information is made available to the problem solver.
- *Polytely.* Sometimes the problem solver must simultaneously pursue multiple and even contradictory goals.

Computer-simulated scenarios are used as a way of translating such complex problems into an assessment context. Subjects have to run a city 'transportation system' (Broadbent, 1977), or manage a small factory. Funke (1991) provides an overview of the various scenarios. Mostly one finds simulations of a 'company' competing on some market with different 'organisational units'

(e.g. research policy, developing and launching new products). For example, the TAILOR SHOP scenario or equivalent derivatives asks participants to maximise benefits and stabilise the company over a period of time. The success of each participant or a small group is judged against 'cumulative gains' and 'future orientation'. Information about these performance criteria is displayed to participants continuously. The scenario as such deals with the production of shirts. However, the system is kept relatively un-transparent, being based on 24 system variables of which only 11 are modifiable by participants. Results emphasize measures such as final value of the company, relative gains with equal weights given to all periods, and relative gains with higher weights given to later periods.

Another system, FSYS (Wagener, 2001) has a different set-up: an agricultural service company dealing with tree plantation and care and marketing lumber. Profits may be maximised by taking action against insects and anti-soil pollution. Participants are evaluated according to the stock value of the company, avoidance of errors, setting priorities, early exploration of system, information retrieval before decision making, attention given to stock value, and assertiveness while controlling company.

The most prominent example is the simulation called 'Lohhausen', where the subject has to act as the mayor of a small simulated town with the name 'Lohhausen' (Dörner et al., 1983). Subjects are able to manipulate taxes, influence production and sales policies of the city factory or the housing policy and so on. They are simply told to provide for the prosperity of the town over a simulated ten-year period within eight two-hour experimental sessions.

Scenario Difficulty

Of course, it is possible to give some ideas as to what moderates a scenario's difficulty:

- *Content:* knowledge in general, knowledge of the subject matter of the scenario, product, market and so on, acceptance of content and task by participants, generality or specificity of content, and impact of lateral knowledge.
- *User interface:* methods of instruction, mode of interaction with other participants or the system, motivational character of the scenario and displays, ergonomics of the interface, mode of acquiring information, note taking system, organisation of interaction, and means of input.
- *Formal structure of system:* number of variables, manipulation of system, number of periods and interaction time, real time simulation, number of nodes between input and internal variables, and nature of interaction of variables.
- *Feedback of information:* information about trends, delayed system information, percentage of trivial rules, assisting functions; information about system status, rules, and success criteria.

- *Evaluation of results:* methods, dimensions, functional structure of systems, randomness, compatibility with prior knowledge, structural redundancy, antagonistic features, number and reducibility of rules, compensatory features, time-lagged controls, side-effects, internal dynamics, reversibility, traps, difficulty range of tasks, and needed forecasting.

These facets may be used to make a scenario easier or harder, but it is not possible to predict the overall difficulty level from the facets before the test.

Advantages and Disadvantages of Using Computer-Based Scenarios for Diagnostic Purposes

In Europe, and especially in the German-speaking countries, computer-based scenarios are used as assessment tools in both research and practice. Some of the complex problem-solving scenarios used in the context of personnel selection are presented by Funke (1995), and a discussion of the advantages and disadvantages of this form of application can be found in Funke (1998). The main advantages of using computer-based scenarios as diagnostic tools are that the tasks (1) are highly motivating and (2) involve novel demands that (3) are deemed to have higher face validity than intelligence tests, and (4) test takers enjoy working with the simulations (see Kersting, 1998).

Much attention is given to face validity in that authors advocate that the simulation scenarios are much closer to the nature and challenges of prospective jobs than ordinary tests. So incumbents are led to see simulation scenarios as representative tasks. The lay public is intrigued by the content and purported relation of the scenarios to real life challenges. This makes it hard, from time to time, to insist on empirical facts. For example, rarely is a thorough task or job analysis presented to verify the cognitive demands of both jobs and simulation tasks in regard to validity; rather, users rely on the nice mock-ups of a scenario per se. There is a strong reliance on 'faith' to demonstrate what one encounters in real life.

