COMPUTERIZED TESTS AND OTHER NEW TYPES OF TEST

Julio Olea*, Francisco J. Abad* and Juan R. Barrada**
*Universidad Autónoma de Madrid. **Universidad Autónoma de Barcelona

The paper provides a short description of some test types that are attracting considerable interest in both research and applied areas. The main feature of a computerized adaptive test is that in spite of the examinees receiving different sets of items, their test scores are in the same metric and can be directly compared. Four other test types are considered: a) model-based tests (a model or theory is available to explain the item response process and this makes possible the prediction of item difficulties), b) ipsative tests (the examinee has to select one among two or more options with similar social desirability; hence, these tests can help to control faking or other examinee response biases), c) behavioural tests (personality traits are measured from non-verbal responses rather than from self-reports), and d) situational tests (the examinee faces a conflictive situation and has to select the option that best describes what he or she would do). The paper evaluates these types of tests, comments on their pros and cons and provides some specific examples.

***Key words:** Computerized adaptive test, Situational test, Behavioural test, Ipsative test, automatic item generation.*

Recientemente se ha producido un considerable desarrollo de los tests adaptativos informatizados, en los que el test se adapta progresivamente al rendimiento del evaluando, y de otros tipos de tests: a) los test basados en modelos (se dispone de un modelo o teoría de cómo se responde a cada ítem, lo que permite predecir su dificultad), b) los tests ipsativos (el evaluado ha de elegir entre opciones que tienen parecida deseabilidad social, por lo que pueden resultar eficaces para controlar algunos sesgos de respuestas), c) los tests conductuales (miden rasgos que ordinariamente se han venido midiendo con autoinformes, mediante tareas que requieren respuestas no verbales) y d) los tests situacionales (en los que se presenta al evaluado una situación de conflicto laboral, por ejemplo, con varias posibles soluciones, y ha de elegir la que le parece la mejor descripción de lo que el haría en esa situación). El artículo comenta las características, ventajas e inconvenientes de todos ellos y muestra algunos ejemplos de tests concretos.

***Palabras clave:** Test adaptativo informatizado, Test situacional, Test comportamental, Test ipsativo y generación automática de ítems.*

It was around two years ago that researchers into the history of psychology in Spain (Quintana, Bitaubé, & López-Martín, 2008) recovered and published some “Notes for a Psychology course applied to professional selection”, dating from 1924 and found among the teaching materials of one Dr. Rodrigo Lavín for his lectures in Experimental Psychology. This veritable gem of a document represents one of the first systematic references in Spain to the types and uses of tests. In the author’s words: “*Since observation provides us with little usable data, and conversation and interviews are insufficient for discovering the abilities of applicants, we must take recourse to tests. It might be said that tests are only at their inception, and yet extraordinary developments have already taken place; this suggests what will happen with the passage of time*”. Dr. Lavín mentioned the existence at that time of tests of skills or abilities, both general and specific, as well as the particular importance in the professional selection context of tests of strength, of resistance to fatigue, of motor control and of mental capacities (attention, sensation and perception, imagination and general intelligence).

The passage of time having brought us to the present, the development of tests has indeed been extraordinary, as Lavín anticipated, with regard to both their variety and their complexity. An indication of this is that simple classifications of test types (for example, that which distinguished between “printed tests” and “manipulative tests”, or those which referred to their different content) have become obsolete due to the emergence of new types of test which were difficult to envisage in the past. This can be attributed to different factors:

✔ *Technical progress*. The development of the psychometric models underpinning the metric properties of tests and the evolution and decrease in cost of computer technology have permitted the incorporation of new psychological attributes to the list of what is measurable; they have also brought about increased efficiency of applications and made it possible to include new functionalities, such as automatic item gen-

Correspondence: Julio Olea. *Facultad de Psicología. Universidad Autónoma de Madrid. Calle Iván Pavlolv 6. 28049 Madrid. España. E-mail: Julio.olea@uam.es*

eration, adaptive application of tests or automatic marking of complex responses.

- ✔ *New social demands*. In the case of Spain, though still some way short of that of many other countries, both psychology professionals and those in responsible positions in public and private organizations are increasingly putting their trust in the utility of tests to achieve certain applied objectives, as reflected in the article by Muñiz and Fernández-Hermida (2010) in this same issue. But not only have we seen an increase in "classic" tests such as the WAIS or 16FP; in a society ever more sensitive to the assessment of intervention outcomes and the accreditation of individual and institutional competences, there has been a huge increase in the types of psychological attributes required to be measured. Whilst a few short years ago the basic applications were confined to tests of cognitive capacities or tests of personality, more and more professionals are demanding reliable tests for a range of specific objectives.
- ✔ *Demand for greater quality*. The consequences of test scores for people and organizations are becoming increasingly important. Hence, the psychometric demands on test scores are also growing. The inevitable requirement of "measuring well" and the need to deal with particular problems in certain assessment contexts (such as faking of responses in selection contexts) are driving the development of new types of tests and new psychometric models for studying the guarantees provided by their application.

COMPUTERIZED TESTS

There are an ever-increasing number of tests whose items are presented, responded to and scored on computer, and this is part of a process of significant changes and progress in applied contests of psychological and educational assessment. For Davey (2005): "*In the last two decades computerized tests have gone from being an experimental procedure to being used by hundreds of assessment programs that assess millions of people each year*" ... "*being assessed by a computer may soon become even more natural than being assessed on paper*" (p. 358).

