

The technique of objective personality-tests sensu R. B. Cattell nowadays: The Viennese pool of computerized tests aimed at experiment-based assessment of behavior

Klaus D. Kubinger

(Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Austria)

Abstract: Considering the current scene of personality testing, in particular in China, discloses that it is primarily personality questionnaires that are being applied, but that so-called Objective Personality-tests *sensu R. B. Cattell* (or, in more modern terminology: tests for experiment-based assessment of behavior), which have lately been experiencing a certain “renaissance” within German speaking countries, are almost not in use. For this reason, a review is given of their typical properties and principle advantages over personality questionnaires. They are primarily less fakeable. As an illustration, several such tests stemming from a Viennese research group are presented, and their psychometric qualities and shortcomings discussed. Finally, several applications of these tests already in practice are listed.

Key words: Objective personality test; psychological assessment; personality questionnaire; unfakeability; computerized testing

Introduction

Recently, Zhou and Zhang (2008) summarized that the use and development of personality testing in China was initially based on an introduction of US-American/British personality questionnaires like the MMPI (Song, 1985) and the MMPI-2 (Zhang, Song, & Cheung, 1999), Eysenck’s EPQ (Chen, 1983), Cattell’s 16 PF (Zhu & Dai, 1988), and the NEO-PI-R (Dai & Wu, 2005) – see for an even earlier report Zhang (1988). As some of those items are obviously not suitable within the Chinese context, the second wave of personality test development in China incorporated different aspects of the traditional Chinese culture (cf. Zhang & Zhou, 2006; Cheung, 2009), resulting, for instance, in the *Qing nian Zhong guo Personality Scale* (cf. Wang & Cui, 2003) – and see for some detailed reported indigenous measures Cheung, Cheung, Wada, & Zhang (2003) and Cheung,

Fan, Cheung, & Leung (2008).

That is, within China there is a focus on personality testing that uses personality questionnaires, which are based on testees’ self-evaluation (cf. Cheung, 2004). Although it is possible that all the problems of personality questionnaires established within so-called Western culture (in particular, the USA and Europe) do not apply to the Chinese culture, this does not seem very likely. In any case, there are no empirical results supporting such a hypothesis (cf. Chen, Lee, & Stevenson, 1995; Hamid, Lai, & Cheng, 2001; Yang, Bagby, & Ryder, 2000). Actually, there is primarily the problem of the fakeability of personality questionnaires. And the phenomenon of “alexithymia,” or in other words the inability to properly recognizing or describing one’s own feelings, also has an effect; not to mention self-deceptive tendencies where people misjudge their attitudes out of naïveté or pig-headedness.

Received: 2009-05-28

Correspondence concerning this article should be addressed to: Klaus D. Kubinger, PhD, Head of the Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria, Europe. Email: klaus.kubinger@univie.ac.at

Concerning the fakeability of personality questionnaires, numerous (Western) experiments with so-called faking instructions have proven that such questionnaires can be faked to one's personal benefit as a result of the "transparency of their measurement intentions and measurement principles" (face validity). One of the most cited overviews on this topic is Viswesvaran and Ones (1999), but see also Viswesvaran and Deller (2006), Kubinger (2002), and Seewald (2002). Additional experiments with persons, for whom the relevant test results would have specific consequences, prove that testees do actually fake the test (see again Kubinger, 2002).

For these reasons, it is preferable to use such means of personality testing which are not based (as strongly) on the testee's self-concept. The primary alternative genre which comes to mind here is that of so-called "Objective Personality-tests" based on the work of *Raymond B. Cattell* (cf. Cattell & Warburton, 1967), though an almost equivalent approach by Cronbach (1970), "performance-tests of personality," should also be mentioned.

Although Cattell's many suggestions for the conceptualization of such tests were not originally implemented for consulting in practice (one reason being the fact that some of the tests were still questionnaire-like in that they asked about subjects' self-concepts), within German speaking countries a kind of "Renaissance of Objective Personality-tests" has recently taken place (see Kubinger, 1997, and, in particular, Ortner, Proyer, & Kubinger, 2006). This is because, in the meantime, the development of tests of this genre focuses on establishing a new (psychological) "technology" which is inseparably linked to the innovative possibilities of the computer. Consequently, there is currently a new definition of how the respective tests do work, and in particular there is a new but more appropriate label for this genre: "The *experiment-based assessment of behavior* as a (psychological) 'technology' refers to techniques that deduce personal characteristics of style from observable behavior on experimentally varied performance tasks, whereby the computer registers the way in which the testee deals with the problem." (Kubinger, 2006, p. 256 – translated by the author).