Candidates are more open in their responses than with traditional 'tests' (Shotland, Alliger, & Sales, 1998). Feedback and administration within organisations is easier as the meaning of the test is 'obvious', face validity increases motivation, face validity correlates with the attractiveness of the organisation, and managers 'love' face valid computer tests (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). In relation to online assessment in general, 84% of users of an on-line procedure reported a 'positive experience' (Vlug, Furcon, Mondragon, & Mergen, 2000) and on-line assessment was rated significantly 'more fair' and 'more satisfying' than the paper–pencil version (Reynolds, Sinar, Scott, & McClough, 2000). Scenario-based tests provide more realistic job preview, require less time for follow-up interviews, and make it easier for candidates to self-select and to accumulate knowledge about job characteristics during the selection procedure. There is nothing

wrong with face validity—it is just the weakest criterion in arguing for the quality of a testing instrument as a predictor of future performance.

However, the diagnostic use of computer-based scenarios also entails serious difficulties that have yet to be overcome.

(1) The central question of appropriate approaches to the operationalization of problem solving quality remains largely unanswered.
(2) The reliability of the measurements obtained with some of the computer-based scenarios is less than satisfactory (see below).
(3) The existence of a task-independent and thus generalizable problem solving ability has not yet been substantiated. This indicates that the ability to steer the system is dependent not only on the skills of the problem solver, but evidently also on the nature of the task in question.
(4) The main problem is construct validity. It is still unclear which skills are actually measured by means of the computer-based scenarios (see Süß, Kersting, & Oberauer, 1992; Süß, Oberauer, & Kersting, 1994). Either the measurement has to be interpreted as an indicator of an independent *ability construct* (as suggested by newly coined concepts such as 'networked thinking', 'heuristic competence', and 'operative intelligence'), or the scenarios are regarded as a new *measurement method* which, in a certain respect, is better able to measure established constructs such as intelligence than has previously been the case (e.g. in a more differentiated manner or with a higher level of acceptance). Beckmann and Guthke (1995) have summarized the European research dealing with the controversial relation between traditional measures of intelligence and problem solving skills.
(5) Evidence for the criterion validity of the measures used is also urgently needed.

Reliability

Theoretically, it is to be expected that the reliability of scenario-based control performance measures will be lower than that of other performance measurements, e.g. intelligence tests, as problem solving scenarios provide more degree of freedom for the test taker and control performance is probably determined by heterogeneous factors. Due to the dependence of each system state on the preceding ones, unintended sources of variance such as motivational fluctuations or fatigue can build up, creating sequential dependencies in processing times. The long duration of response times, as compared to intelligence tests, does not increase the reliability. As a rule, each run of the simulation merely results in 'single act' criteria (Fishbein & Ajzen, 1974).

Psychometric results from simulation games yield in some cases estimates of different kinds of reliability. To give some examples, Köller, Strauß, and Sievers (1995) used three variants, all of which are based on a 'tailorshop' scenario: Textile Production, Fuel Delivery, Coal Distribution. As

dependent variable the 'number of periods with capital gains' was used. For Fuel and Coal scenarios the number of periods with capital gains correlated at $r = 0.69$, for Textile with Fuel $r = 0.41$ and for Textile with Coal $r = 0.44$. Funke (1995) used a parallel version of a video recorder production scenario and found an $r = 0.83$.

As for *internal consistency*, Müller (1993) found $r = 0.83$ to $r = 0.86$ for two independent parts of a simulation game. In contrast, the test–retest reliability, which Müller based on a repeated test after 5 months, was $r = 0.53$ and thereby clearly lower. In the investigations by Süß et al. (1992, 1994), 137 test takers repeated the 'tailorshop' after an interval of one year. The retest stability was $r = 0.46$.

According to the reviewed studies, in certain scenarios at least some specific measures satisfy the reliability requirements necessary for diagnostic use, in spite of unfavourable theoretical prerequisites (e.g. multiple performance conditionality, learning effects, and the low level of aggregation). Funke (1995, S. 189 f) states that the reliability of performance on scenario-based problem solving tasks is on the same level as the reliability of so-called 'simulation oriented' methods (such as group discussions). In general, performance on tasks such as these shows lower reliability than intelligence tests.

Criterion Validity

Thus far, only a single study (Kersting, 1999, 2001) has directly compared the predictive criterion validity of computer-based scenarios with the validity of existing procedures deemed to have overlapping coverage. A total of 104 police officers were tested. For a subgroup of 26 participants it was not possible to analyse criterion validity because they attended a police academy.