Strictly speaking, a computerized test must meet two requirements (Olea, Ponsoda, & Prieto, 1999): a) that the psychometric properties of its items are known, on the basis of a mathematical model, and b) that the items are presented and answered by means of a computer. The first of these requirements thus excludes from consideration as "computerized tests" many of those without the necessary guarantees available on the Internet.

Computers permit the application of tests in various forms. First of all, there are *static computerized tests*, in which the items are applied in the same sequence to all respondents. A second type are so-called *computerized adaptive tests*, which permit the presentation of the most appropriate items for each examinee. Given their importance, we shall devote a separate section to this latter type.

In general, computerizing a test has certain advantages:

- ✔ It helps to achieve better standardization of the application conditions of tests for all examinees: common instructions, control of application time, reduced possibilities for copying and leaking of information, efficiency in marking, and so on.
- ✔ It is necessary for applying the complex estimation procedures required in Item Response Theory (IRT) (see in this issue Muñiz, 2010), making it possible to apply new psychometric models with all their potential advantages.
- ✔ It permits the immediate provision of quantitative, verbal and graphic information on the position of a respondent with respect to a group on a given scale; in other words, it allows the production of automatic reports. It is also possible to continuously update the scales, incorporating the scores of new examinees.
- ✔ Computers are necessary to apply *new item formats* (e.g., dynamic visual presentations, auditory items or simulated sequences recorded on video), which have enabled a substantial expansion of the traits, competences and behaviours that can be assessed in psychology, to include, for example, musical ability, the performance of air traffic controllers, or conflict-solving skills (see Drasgow & Olson-Buchanan, 1999). Thus, the range of attributes that can be assessed is extended, increasing the similarity between the assessment task and the criteria to be predicted on the basis of the test scores (for example, activities to be carried out by the applicant in the job applied for). Moreover, tests can break with the traditional response format (multiple-choice or ordered categories), to include tasks as diverse as marking certain locations on a map, following the movement of an

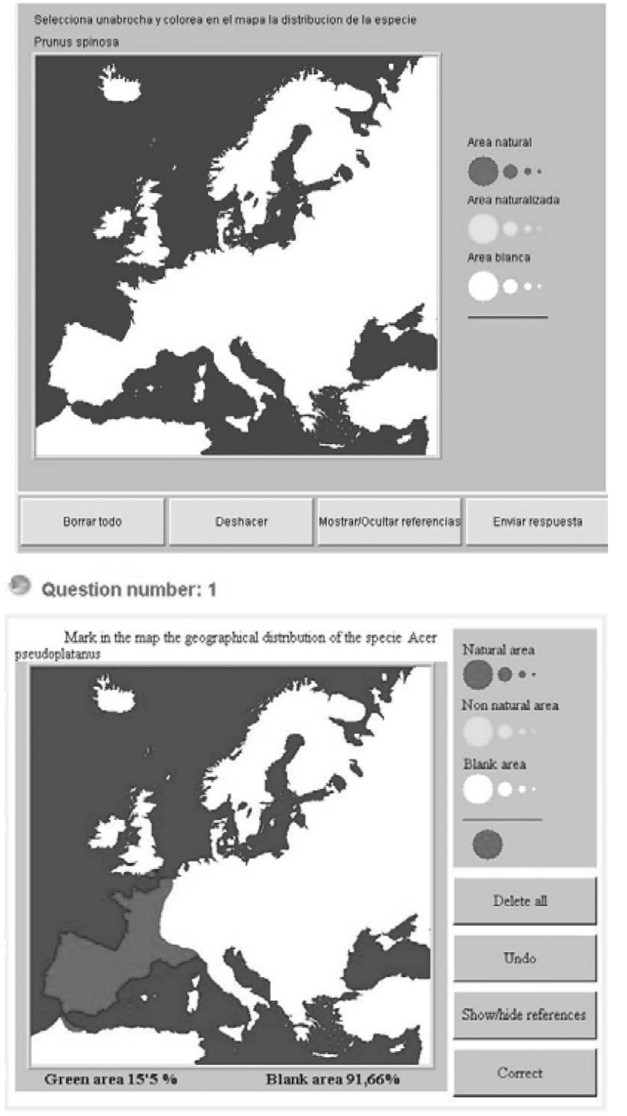
object using the mouse, rotating three-dimensional figures by degrees, detecting and changing grammatical errors in diverse types of texts, writing with an equations program the simplified result of a mathematical formula, recording a verbal response into a microphone, giving a medical diagnosis after collating diverse information on a patient's symptoms, or locating the architectonic components of a building. With these types of item, as well as recording right and wrong answers, the computer can measure other types of performance-related variables (such as reaction times or physical distances with respect to the optimum solution in a visuo-motor task).

✔ Some computerized assessment systems now permit the automatic marking of performance in a specific task. Figure 1 shows an example of an item on botanical knowledge, consisting in shading the regions of distribution of a given species, and whose correction is automatic (taken from Conejo, Guzmán, Millán, Trella, Pérez de la Cruz, & Ríos, 2004). For scoring this item the program uses a map with the correct shading as a template. If the student marks approximately the right region (with a certain margin of error), the response is scored as correct. Moreover, the proportion of area correctly identified is indicated. For example: 15.5% of the shaded area is correctly shaded and 91.66% of the non-shaded area is correctly non-shaded.