In order to comprehend the typical advantages of this genre's approach and to illustrate the given definition, an exemplification of tests is needed. To this end this paper gives a review of a representative sample of respective test developments, some of them published and already in daily use by psychological practitioners. As a matter of fact, they actually incorporate – so to speak – a Viennese test pool, as they all stem from the Division of Psychological Assessment and Applied Psychometrics of the University of Vienna.

Representative examples of tests for experiment-based assessment of behavior

The most well known and most widely used product in practice is probably *Work Styles* (Kubinger & Ebenhöf, 1996, 2002¹). *ILICA* (Möseneder & Ebenhöf, 1996; "Ilica" is an invented name) is in fact a computer simulation, but nevertheless a typical Objective Personality-test. Then there is *BACo-D* (translated as: *Stress Resistance Assessment: Computerized Objective Test-Battery*; Ortner, Kubinger, Schrott, Radinger & Litzenberger, 2006). With *LEWITE* (translated as *Lexicon Knowledge Test*; Wagner-Menghin, 2004), the author created an entirely new concept in psychological assessment, namely that of "multifunctional testing," in other words measuring "personality" whilst testing achievement performance. Several other Objective Personality Tests have been constructed, tested (in some cases multiple times), and introduced to the scientific community in journals or at scientific meetings, though they are still not used routinely nor marketed for publication despite their fundamentally useful test concepts: *LAMBDA* (translated as: *Learning by Heart, Memory, Stress Resistance, Reasoning*); *Rigidity Test*; *Repressor-Sensitizer Test*; and *EUcO* (translated as: *Decisions in the case of Uncertainty: Computerized Objective Test-Battery*). The same is true as concerns *IACo* (translated as: *Interests Assessment: Computerized Objective Test-Battery*),

¹ Apart from its original in German, the following language versions are available: Arabic, Bulgarian, Chinese, Czech, Dutch, English, Finnish, French, Hungarian, Italian, Portuguese, Romanian, Slovakian, Spanish, and Swedish.

although its test concepts are partially published in MOI (translated as: *Multi-methodical Objective Test-Battery of Interests*; Proyer & Häusler, 2008): It is not explicitly personality, but rather interests, which are measured.

Each of these will be introduced, but in the case where there are several subtests, only one or two of them will be illustrated for typical characterization. After their introduction they will be discussed with respect to their use for consulting in practice and their psychometric qualities and shortcomings as a pool.

Work Styles

Here only the subtest “Coding symbols” will be illustrated, but there are two more subtests, “Comparison of surfaces” and “Figure differentiation,” which measure impulsiveness/reflectiveness and achievement motivation.

The basic nature of the task in this subtest, as well as the idea of presenting this task several times in conjunction with respective feedback and with the request that the testee give a prediction of his/her next test performance, is taken from Cattell’s original test-battery. In five stages, each with a 50-second time limit, the testee has to match four different abstract black-white figures, which are continuously shown on the screen in a random order, with the appropriate symbol (again an abstract, but colored figure) by using a coding scheme provided (cf. Fig. 1). Feedback on the number of (correctly) coded symbols is given from the second stage onwards, where the testee is asked to predict the number of codings he/she expects to achieve in the next stage. What is

clearly different in comparison to Cattell’s test-battery, from an innovative perspective (and realizable only by using a computer), is that from the third stage onwards, after deliberate feedback and before the required prediction, the testee is exposed to a frustration condition: he/she is wrongly informed that on average other people have yielded a 10% higher performance on this task (compare the schematized test procedure in Fig. 2)

This subtest measures, on the one hand, the testee’s aspiration level by calculating the difference between actual performance and predicted performance, and, on the other hand, the frustration tolerance by calculating the difference between the initial and final prediction.

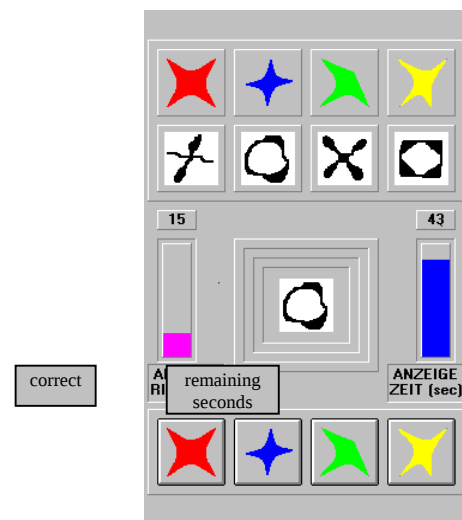


Figure 1: Task from the subtest “Coding Symbols” in *Work Styles* (Kubinger & Ebenhöf, 1996). The upper area provides the coding scheme with different abstract black-white figures in the upper row and abstract, but colored figures in the row below it; on the bottom are the reaction buttons for the testee, in the middle one randomly shown stimulus figure.