For all participants general intelligence according to the Berlin Model of Intelligence by Jäger was assessed (BIS; Jäger, 1982, 1984; see Bucik & Neubauer, 1996; Wittmann, 1988). Before or after completing the intelligence test all participants worked on two computer-based problem solving scenarios. The scenarios that were used both referred to an economic context. One was the so-called 'tailorshop' (where subjects have to manage a tailoring factory); the other was the scenario 'Disco', which required the management of a computer chip factory. In both these scenarios the goal was to manage the company in a way that maximized the company's assets at the end of the game. Control performance is – according to the instructions given – measured as 'final asset value'. Both scenarios generated similar findings. Due to restrictions of printing space, only the results of the 'tailorshop' scenario will be presented here. After having worked on the 'tailorshop' task subjects had to work, among other things, on a paper–pencil test to assess system-specific knowledge. This paper–pencil test asks, for example, about the relationship of distinct key variables in the 'tailorshop'.

For the purpose of criterion validity a series of retrograde and concurrent criterion measures were assessed. The present account focuses on predictive

validation. A postal questionnaire of criterion data on participant's job performance was conducted on average one year and seven and a half months after the initial testing. For some 73 out of the 78 initial participants the questionnaires were returned; this equals a response rate of 94%. Their age ranged from 28 to 57 (median = 36). In most cases job performance ratings were obtained from their direct supervisors, who were asked to rate their subordinates on a series of job-related performance dimensions, all of which could be attributed to intelligence, problem solving ability, and cooperative ability. The latter was used for the purpose of discriminant validation.

Raters were at first requested to mark for each of the three dimensions on a four-point scale how in terms of the respective dimensions the ratee, in comparison to colleagues, belonged to different quartiles of the comparison group. After these initial questions, which assess the extent of general abilities, 15 items tapped specific behaviours considered to be relevant to everyday job performance. They were indicators of constructs such as intelligence, problem solving ability, and cooperative ability. To give an example, in one of these items related to problem solving ability the term 'sensibility and flexibility towards changes' has been explained as follows: 'The achievement to continously confront one's own actions with reality and to adjust one's own decisions flexibly to feedback'.

However, the rating scales measuring problem-solving behaviour and intellectual behaviour on the job were highly correlated, and were thus combined into a scale called 'quality of problem solving behaviour and intellectual behaviour in daily job performance'.

Table 8.1 shows correlations of these criteria with intelligence, problem solving and knowledge. The table shows the correlations of these predictors with the supervisor assessments made about one and a half years later. Problem solving and intelligence related job performance could best be predicted by intelligence ($r = 0.39$). The problem solving scenario and the knowledge test also proved to be valid predictors.

Table 8.1 Predictive criterion validity—comparing problem solving scenarios and intelligence tests

	Quality of problem solving behaviour and intellectual behaviour in daily job performance	Partial correlations, controlling for...			Job performance: cooperative ability
		Intelligence	Problem solving	Knowledge	
Intelligence	0.39**		0.33**	0.32**	0.07
Problem solving	0.37**	0.29*		0.32**	0.19
Knowledge	0.30*	0.23	0.23		0.06

Source: From Kersting (2001). *Diagnostica*, 47, 67–76.

For all predictors the assessment of cooperative ability demonstrated discriminant validity. Additionally, partial correlations were computed. It is immediately obvious that the different variables predict similar variance to a large extent. This leads to the question of whether a combination of predictors is able to raise the validity of the prediction, i.e. incremental validity. This was tested by a hierarchical regression analysis. In the first step, based on the highest bivariate correlation, intelligence was included, yielding an R of 0.39. By also including system specific knowledge about the 'tailorshop' as a second predictor, R was raised to 0.46. This equals an incremental percentage of explained variance of 7%. Another increment in prediction was achieved by including the third predictor, control performance on the 'tailorshop', but was significant only at the 10% level. After inclusion of all predictors a multiple correlation of $R = 0.50$ was obtained.

Considering a predictor above and beyond intelligence thus provides a substantive contribution to predictive power. On purely statistical terms control performance and knowledge proved to be similarly adequate in achieving significant prediction increments. The variable that was added to intelligence on the second step led to a statistically significant increase of the multiple R. The variable that was included on the third step did not. The sequence in which the predictors were included in the hierarchical regression shown here was based on the theoretical assumption that the systematic variance in control performance on problem solving scenarios can essentially be attributed to intelligence and knowledge. Accordingly, based on these theoretical assumptions, intelligence and knowledge were given priority in the hierarchical regression analysis.