The recent work "*Automated Scoring of Complex Tasks in Computer Based Testing*" (Williamson, Mislevy, & Bejar, 2006) presents numerous examples of automatic marking in items with complex responses. The book recommends the use of *Evidence-Centred Designs* (ECDs), in which a schema is drawn up to be followed in this type of procedure. The ECD methodology starts out from a model of the examinee (exhaustive description of the constructs, abilities or skills to be measured) and a model of the task or family of tasks (with an exhaustive description of the task characteristics that permit the automatic generation of the item). The evidence-centred model connects the two models by including the relations between the respondent's performance in the task and the construct or decision about the respondent (e.g., suitable or unsuitable). In ECDs a distinction is drawn between evidence or scoring rules (which transform the examinee's task performance into numerical scores) and a measurement model (which connects the numerical scores with the construct scores and with the decisions to be taken on their basis).

One of the first important attempts to develop tests with automatic marking was the ARE (Architectural Registration Examination), an assessment battery with a substantial role in the accreditation process for architects to obtain a licence to practice in Canada. Some of the items require the examinee to use some basic functions of a computerized graphic design tool (see Figure 2). The applicant's task is to design a house, a clinic or any other type of building which fulfils a set of requirements. The designs produced by the examinee are scored

FIGURE 1



Selecciona unabrocha y colorea en el mapa la distribucion de la especie Prunus spinosa

Question number: 1

Mark in the map the geographical distribution of the specie Acer pseudoplatanus

Green area 15'5 %

Blank area 91,66%

automatically by an algorithm taking into account safety, functionality, the consideration of restrictions (geographical, environmental, climatic, etc.), accessibility, and so on. The development of these automatic procedures requires the collaboration of experts and the formation of discussion groups for designing an algorithm that provides scores similar to those a human assessor would provide. Paradoxically, although the experts provide the scoring rules incorporated in the algorithm, automatic marking may actually be more efficient than human marking, due to a more systematic application of the criteria. The ARE test is a performance assessment, and these are explored in detail in the article by Martínez-Arias in this same issue.

FIGURE 2

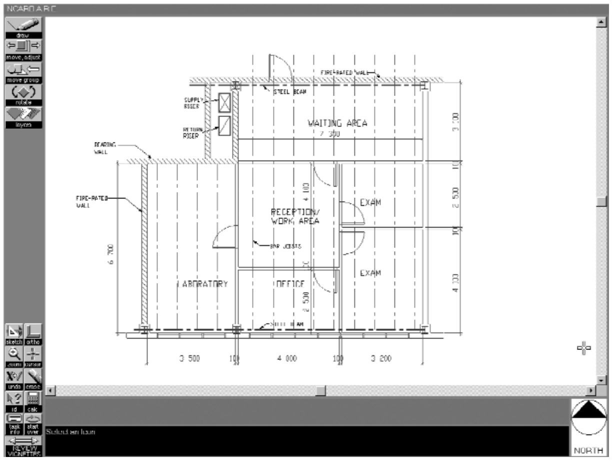
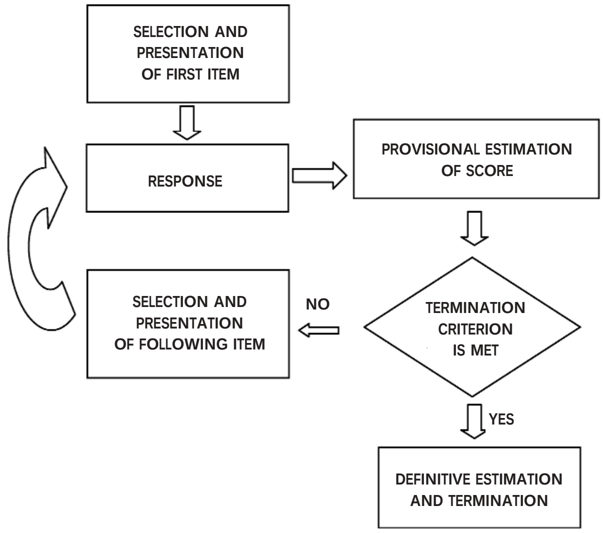


FIGURE 3


SELECTION AND PRESENTATION OF FIRST ITEM
RESPONSE
PROVISIONAL ESTIMATION OF SCORE
TERMINATION CRITERION IS MET
NO
SELECTION AND PRESENTATION OF FOLLOWING ITEM
YES
DEFINITIVE ESTIMATION AND TERMINATION

Computerized adaptive tests

The use of computers combined with IRT permits the construction of *computerized adaptive tests* (CATs), whose principal characteristic is that the items to be administered are continually adapted to the level of competence shown by the examinee according to his or her responses to the previous items. Starting out from a large item bank, different items from the bank are selected for each person. Thanks to IRT, ratings of the trait level obtained in different tests will be comparable (they will be in the same metric). The basic idea consists in presenting only those items which are highly informative for rating the level of each examinee in a given trait. Once the item bank has been calibrated, the process of applying a CAT to an examinee can be summarized, in simplified form, by a flow diagram like the one in Figure 3 (Olea & Ponsoda, 2003).