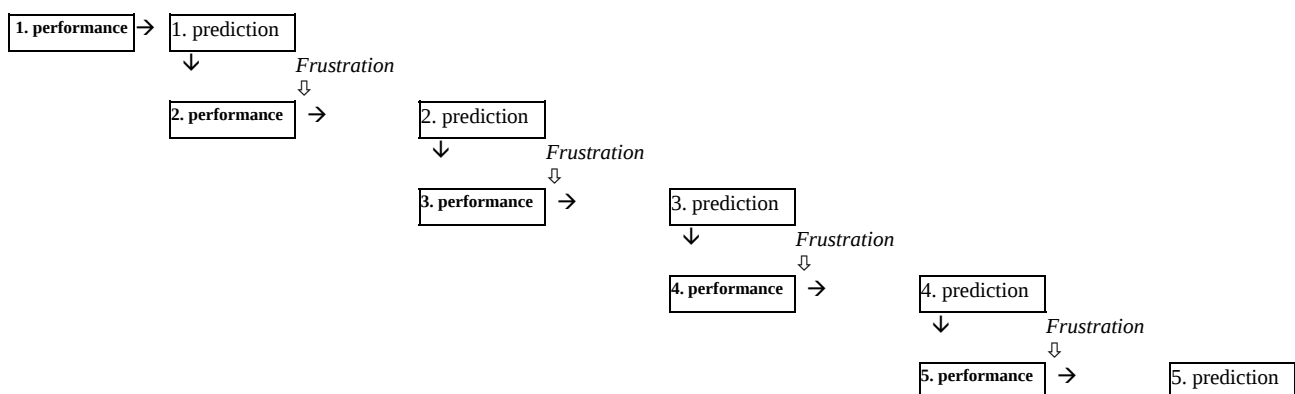


Figure 2: Schematized test procedure from the subtest „Coding Symbols“ from *Work Styles* (Kubinger & Ebenhöf, 1996).

ILICA

So-called “computer simulations” try to measure the behavior of a testee as realistically as possible by using more or less complex job-related tasks. That is, they follow earlier “work samples”. This test now endeavors to test the problem-solving and decision-making behavior of a person with rather simple tasks – from a content point of view. On the one hand, the testee is given various options to “virtually” arrange a “day off” (e.g. different leisure activities; cf. Fig 3), on the other hand, there are a few obligations that he/she has to fulfill (e.g. visiting his/her sick mother in the hospital). In addition, the testee comes across

various “surprises” in the course of testing which influence the day’s routine (e.g. a neighbor asks if the testee can look after her children).

Amongst other things, the number of activities performed in relation to the number of planned activities, as well as the number of goal-directed vs. socially related reactions in situations which conflict with planning and are subject to external influences, measure the way in which the testees conduct themselves. In other words, it is a question of testees’ typical planning strategies and their problem-solving behavior in organizational, less demanding tasks, and also their distractibility whilst trying to carry out a plan in reality.

Introduction

Holiday destination

Free day

You can select up to 15 goals for the day (notes).

- ☐ visit mother in hospital (broken leg)
- ☐ collect brochures
- ☐ go for a stroll
- ☐ read the newspaper
- ☐ find literature on holiday destination
- ☐ sit in the park
- ☒ speak to Susi (who was in Illica the year before)
- ☒ buy films
- ☐ buy sports equipment
- ☐ have a ride on the bicycle
- ☐ buy a language guide and read a bit
- ☐ take a nice long bath
- ☐ go to the hairdressers
- ☐ consider how to finance the holiday
- ☒ arrange to meet Doris tomorrow
- ☐ write letters
- ☐ put together a list of what to pack
- ☐ listen to music
- ☐ water flowers
- ☐ take care of health insurance for the holiday
- ☐ wash the dishes from yesterday
- ☐ prepare medicines
- ☐ buy shoes for the holiday
- ☐ hang up new pictures
- ☐ buy birthday present for your sister
- ☐ vacuum-clean the bedroom
- ☐ talk to your sister about looking after the plants
- ☐ finish reading the exciting murder mystery
- ☐ repair broken lamp
- ☒ go swimming
- ☐ look for tenant for apartment
- ☐ feed the dog
- ☐ book flight
- ☐ play with Claudia (neighbour's child)
- ☐ go to bank (change money, travellers checks)
- ☐ have a short (midday) nap
- ☒ pick up passport application (police,...)
- ☐ drive to Hans and Gabi (invitation)
- ☐ go to the Institute for Tropical Medicine
- ☐ go to dentist

Figure 3: Schedule of leisure activities that the testee can do a “virtual day off” from ILICA (Möseneder & Ebenhöf, 1996).