Problem solving scenarios are diagnostically interesting, because it is not only intelligence *but also knowledge* that is required for managing these scenarios. Inclusion of knowledge in job performance may yield an increment in overall validity. For practical purposes it may be worthwhile to use both intelligence and knowledge assessment. Problem solving scenarios will be helpful in this regard, because managing complex problem solving scenarios demands acquiring knowledge, which could subsequently be tested.

Significant progress in the domain of problem solving assessment cannot be expected until both the operationalization of problem solving quality and the psychometric quality of assessment instruments are improved. Above all, it is essential to classify the ability tapped by the performance measures within an existing nomological network. Studies are required in which sufficiently reliable measures are implemented by means of *different* computer-based scenarios, and differentiated measures of intelligence are administered in sufficiently large samples. At the same time, tests of additional theoretically relevant constructs such as knowledge also need to be administered. In investigations of this kind (see Wittmann & Süß, 1999) it was shown that the systematic variance captured by problem solving scenarios can mainly be attributed to intelligence and prior knowledge. There is no empirical evidence for the existence of something like a problem solving ability as an independent construct.

QUALITY ASSURANCE

Returning to the initial argument for a controlled psycho-diagnostic experiment, it seems obvious that the test environment for unsupervised web-based testing in particular is quite uncontrolled. There may be background proctors such as friends and relatives if testing is done at home. Moreover, participants may use books, information materials, or the 'internet' as an aid in responding, not to mention unobserved effects resulting from emotional and physiological states such as fatigue, boredom, and so on. Even the differences in technical equipment, browsers, operating systems, and displays may cause variations in results that are error variance. Participants may 'fake', repeat tests, or return to previous items, if no precautions are taken. This can be remedied to some degree by sophisticated software, but ultimately full control is only possible by using professional testing centres as is done by ETS and other big test delivery companies. Even in self-assessments for vocational and educational purposes, some minimal standards for controlling the testing situation must be ensured.

It should be mentioned here that simulations and games have their own drawbacks and lack of proper control too. Typical problems include misunderstanding instructions and lack of personal contact or emotional 'rapport' with assessors, which is hardly compensated by an impersonal FAQ list, wordy reports based on few items, impersonal transfer of results, no personal feedback in case of unfavourable results, lack of acceptance of the psychometric procedures, security of data transmission, anonymity, and test fairness, to name just a few.

All the various drawbacks and problems of control discussed in this chapter are nothing but a challenge to improve quality. It is essential that several qualitative and quantitative studies are carried out to identify and remove flaws. With modern IRT it is not only possible to reach an economical score estimate, but also to use statistical fit indices in order to gauge whether the participant is responding in line with a theoretical model. Whenever there is a low fit for a series of responses, it becomes imperative to find out how to improve the test administration algorithm. The same holds true for timing controls and protocols. They tell, on an item by item basis, how long a participant has been working on each task. They provide the basis for detecting aberrant response patterns. Having a cup of tea during the test may be fine, but a ten minutes interruption of a cognitive task lowers validity and reliability!

Intelligent programming of computer and web based test administration routines for the future poses a big challenge. Group testing in the past showed that participants did not always function at the level for which the test was designed, frequently leading to false decisions. To overcome such problems in the open environment of web testing is the challenge that lies ahead of us. ETS, for example, have used video cameras to monitor participants in order to safeguard against 'false test takers' and 'helpers'. Modern technology will bring further support in this respect.