The application of a CAT begins with a particular starting-point strategy, which consists in establishing in some way the initial trait level assigned to the examinee (e.g., average level in the population). After each response from the examinee, his or her level in the trait is assessed by means of statistical Bayesian or maximum-likelihood procedures. An algorithm is also required for the successive selection of items. Generally, procedures based on the information function I(θ) are used; for example, a candidate for the second item would be that which is most informative for the θ level estimated after the first response. In contexts of accreditation, promotion or selection it is important to sample the content adequately or for examinees to receive, as far as possible, different items. In such cases, an appropriate selection algorithm must include restrictions in the rate of exposure to the items (for example, that no item is administered in more than 20% of the tests) or other restrictions, so as to guarantee an adequate sampling of content. Finally, some criterion for terminating the item presentation sequence is required, normally based on the attainment of a certain level of accuracy or on having applied a pre-set number of items; the latter tends to be necessary in order to maintain the balance of content in the test, and preferable for avoiding CAT users having the feeling that they have been assessed with few items. As shown in the diagram, the cycle “select item - apply item – register response – assess trait” is repeated until the termination criterion is met.

CATs, given their adaptive nature, have at least three important advantages with respect to any other computerized test:

- ✔ They improve the security of the test, since a large part of the items presented to examinees is different. This is one of the prime concerns for those responsible for assessment in applied contexts because, even when they opt for conventional tests, one of the major obstacles to the validity of tests is that examines can find out the test items in advance.
- ✔ They reduce application time (sometimes by more than half), as they attain similar levels of accuracy to those of conventional tests with a smaller number of items.
- ✔ They permit – with the same numbers of items as conventional tests – more accurate assessments. Under similar conditions to those of a conventional test (in time required and number of items applied) a CAT offers better guarantees (less measurement error) with respect to the trait levels assessed, and therefore with respect to the decisions made on the basis of test scores.

These three aspects are especially relevant where mass applications of performance or knowledge tests are carried out – for example, in contexts of personnel selection or educational assessment, or in tests for obtaining professional certification or licenses. To cite some examples, in the United States there are computerized adaptive versions of the TOEFL (Test of English as a Foreign Language), the GRE (Graduate Record Examination, for access to higher education), the GMAT (Graduate Management Admission Test, for access to Business schools), the ASVAB (Armed Services Vocational Aptitude Battery) and various examinations for professional accreditation (e.g., in Medicine and Nursing) or for level assessment in primary and secondary schools. In Spain, several CATs are available, such as the TRASI (Rubio & Santacreu, 2003), for measuring sequential and inductive reasoning capacity; the eCAT (Olea, Abad, Ponsoda, & Ximénez, 2004), which measures comprehension level of written English; and the CAT-Health (Rebollo, García-Cueto, Zardaín, Cuervo, Martínez, Alonso, Ferrer, & Muñiz, 2009), for the assessment of health-related quality of life; others are currently under construction for the assessment of examinees’ level in Catalan, Basque, other languages, emotional adjustment, satisfaction with health services, and so on.

Web applications

Information technology has for many years now permitted the *application of tests via Internet*. Examples of instruments applied in this way would be certain neuropsychological batteries, English language tests, predictive tests of job performance, academic school tests, personality questionnaires applied in clinical contexts or questionnaires on drug addiction (for a fuller treatment of this aspect, see Bartram & Hambleton, 2006).

Not only the test but also the presentation algorithms and the results are stored and distributed from a server, which gives more control over the application processes and immediate information about the results. Connection via Internet also brings significant logistical benefits: greater accessibility of examinees (for example, in recruitment processes for personnel selection or in cases of psychological intervention in individuals who live at some distance from the treatment services) and, in some cases, lower costs (consider, for instance, the application of tests to numerous examinees living in different regions of a country).

Application through the Internet is also advantageous for test publishers, as it gives them direct access to databases which make possible the essential studies of score validity and “monitoring” of the test’s psychometric properties. Moreover, it means test providers can ensure that the “client” (e.g., the company or institution commissioning the application) has access only to the pertinent information. For example, there is no need to include marking templates, and this gives greater guarantees of security.

However, the use of Internet as a “means of transport” for tests and for examinees’ responses means taking into account several possible sources of risk:

- ✔ *Quality*. Anyone can access hundreds of tests offered throughout the world, and whose psychometric properties are unknown. As in many other contexts, the competent psychologist should be capable of sorting out those assessment instruments available on the web that have truly shown their utility from those which serve merely as entertainment.
- ✔ *Security*. A substantial problem is that of the security of

the test itself, especially when test scores have important consequences for examinees (admission to an educational institution, securing a job, professional accreditation, etc.). In the case of the GRE, applied some years ago via the Internet, the company responsible for the test decided to return to pencil and paper versions after discovering the large numbers of items that applicants in certain Asian countries knew in advance, as a result of their being revealed in web forums. Obviously, access to the test content and the information provided by examinees must be secure and controlled. In some cases, moreover, the Internet can come into conflict with data protection legislation.

✔ *Control*. Another important problem concerns the possibilities of impersonation – that someone other than the named examinee takes the test in their place. A possible solution would be controlled application by supervisors who confirm the identity of the examinees, assign the appropriate passwords and control the fulfilment of application conditions.

✔ *Technical guarantees*. Computerized application can represent a threat to score validity if the assessment conditions are not standardized. For example, some tests which include dynamic information and limited response times are highly dependent on the speed of transmission of the information and the characteristics of the examinee's computer and connection.

On the other hand, it should be borne in mind that the properties of a test depend not only on the items applied but also on how they are applied (that the assessor creates an adequate assessment situation, that he/she responds to any doubts arising, that he/she ensures that examinees devote adequate time to reading the instructions, etc.). The need for a direct supervisor of the application may depend on the type of test (optimum performance vs. typical performance) and of the consequences of the assessment, among other aspects.