BACo-D

Only one of the six subtests – “Hindrance of scheduled course of action” – shall be briefly outlined. The five other subtests “Time pressure,” “Working under observation,” “Task collision,” “Work with inappropriate feedback,” and “Unfavorable working conditions” measure respective stress resistances.

The testee, who has a bird’s eye perspective of a labyrinth on the screen, has the task of moving a figure from start to finish with the help of the mouse (or with help of a navigation panel: up – down – left – right). However, the most direct paths keep closing off spontaneously at regular intervals until the testee has overcome, i.e. avoided, a certain pre-set number of such hindrances.

The ability to work under pressure as concerns hindrance of attempted actions is measured using the relative number of hindrances overcome within a

fixed amount of time.

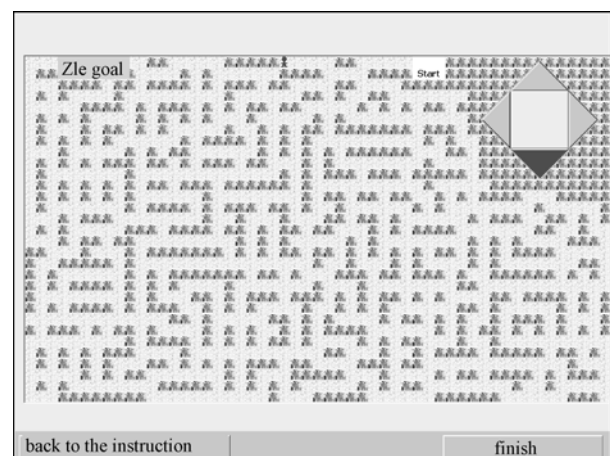


Figure 4: Subtest “Hindrance of scheduled course of action” from BACo-D (Ortner et al., 2006). The testee has to move the figure towards the goal with the help of a navigation panel, but the most direct paths keep closing off spontaneously until the testee has overcome a certain number of hindrances.

LEWITE

In the first part of the test, the testee is presented with a list of concepts (words) and has to then determine whether he/she knows what they mean and also whether he/she can explain them. In the second part of the test, which examines lexicon knowledge (crystallized intelligence) and is therefore a performance task, the testee must master items with respect to each of the same concepts by correctly filling in blanks in a text in order to come up with a correct definition of the concept in question (cf. Fig. 5).

A person's individual style of overestimating him- or herself on performance tasks is measured by comparing the self-assessed performance of knowing a certain concept or not, and the actual performance.

Please now try to complete the sentence

The alphabet refers to of the .

Possible answers:

- given order
- young convolution phase
- quick reaction
- alternative sequence
- characters of a language
- formation of mountains
- digits and letters
- alpha particles in atom nucleus

Next >>

Figure 5: Item from the second part of LEWITE (Wagner-Menghin, 2004). The testee has to correctly fill in blanks in a text in order to come up with a correct definition of the concept in question.

IaCO

It should suffice in the following to illustrate a combination of two subtests: “Observation Time” and “Memory.”

The first subtest consists of pictures with lots of details which are presented for as long as it takes the testee to “look at everything” (cf. Fig. 6). The themes of the pictures have been constructed according to *Holland's Hexagon model* (Holland, 1997). In the next subtest, the testee has to answer questions related to specific contents in these pictures, which are no longer shown.

The length of the observation time for each picture theme, as well as the number of correct answers relating to the content for each picture theme, indicates the main focus of the testee's interests. As indicated above, this test focuses rather on interests and motives than on personality traits.

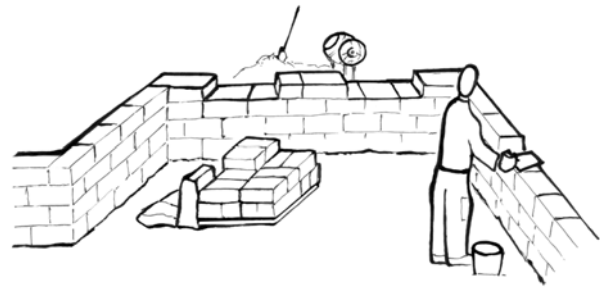


Figure 6: Item from IaCO (unpublished) referring to practical-technical interests. The picture is presented for as long as it takes for the testee to “look at everything”.

LAMBDA

This test-battery requires the testee to perform under stress: the reasoning subtest “Calculating in symbols” is first presented in a neutral condition. Afterwards, halfway through the items, the testing phase from the previous learning aptitude subtest “Organizational chart” is inserted as a stress factor while the testee is attempting to solve each individual reasoning item. The subtest “Organizational chart” is about learning a schematized structure of a business company by heart; whether or not the learning process is successful or not is tested by presenting an organizational chart which is only partly correct and which the testee then has to correct (cf. Fig. 7).