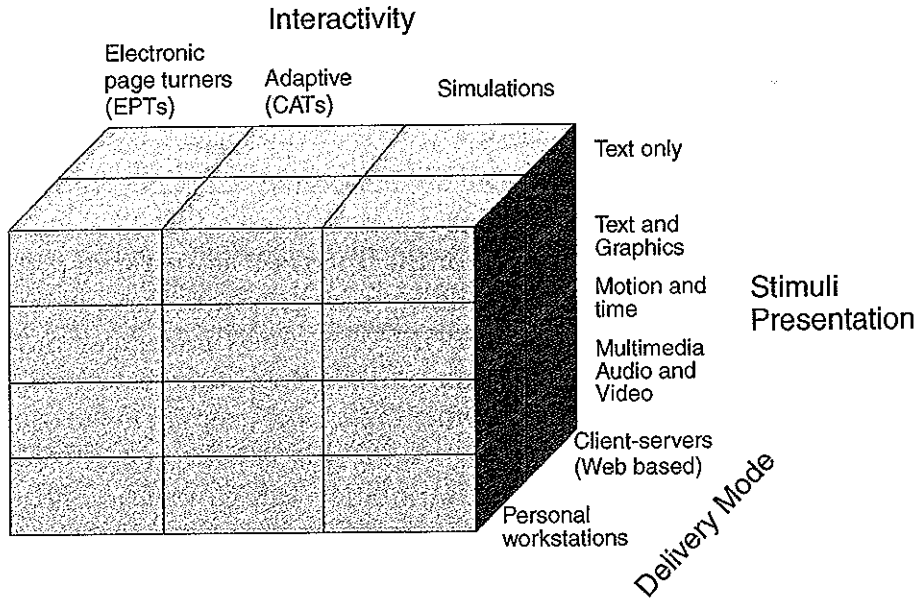

Figure 8.1 Three mode model of modern tests and their future challenges (in red)

It seems that internet-based tests lead to more self-disclosure (Locke & Gilbert, 1995), an aspect assessors are looking forward to with some excitement. The more we invest in quality assurance and personally appealing instructions, the more the people will come to value these assessments.

Web-based tests in the future will be much more than they are today (see Figure 8.1). They are communications with people who are very eager to learn about themselves and the demands of their job, a company, a university, and the like. Cheating is deplorable and exercised in order to attain a particular goal. Thus future tests need to be subjected to the best quality assurance procedures available, well marketed, and presented in an intelligent way so that everyone understands that cheating is counterproductive and works against the interests of both the individual and the organisation. Those who understand this will respond to web-based testing in the controlled manner the program and the assessors require. Participants should enrich their test taking by exercising internalised control, so that the test enables them to show their true behaviours and characteristics. In this regard web-based testing, apart from posing a challenge in terms of experimental controls, psychometric intricacies, and measurements, also opens up fresh opportunities for a gaining a wider recognition of psychological testing.

REFERENCES

Beckmann, J. F., & Guthke, J. (1995). Complex problem solving, intelligence, and learning ability. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 177–200). Hillsdale, NJ: Erlbaum.

Broadbent, D. E. (1977). Levels, hierarchies, and the locus of control. *Quarterly Journal of Experimental Psychology*, **29**:181–201.

Bucik, V., & Neubauer, A. C. (1996). Bimodality in the Berlin Model of Intelligence Structure (BIS): A replication study. *Personality and Individual Differences*, **21**:987–1005.

Dörner, D., Kreuzig, H. W., Reither, F. & Stäudel, T. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität* [*Lohhausen. On dealing with uncertainty and complexity*]. Bern: Huber.

Fishbein, M., & Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review*, **81**:59–74.

Funke, J. (1991). Solving complex problems: Exploration and control of complex systems. In R. J. Sternberg & P. A. Frensch (Eds.), *Complex problem solving: Principles and mechanisms* (pp. 185–222). Hillsdale, NJ: Erlbaum.

Funke, J. (1998). Computer-based testing and training with scenarios from complex problem solving research: Advantages and disadvantages. *International Journal of Selection and Assessment*, **6**:90–96.

Funke, U. (1995). Using complex problem solving tasks in personnel selection and training. In P. A. Frensch & J. Funke (eds.), *Complex problem solving: The European perspective* (pp. 219–240). Hillsdale, NJ: Erlbaum.

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells [Multimodal classification of intelligence tests: Experimentally controlled development of a descriptive model of intelligence structure]. *Diagnostica*, **28**:195–225.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Research on intelligence structure: Competing models, new developments, perspectives.] *Psychologische Rundschau*, **35**:21–35.

Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren [Differential-psychological aspects of applicants' acceptance of intelligence tests and problem solving scenarios as diagnostic tools for personnel selection]. *Zeitschrift für Arbeits- und Organisationspsychologie*, **42**:61–75.

Kersting, M. (1999). *Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien? [Assessment and personnel selection with computer-simulated problem solving scenarios?]* Göttingen: Hogrefe.

Kersting, M. (2001). Zur Konstrukt- und Kriteriumsvalidität von Problemlöseszenarien anhand der Vorhersage von Vorgesetztenurteilen über die berufliche Bewährung [On the construct and criterion validity of problem solving scenarios based on the prediction of supervisor assessment of job performance]. *Diagnostica*, **47**:67–76.