These and other problems have led to the drawing up of guidelines on good practice in the design and application of computerized tests, with specific recommendations for those applied via Internet (ITC, 2005), which involve additional quality control requirements. It is necessary to determine the minimum software and hardware requirements, set up mechanisms for the prevention and detection of errors in the administration, prevent and detect breaches of security, determine the level of supervision of the application, establish identification checks for examinees, guarantee the secure storage of responses, periodically check the psychometric properties of the items, and so on. In the more strictly psychometric context, the guidelines stipulate that a computerized test should incorporate the appropriate psychometric information (reliability and validity) and it should be guaranteed to require no extra knowledge or skills (such as familiarity with computers) on top of those demanded by the test itself. These guidelines can be consulted on the ITC website: http://www.intestcom.org/guidelines/index.php.

OTHER NEW TYPES OF TESTS

Model-based tests

One way of obtaining information on the inferences we can draw from test scores is to analyze the processes, strategies and knowledge structures involved in resolving items. Bejar (2002) use the term *model-based tests* to refer to the design of assessment instruments guided by a psychological theory on the processing of responses.

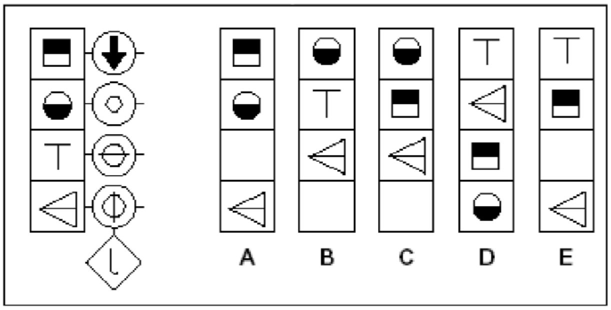
An excellent selection of these types of test is included in Irvine and Kyllonen's (2002) book "*Item generation for test development*", which explores the progressive rapprochement of Cognitive Psychology and Psychometrics, which has given rise to the development of tests of quantitative reasoning, analytical reasoning, visualization, verbal analogies, and so on. The first step in the construction of this type of test is an analysis of the cognitive processes needed for resolving the task and a detailed study of the characteristics of the item which, as a function of these processes, determine its particular level of cognitive demand and, therefore, its difficulty. For example, Hornke (2002) describes a figure-rotation test involving the manipulation of variables such as the number of elements to be processed, whether the figures are two- or three-dimensional, the angle of rotation or the number and type of rotations (from right to left, from top to bottom, etc.). The same author also describes a visual memory test in which the items are maps of a city on which there appear a series of icons representing public services, being manipulated in each case the number of icons, their size or their level of dispersion on the map.

In Spain, Revuelta and Ponsoda (1998) developed a test based on a cognitive model for the DA5 test. The 50 test items are designed to measure logical reasoning ability by means of tasks that include a set of instructions (symbols inside circles and a diamond) about what to do mentally with the corresponding adjacent figure (see Figure 4). An item consists of a series of figures (column of four squares on the left of Figure 4, each of which contains a figure), the instructions on the changes to be made to each figure (column of circles and diamond), and the five possible responses (columns A, B … E). The examinee's task is to apply the instructions to the figures and select the correct option from the five available. The instructions may require, for example, revolving the figure by a certain number of degrees, exchanging its position with that of the previous figure, omitting it, ignoring other instructions or rearranging all the figures in a certain way.

A processing model assumes that the examinee codes the first figure (that which appears in the first square of the first column) and the instruction, applies the instruction to the figure (in the example, the instruction indicates that the figure must be moved one square down, so that only options C and E could be correct), and follows on sequentially with the remaining figures until reaching the solution. Study of the contribution of each instruction (and of the number of times it is necessary to apply them) to item difficulty revealed that the instructions which required mentally rearranging the 4 figures of the items carried most weight in the prediction of difficulty.

A novel contribution of this way of working is that if we know which variables are involved in the response processes, a method can be established for constructing the entire possible universe of items governed by those variables. The procedure, called "*automatic item generation*" (AIG), consists in the construction of item banks by means of algorithms. In AIG a set of explicit rules is drawn up, capable of being programmed into a computer, which determines how the items should be constructed. For example, Revuelta and Ponsoda (1998) generated the 4,242 possible items whose basis is the DA5, combining type of figures, the instructions to apply and certain criteria for generating incorrect response options. If the model that describes the item response processes is correct, it will be possible to know the difficulty level of new items before they have been applied to anyone. The advantages of having available the entire bank of possible items are enormous, mainly in the sense of guaranteeing high accuracy in the measurement of any ability level.

FIGURE 4



Ipsative tests

In contexts of personnel selection above all, the faking of responses in personality tests is problem for which various solutions have been tried. One of the most encouraging developments in this regard has been the creation of ipsative tests, which oblige the examinee to choose between response options with similar levels of social desirability referring to different personality dimensions. For example, the applicant may have to choose between "I am a hard-working person" (responsibility) and "I am an outgoing person" (extraversion). The design process of an ipsative test is basically as follows:

a. Determine the dimensions to be assessed and the initial items that define them.
b. With these initial items, design a conventional normative test. It is recommended to carry out factor studies to determine empirically the items making up each dimension, and if necessary to eliminate those items that do not saturate in the proposed factor.
c. Establish the number of options for each ipsative item. The simplest approach is to establish binary items, each one made up of two initial items.
d. Carry out an empirical study in which an appropriate sample of judges rates the desirability level of each initial item. On the basis of these ratings, desirability values are obtained for each one of the initial items.
e. Design an ipsative test, considering that the items must include all the possible combinations of dimensions. Each ipsative item must include options (initial items) of similar desirability. Each dimension must be compared with any other a similar number of times.
f. Establish the scoring system for examinees – for example, counting the times they choose the options of each one of the dimensions.