By comparing the actual performance on those items in the subtest “Calculating in Symbols” with and without the additional disruptive tasks, (a specific aspect of) a testee's ability to work under pressure is measured – in this respect, LAMBDA is a precursor to BAcO-D.

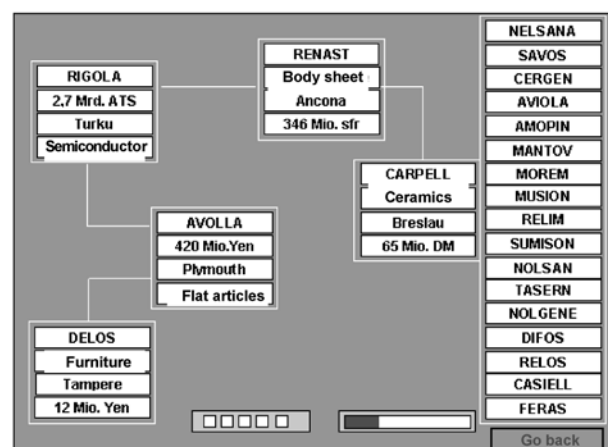


Figure 7: Only partly correct “Organizational chart” from LAMBDA (unpublished). The testee has to correct the wrong elements with help of the right-hand column.

Rigidity Test

It should suffice here to describe only one of the six items.

Based on *Luchins'* (cf. Luchins & Luchins, 1959) well-known water jar problems, the test first presents sub-items, whose solution always results from the same, relatively complex formula (for example, according to: $\text{goal} = B - A - 2 \times C$, with goal, A, B, and C being concrete but different numerical jar volumes; cf. Fig. 8). After such sub-items, other sub-items are presented that can still be solved according to the (already learned) formula, but which can alternatively be solved according to another, much simpler solution formula (e.g. $\text{goal} = A - C$).

A	B	C	→	Goal
11	163	75	→	2
36	86	22	→	6
28	58	6	→	18
20	60	5	→	30
46	98	6	→	40
45	117	9	→	54
36	96	8	→	44
7	18	4	→	3

Figure 8: Item from the *Rigidity Test* (unpublished). Every sub-item has as a solution, the relatively complex formula $\text{Goal} = B - A - 2 \times C$, however starting with the fifth sub-item there is always a much simpler solution formula, too.

The number of these sub-items that are solved with the alternative solutions and not according to the (already learned) formula, is meant to measure a person's flexibility in thinking.

Repressor-Sensitizer Test

This test is based on the perception experiment on the phenomena of perceptual defense by Bruner and Postman (1947) in which individuals resulted who, on the one hand, typically perceive emotionally filled stimuli more slowly than neutral stimuli, and individuals resulted who, on the other hand, perceive emotionally filled stimuli faster than neutral stimuli. Universally neutral and universally (negative) emotional pictures are presented, which are depicted in continually refining grids so that they become more and more easy to identify (cf. Fig. 9). The testee stops further refinement of the grids as soon as he/she believes to have identified the content of the picture.

The construct "Repressor vs. sensitizer" – in other words, the individual (cognitive) style of a person to either repress (negative) emotional content or to purposefully search for it – is measured by comparing the time it takes for a testee to correctly identify (negative) emotional content of the picture as opposed to neutral content.

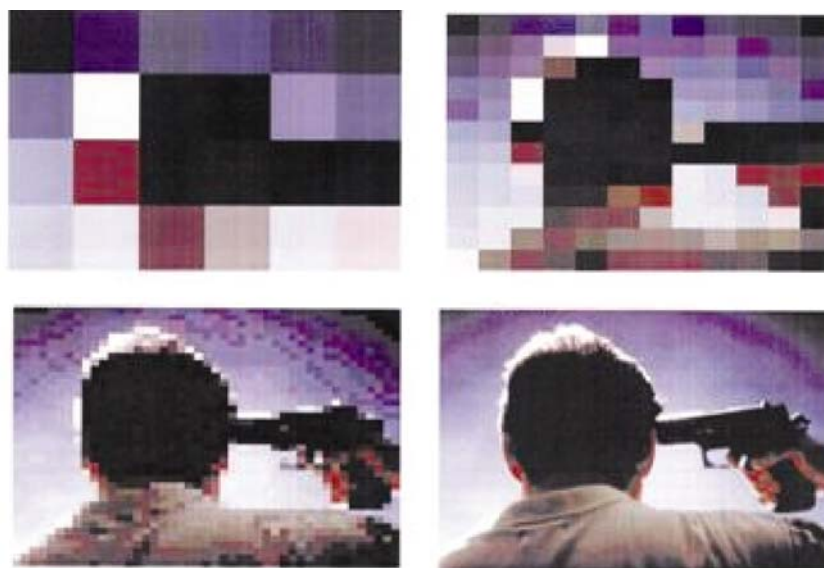


Figure 9: Four different refined grids of an item from the *Repressor-Sensitizer Test* (unpublished) – from up-left to right-down the grades of grids 1 (lowest), 3, 8, and 15 (highest). The testee stops further refinement of the grids as soon as he/she has identified the content of the picture.