Köller, O., Strauß, B., & Sievers, K. (1995). Zum Zusammenhang von (selbst eingeschätzter) Kompetenz und Problemlöseleistungen in komplexen Situationen [Correlation of (self assessed) competence and performance of problem solving in complex situations]. *Sprache & Kognition*, **14**:210–220.

Locke, S. D., & Gilbert, B. O. (1995). Method of psychological assessment, self disclosure, and experiential differences: A study of computer, questionnaire, and interview assessment formats. *Journal of Social Behaviour and Personality*, **10**:187–192.

Müller, H. (1993). *Komplexes Problemlösen: Reliabilität und Wissen* [*Complex problem solving: Reliability and knowledge*]. Bonn: Holos.

Pressey, S. L. (1926). A simple apparatus which gives tests and scores – and teaches. *School and Society*, **23**:373–376.

Reynolds, D. H., Sinar, E. F., Scott, D. R., & McClough, A. C. (2000). Evaluation of a Web-based selection procedure. In N. Mondragon (Chair), *Beyond the demo: The empirical nature of technology-based assessments*. Symposium conducted at the 15th Annual Society for Industrial and Organizational Psychology Conference, New Orleans, LA.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized Adaptive Testing*. Washington, DC: American Psychological Association.

Shotland, A., Alliger, G. M., & Sales, T. (1998). Face validity in the context of personnel selection: A multimedia approach. *International Journal of Selection and Assessment*, **6**:124–130.

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, **46**:49–76.

Suppes, P., & Morningstar, M. (1972). *CAI at Stanford 1966–68. Data, models, and evaluation of arithmetic programs*. New York: Academic.

Suppes, P., Jerman, M., & Brian, D. (1968). *Computer-Assisted Instruction: Stanford's 1965–66 arithmetic program*. New York: Academic.

Süß, H. M., Kersting, M., & Oberauer, K. (1992). The role of intelligence and knowledge in complex problem-solving. *The German Journal of Psychology*, **16**:269–270.

Süß, H. M., Oberauer, K., & Kersting, M. (1994). Intelligence and control performance on computer-simulated systems. *The German Journal of Psychology*, **18**:33–35.

Vlug, T., Furcon, J. E., Mondragon, N., & Mergen, C. Q. (2000). Validation and implementation of a Web-based screening system in the Netherlands. In N. Mondragon (Chair), *Beyond the demo: The empirical nature of technology-based assessments*. Symposium conducted at the 15th Annual Society for Industrial and Organizational Psychology Conference, New Orleans, LA.

Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios: Taxonomie, Entwicklung, Evaluation* [*Psychological assessment with complex scenarios: Taxonomy, development, evaluation*]. Lengerich: Pabst.

Wittmann, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505–560). New York: Plenum.

Wittmann, W. W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem solving via Brunswik Symmetry. In P. L. Ackerman, P. C. Kyllonen & R. D. Roberts (Eds.), *Learning and individual differences* (pp. 77–104). Washington, DC: APA.

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA
Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA
Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany
John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia
John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809
John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Computer-based testing and the internet: issues and advances/edited
by Dave Bartram, Ronald K. Hambleton.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-470-86192-9 (cloth : alk. paper)
ISBN-10: 0-470-86192-4 (cloth : alk. paper)
ISBN-13: 978-0-470-01721-0 (pbk. : alk. paper)
ISBN-10: 0-470-01721-X (pbk. : alk. paper)
1. Psychological tests—Data processing. I. Bartram, Dave, 1948-
II. Hambleton, Ronald K.
BF176.2.C64 2005
150'.28'7—dc22 2005011178

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13 978-0-470-86192-9 (hbk) 978-0-470-01721-0 (pbk)
ISBN-10 0-470-86192-4 (hbk) 0-470-01721-X (pbk)

Typeset in 10/12 pt Palatino by Thomson Press (India) Limited, New Delhi
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Edited by
Dave Bartram and
Ronald K. Hambleton

Computer-Based Testing and the Internet

Issues and Advances

Computer-Based Testing and the Internet

Issues and Advances

Edited by

Dave Bartram
SHL Group plc, Thames Ditton, Surrey, UK

Ronald K. Hambleton
University of Massachusetts at Amherst, USA

John Wiley & Sons, Ltd