Example of the development process of an ipsative test
(Abad, Olea, Ponsoda, & Garrido, 2007)

1) Dimensions to be assessed: the 5 personality dimensions defined in the Big Five Model, each one assessed by means of 18 adjectives.
2) Normative test: application of the 90 items to a sample according to a format of 5 order categories, asking for the extent to which each one describes the person.
3) Factorial study: the 12 items of each dimension showing the best saturation in the proposed factor were retained, so that the definitive test comprised 60 items.
4) Obtaining desirability indices (DI): a sample of people rated (from 1 to 4) the extent to which each adjective indicated a positive quality for being efficient in a given job. The means of these ratings were considered as indices of desirability of the items. The adjective with the lowest mean was “ordinary” (DI = 1.93) and that with the highest mean, “organized” (DI =3.87).
5) Design of the ipsative test: it was decided to construct a test with 30 ipsative items, each one made up of two adjectives from different dimensions and with similar DIs. For example, one of the items was “stable-energetic” which refer, respectively, to the dimensions of emotional stability and extraversion, and which obtained DI values of 3.71 and 3.43. In accordance with this design, each dimension was compared 3 times with the other 4 personality dimensions.
6) Score on the ipsative test: to score each individual on each of the 5 dimensions, the sum was obtained of the number of times in the pairs of adjectives the items of each dimension were chosen. Therefore, the theoretical maximum score in a dimension was 12, whilst the minimum was 0.
7) Studies of convergent and predictive validity were carried out (correlations with grades in training courses). The best predictive capacity was shown by some ipsative items that combined adjectives from the dimensions of emotional stability and responsibility.

Over the last couple of decades ipsative tests have had their ups and downs, with advocates and critics disputing with equal ferocity their pros and cons. The drawbacks with these tests include the following:

a. The method of scoring an examinee ipsatively in the different dimensions leads to interdependencies between them: a high score in one dimension necessarily entails low scores in the others. This problem is greater the smaller the number of dimensions. More generally, the average of the correlations between m dimensions approaches $-1/(m-1)$, m being the number of dimensions (Meade, 2004). In the case of measuring just two dimensions, the correlation between them would necessarily be -1. The ipsative form of scoring would lead, moreover, to the sum of the covariances of dimensions with an external variable (for example, a criterion) being zero, and to distortions in reliability coefficients for the scores in the dimensions. All of this implies the need for specific psychometric treatment of ipsative data (it is not unusual, for example for the factorial solutions of normative and ipsative data from the same test to be different), which is currently the object of research.

b. Conceptually, an ipsative test presents a preferences task, and hence permits comparison between scales within a single person (for example, it could be said that a person is more responsible than extraverted) but not between different persons (that one person is more responsible than another). Therefore, it is more suitable for use with measures of attributes that involve preferences, which is customary in the measurement of interests.

c. It is not clear that they are actually resistant to faking, since applicants may be aware of which dimensions are desirable for the position in question.

It does not seem to us advisable for now to apply ipsative tests if the aim is to make comparisons of performance between different examinees, given the difficulty of studying their psychometric properties by means of the usual models and techniques. Nevertheless, we consider these types of test to have great potential (some studies have already shown them to have greater predictive validity than traditional personality tests) once it becomes possible to theoretically model responses to this type of item, which is indeed the aim of some current work in the field of psychometric research (Stark, Chernyshenko, & Drasgow, 2005). In any case, the question is far from settled at the moment.

Behavioural tests

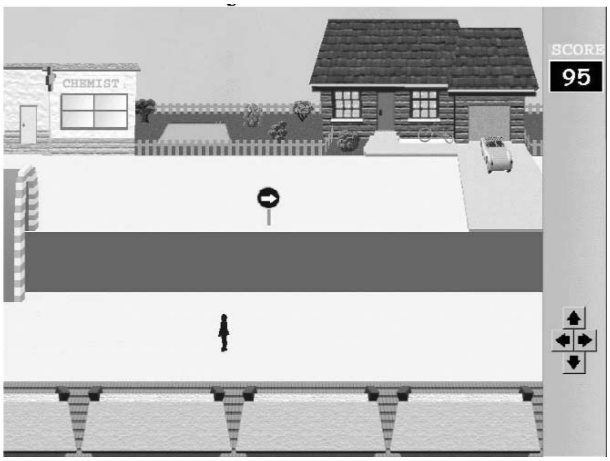
In the context of personality measurement, there is a theoretical approach to behavioural personality assessment involving the study of interactive styles or consistent behavioural tendencies in the face of certain situations (Santacreu, Rubio, & Hernández, 2006). From this perspective, computerized behavioural tests are designed for measuring, for example, tendency to take risks (propensity to choose more rewarding options despite their being improbable) by means of simulated roulette or darts games, or using tasks based on decisions that are more likely or less likely to lead to accidents. Figure 5 shows a task which involves deciding when to cross the road to get to the pharmacy as quickly as possible, changing the person's position in successive trials and in the knowledge that a car could come out of the tunnel. If the pedestrian is too far to the left, he or she has more chance of being run over (poorer visibility), but it would also take less time to reach the pharmacy. The safest way is to move to the right and cross, but that takes more time. After each trial, the examinees receive feedback about the time they have taken to reach the pharmacy, but not about whether they have been run over by a car. Tendency for risk is obtained by calculating the mean in successive trials of the distance between the person and the pharmacy (the higher the mean, the lower the tendency to take risks). Obviously, this procedure is very different from that of traditional personality tests in which respondents themselves provide information about their tendency to seek sensations or their level of openness. Professionals who opt for this type of test consider one of the advantages to be that they avoid problems of social desirability.