EUcO

On the one hand, this test is designed to measure the ability of a person to assess risk (intuitively, from experience and by cognitive anticipation). Typical tasks from the theory of probability following the so-called “Bayes-theorem” (Bayes, 1764) are presented. That is, for each item, the testee has to decide from which of two containers with visible black and white balls a randomly drawn (black or white) ball has most probably been taken. On the other hand, the testee has specific options how to

behave in this situation or the option to gather more or less relevant information. These options are, among other things: asking for help, delegating the decision-making to the computer, or obtaining the advice of an expert. However, there is also the option of postponing or avoiding the decision (cf. Fig. 10). As a matter of fact, not all options are available each time.

The type and frequency of the options used is meant to measure a testee’s typical decision-making behavior in situations with uncertain parameters.

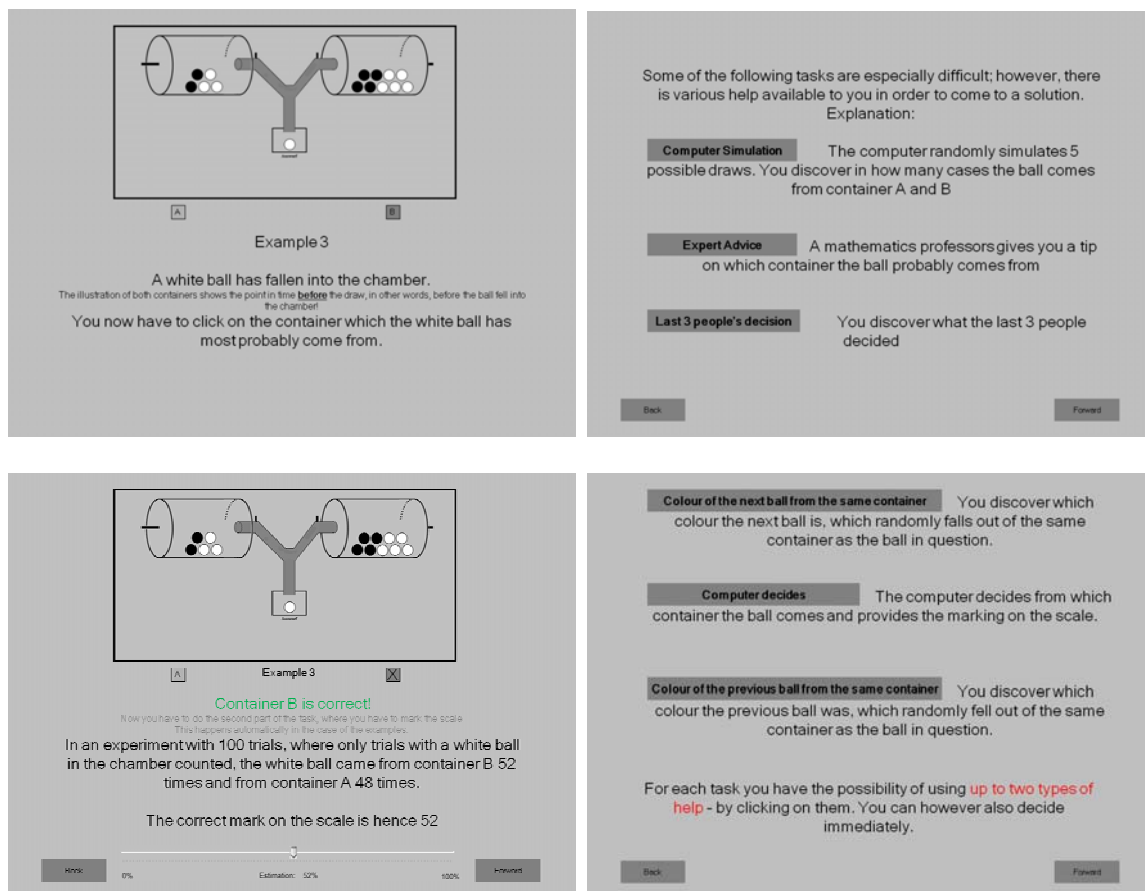


Figure 10: Screen-shots of the instruction from EUcO (unpublished).