Situational tests

These consist in describing certain situations (e.g., in the job context) and asking examinees to say how they would react to those situations. It appears that these types of test increase predictive power about job effectiveness compared to traditional tests of cognitive ability and personality (which is why they are becoming more and more popular), though little research has been carried out on their effectiveness for reducing response faking. Examinees may provide descriptions in open response format or, as is more usual, choose between several predefined options. Below we present an example of a situational item on integrity (Becker, 2005). In brackets we show how the responses are scored, these values being set on the basis of experts' opinions:

> Your work team is in a meeting discussing how to sell a new product. Everyone seems to agree that the product should be offered to customers this month. Your boss is keen for this to happen, and you know that he doesn't like public disagreements. However, you have some reservations because a recent report from the research department identified various potential safety problems. What do you think your reaction would be?
>
> *A.*Try to understand why all the others want to offer the product to customers this month. Perhaps your concerns are unfounded. [-1]
>
> *B.*Express your concerns about the product and explain why you think the safety issues need to be addressed. [1]
>
> *C.*Show agreement with what the others want so that everyone feels good about the team. [-1]
>
> *D.*After the meeting, talk to some of the other team members to see whether they share your concerns. [0]+

FIGURA 5



From a psychometric point of view, a particularly relevant question is how to best score the response to these types of item. Bergman, Donovan, Drasgow, Henning and Juraska (2006) studied the different effects of 11 different ways of scoring the items of a situational test for assessing leadership capacity, made up of 21 items presented via video, and with four different response options depending on the degree of participation in decision-making.

SOME ADDITIONAL RISKS, SOME RESOURCES

While they should not be allowed to detract from the considerable advantages for psychologists of the new types of tests, it is appropriate to mention some of the potential risks associated with them.

First of all, it should be stressed that the new technologies are not in themselves guarantees of better measurement. The efficiency of the new response and information-processing procedures cannot substitute the necessary psychometric scrutiny of the scores assigned. The appearance of validity of the new item formats must be accompanied by empirical evidence of validity. For example, a "multimedia" item may be more informative than a classical multiple-choice item, but much more time may be required for its resolution. Also, the accuracy required may be different if the aim is to classify a person (suitable vs. unsuitable) or if it is to quantify his or her trait level. Likewise, it is necessary to reflect on when it is worth applying a CAT and when it is not (Wainer, 2000). For example, if the test has significant consequences for examinees, it will be applied once or twice a year and the content does not require computerized application, the cost of a CAT (need to create and maintain large item banks, to develop, assess and update software, availability of computers for the application, etc.) may exceed the benefits.

A second risk has to do with the possibility that certain areas of test application are overlooked. Developing new tests is costly, and there is a risk of progress being made almost exclusively in applied contexts (organizational or educational) in which more economic resources are invested or efficient technological solutions are in higher demand. But advances in this area should not ignore certain more strictly "psychological" measurement contexts, such as clinical assessment or the evaluation of psychosocial intervention programmes. In this regard, the conjunction of new types of tests, new psychometric models and statistical models for measuring the changes obtained by interventions should be seen as fruitful territory for R+D projects.

A third risk concerns the improper use of the new tests. Given their immediate availability, the new types of tests might easily be applied in inappropriate contexts and by untrained or unqualified personnel, and erroneous inferences may be drawn from the scores they provide.

What can psychology professionals do to increase their competence in the new forms of measurement? More than ever before they need ongoing education and training in these aspects throughout their professional career to keep up to date with the increasingly frequent innovations aimed at improving psychological measures. Reluctant as we are to give advice, we might suggest beginning with some of the recent books on psychometrics and the specialized journals which publish articles and reports on advances in psychometrics and experiences in the application of new types of tests (the Spanish journals most likely to focus on these issues include *Psicothema*, *Psicológica*, *Revista Electrónica de Metodología Aplicada* and *Spanish Journal of Psychology*). In 2000, *Psicológica* published a special issue on CATs. Information in Spanish on CATs can be found in books (Olea & Ponsoda, 2003; Olea et al., 1998) and book chapters (Olea & Ponsoda, 1996); there are many books in English on computerized tests (Bartram & Hambleton, 2006; Eggen, 2004; Mills, Potenza, Fremer, & Ward, 2002; Parshall, Spray, Kalohn, & Davey, 2002; Sands, Waters, & McBride, 1997; van der Linden & Glas, 2000; Wainer, Dorans, & cols., 2000).

A simple tutorial on CATs is available at http://edres.org/scripts/cat/catdemo.htm and an essential site for researchers, offering extensive theoretical and applied information on CATs, is the following: http://www.psych.umn.edu/psylabs/catcentral/.

The interested reader may also consult the catalogues of tests available on the websites of the principal publishing companies. Professionals interested in improving their education in this area can also attend specialized courses – many are on offer at Spanish universities. A further possibility is to begin working with some of the available programmes for the development and analysis of computerized tests. If this is the option you take, you should make sure to consult the following site, run by the University of Málaga (http://jupiter.lcc.uma.es/siette.wiki.es/index.php/Portada) or that of the principal distributor of psychometric software in North America: http://assess.com. In addition to the general software for the application of IRT, there is software for implementing CATs (FASTEST and POSTSIM 2.0); whilst the former (ASC, 2001) permits the organization of item banks, test assembly and application of CATs, the latter allows assessment of the psychometric functioning of a CAT by means of simulation and under different conditions (of item selection, of ability level estimation and of termination criterion). For a "first contact" the ADTEST program (Ponsoda, Olea, & Revuelta, 1994) is also worth a look.