The characteristics of tests for experiment-based assessment of behavior

What is common to all the tests mentioned is that they always demand a certain performance. Whether it is coding as many symbols as possible (*Work Styles*), carrying out certain tasks (*ILICA*; *BACO-D*), explaining as many concepts as possible (*LEWITE*), perceiving and/or learning certain con-

tents as quickly as possible (*IACO*; *Repressor-Sensitizer Test*) or solving as many numerical problems as possible (*LAMBDA*; *Rigidity Test*; *EUcO*), the tests always involve the examination of performance. The testee’s actual behavior before, after, or while performing the presented tasks – particularly the behavior that goes beyond the quantity/quality of the performance delivered – is also always “observed” by the computer. This refers to self-defined

goals (*Work Styles*, *ILICA*, *LEWITE*), the change in performance as a result of experience or stress (*Work Styles*, *BACo-D*, *LAMBDA*, *Rigidity Test*), the variation of performance determined by the contents (*IACo*, *Repressor-Sensitizer Test*), and the use of the offer of assistance or task avoidance (*EUCO*). Of course, certain information is again only inquired by questions, that is specifically about some self-defined goals as well as about the use of the offer of assistance or task avoidance. What is additionally common to all the tests mentioned is that they are administered at a computer. For these tests, a computer is not only a practical, but also a suitable medium. *LEWITE* could also possibly be administered as a paper-pencil test, as could the *Rigidity Test*. All other tests use the technical possibilities of the computer, namely for the application of specific test materials, for the (accompanying) measurement of the item-reaction times and especially for the necessary computing power in order to determine a test value.

What is less obvious, but nevertheless essential for the tests mentioned, is that all the tests include the manipulative element of psychological experiments; the conditions under which the testee has to generate the required performance are systematically varied by the test author according to theory. The feedback on the performance generated is controlled accordingly (*Work Styles*, *BACo-D*), stressors are intermittently inserted (*ILICA*, *LAMBDA*), certain expectations are systematically established in the testee in order to then present requirements that contradict these expectations (*LEWITE*, *Rigidity Test*), or various other conditions are systematically inserted (*IACo*; *Repressor-Sensitizer Test*, *EUCO*).

Psychometric qualities and shortcomings

Instead of giving a detailed consideration we would like to summarize the psychometric qualities and shortcomings of all the tests spoken about in general – partially supported by empirical evaluation evidence. Referring to the standards of psychological test construction there are several important advantages of the test genre in question.

First of all, the criterion of unfakeability is a point. At least such tests appear to be much less fakeable than personality questionnaires. There is a

certain lack of transparency of the measurement principle for the testees (what exactly and how it is being measured are not easily transparent), which consequently means that testees are hardly able to intentionally fake their responses – only faking bad or at least an intentional destructive behavior of a testee, that does not conform with the instruction, is possible, of course. Actually, Baldinger (2006), for instance, proved that *Work Styles* are much less sensitive to the testee's manipulation of the test result at his/her discretion than traditional personality questionnaires.

Given less fakeability, the criterion of fairness becomes more likely to be fulfilled: Faking the test behavior in such a way that the test results serve for the testee's benefit needs a proper (social) intelligence or a proper educational standard; hence, given a too low level of both of them would handicap a testee by applying fakeable instruments.

The criterion of utility is met as a lot of psychological constructs are not at all (ecological) feasible otherwise, e.g. stress resistance. The criterion of reasonableness seems to be realized by tests of this genre rather than by traditional personality questionnaires as the latter impose energetic-motivational demands only due to their banality and hackneyed occurrence, perhaps also simply due to their fakeability – according to experience the tests of the genre of *experiment-based assessment of behavior* enjoy a testee's acceptance because of their face-validity, in particular if the testee is a medium or top manager and the tests are announced as a "computerized single-assessment task". The acceptance by testees has been explicitly established recently even for 7- to 8-year old children as concerns the test-battery *Work Styles* (Nieszner, 2008), which is significant as no personality questionnaire for this age group is available to practitioners.

As all the tests are computer tests and administration, as well as test scoring, is thus computer-assisted, objectivity and economy is given in comparison to other approaches, in particular the assessment centre.

However, there are also several shortcomings, or in other words developmental deficits in their psychometric foundation. This is because the *status quo*

of the Objective Personality-tests and the tests of the genre of *experiment-based assessment of behavior*, respectively, are determined by their original intention; which is fulfilling the practical demands of measuring certain personal aptitudes – and this in a most likely unfakeable manner. That is to say, those aptitudes which when absent regularly give rise to problems in job achievement, however which are not covered properly by traditional personality questionnaires, were selected for measurement. But the genre is hardly determined by any theory of personality. And, finally, it is also not determined by some head start research program in order to evaluate every facet of psychometric quality.