REFERENCES

Abad, F.J., Olea, J., Ponsoda, V. & Garrido, L. (2007). *Test POLIPSA: Informe técnico y propiedades psicométricas* [*POLIPSA Test: Technical report and psychometric properties*].

ASC (2001). The FastTEST *Professional Testing System, Version 1.6.* [*Computer software*]. St. Paul, MN: Author.

Bartram, D. & Hambleton, R. K. (2006). *Computer-based testing and the internet issues and advances.* Chichester, West Sussex: Wiley.

Becker, T. E. (2005). Development and validation of a situational judgment test of employment integrity. *International Journal of Selection and Assessment, 13*(3), 225-232.

Becker T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13*, 225–232.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Mahwah, NJ: LEA.

Bergman, M.E., Drasgow, F., Donovan, M.A., & Henning, J.B. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14*, 223-235.

Conejo, R., Guzmán, E., Millán, Trella, M., Pérez-de-la-Cruz, L. & Rios, A., (2004): SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education, 14*, 29-61.

Davey, T. (2005). Computer-based testing, In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Hoboken, NJ: Wiley.

Drasgow, F. & Olson-Buchanan, J. (1999). *Innovations in computerized assessment*. Mahwah, NJ: LEA.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Arnhem, Holland: Citogroep.

Hornke, L. F. (2002). Item-generative models for higher order cognitive functions. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development.* New Jersey: Lawrence Erlbaum Associates.

International Test Commission (2005). *International Guidelines on Computer-Based and Internet Delivered Testing*. Retrieved 30th June 2005, http: / / www.intestcom.org.

Irvine, H. & Kyllonen, P.C. (2002), *Item generation for test development*. New Jersey: LEA.

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531-552.

Mills, C. N., Potenza, M. T., Fremer, J. J. & Ward, W. C. (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: LEA

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems [Test theory: Classical Theory and Item Response Theory]. *Papeles del Psicólogo, 31*(1), 57-66.

Muñiz, J. & Fernández-Hermida, J.R. (2010) La Opinión de los Psicólogos Españoles sobre el Uso de los Tests [The Opinion of Spanish Psychologists on the Use of Tests]. *Papeles del Psicólogo, 31*(1), 108-122.

Olea, J. & Ponsoda, V. (1996). Tests adaptativos informatizados [Computerized Adaptive Tests]. In J. Muñiz (Coord.), *Psicometría*. Madrid: Universitas.

Olea, J. & Ponsoda, V. (2003). *Tests adaptativos informatizados* [*Computerized Adaptive Tests*]. Madrid: UNED.

Olea, J., Ponsoda, V. & Prieto, G. (1999). *Tests informatizados: Fundamentos y aplicaciones* [*Computerized tests: Foundations and applications*]. Madrid: Pirámide.

Olea, J., Abad, F. J., Ponsoda, V. & Ximenez, M. C. (2004). A computerized adaptive test for the assessment of written English: Design and psychometric properties. *Psicothema, 16*, 519-525.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Ponsoda, V., Olea, J., & Revuelta, J. (1994). ADTEST: A Computer adaptive test based on the maximum information principle. *Educational and Psychological Measurement. 54*, 3, 680-686.

Quintana, J., Bitaubé, A. & López-Martín, S. (2008). *El lugar de la Psicología en la universidad española del siglo XX* [*The place of Psychology in Spanish universities in the 20th century*]. UAM Ediciones: Madrid.

Rebollo, P., García-Cueto, E., Zardaín, J.C., Martínez, I., Alonso, J., Ferrer, M. & Muñiz, J. (2009). Desarrollo del CAT-Health, primer test adaptativo informatizado para la evaluación de la calidad de vida relacionada con la salud en España [The development of CAT-Health, first computerized adaptive test for the assessment of health-related quality of life in Spain]. *Medicina Clínica, 133*, 7, 241-251.

Revuelta, J. & Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems [A computerized adaptive test of logical analysis based on automatic item generation]. *Psicothema, 10*, 3, 709-716.

Rubio, V. & Santacreu, J. (2003). *TRASI. Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción como factores de la habilidad intelectual general* [*TRASI. A computerized adaptive test for the assessment of sequential reasoning and induction as factors of general intelligence*]. Madrid: TEA ediciones.

Sands, W. A., Waters, B. K. & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Santacreu, J., Rubio, V.J., & Hernández, J.M. (2006). The objective assessment of personality: Cattell's T-data revisited and more. *Psychology Science, 48*, 53-68.

Stark, S, Chernyshenko, O.S., Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*, 184–203.

van der Linden, W. J. & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. London: Kluwer Academic.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. & Thissen, D. (2000). *Computerized adaptive testing: A primer (2nd ed.)*. Mahwah, NJ: LEA.

Wainer, H. (2000). CAT: Whether and Whence. *Psicologica, 21*, 121-133.

Williamson, D. M., Mislevy, R. J. & Bejar, I. I. (2006). *Automated Scoring of Complex Tasks in Computer Based Testing*. Mahwah, NJ: LEA.