Bearing in mind that this genre and its revival, respectively, have only existed for about 10 years the shortcomings of standardization sample's representativeness, as given for all the tests introduced, are comprehensible – the more, well established approaches, in particular projective techniques, have suffered for almost 90 years from the lack of any standardization at all. However, the criteria of reliability, validity, and calibrating/scoring are more crucial, though again personality questionnaires, for instance, suffer since ever from a low predictive validity as concerns some forthcoming critical behavior (cf. Schmidt & Hunter, 1998). As a matter of fact, most of the tests introduced in this paper lack sufficient validation studies. It is primarily studies, which prove the generalizability of a testee's registered behavior from the specific given task to construct intended other tasks, that are missing.

However, there actually are several predictive validity studies with respect to vocational and educational training benefits. In particular, a combination of *Work Styles* and *ILICA* was intensively investigated with respect to job achievement and the benefit of educational trainings (cf. Frebort, Kubinger & Litzenberger, in prep.): For example, using a linear combination of test-scores (these both tests included) diminished the drop-out rate of the students from a training school for stockmen within the first two terms from 42 to 10 percent in comparison to no test application at all. The same is true with respect to BAcO-D (Ortner & Kubinger, in prep.): For example,

the high-lightening result is a correlation coefficient of .67 between the test score in subtest “Unfavorable working conditions” and the average grade of financial law and taxation students in the following term – as concerns the subtest dealt with here, “Hindrance of scheduled course of action”, the respective coefficient amounts to .45.

Coming closer to measuring construct validity, the comparison between extreme groups, split on the basis of an external construct-close criterion, was used to analyse the *Rigidity Test* (accountants vs. advertising experts) as well as EUcO (market makers vs. military examinees): The estimated relative effect size amounts to 1.36 in the first case and 1.92 in the second case. The discriminant validity of *ILICA* was investigated mainly with respect to item processing (*Number Linking Test*; that is drawing a path as fast as possible from 1 to 90 of all these numbers unsystematically arranged on a sheet) and reasoning (*Calculating in Symbols*): Both correlation coefficients do not differ significantly from null.

Some of the test conceptualizations only purport content validity, for example *Luchins-based Rigidity Test*. This might initially provoke a certain amount of scepticism as testees may potentially realize that there are alternative, simpler solution strategies, but nevertheless apply the previously learned strategies for whatever reasons. Luckily, a study where the testees had to process the test by thinking out loud proved that they indeed almost always do not discover the simpler strategy (Krapf, 2003).

Concerning reliability and precision of measurement, respectively, an internal-consistency method such as split-half reliability often does not work because scoring refers to the ratio of the frequencies of several kinds of actions and reactions under different conditions, the latter not being able to be separated into certain parts (bear in mind that ratios are always critical if the denominator tends to zero). However, Kubinger and Draxler (2006) have already suggested what kinds of modifications of scoring rules would have to be applied in order for such tests to be calibrated according to item response theory (IRT). That is, if tests would be re-conceptualized so that a calibration of their items for instance along the Rasch

model would be possible, then the calculation of reliability would be feasible.

And re-test reliability is not always meaningful because of an “aha-experience” after working through the test. As concerns *ILICA*, in particular, very low coefficients have been established unless the test scores were pooled into very crude categories (cf. Herle, 1999): testees who are not distracted easily during the first test run are in no case revealed as being easily distracted during the second run (four weeks later), and vice-versa with only 10%. However, recently Maierhofer (2008) disclosed that – using BAcO-D, the subtests “Comparison of surfaces” from *Work Styles* in a new version and the *Repressor-Sensitizer Test* – the highest coefficient of re-test reliability (6 to 8 days between test and re-test) resulted for BAcO-D’s subtest “Hindrance of scheduled course of action” (.872), though generally questionnaires test scores prove to have coefficients which significantly exceed .50.

But to summarize, the limitations in the current research findings as concerns reliability and validity must be acknowledged for the while.

Practical value

Perhaps the most valuable empirical result supporting the use of the genre of *experiment-based assessment of behavior* is the following. A confirmatory factor analysis was carried out in order to test which of the Big Five-factors are possibly covered by the test-batteries *Work Styles* and BAcO (Kubinger & Litzenberger, 2003). This study discovered that the (Big Five) traits model did not prove to fit the data, but by modeling so-called method factors instead of trait factors did indeed prove to fit: That is, questionnaires as the one factor level and Objective Personality-tests as the other factor level were hypothesized to explain all the variance in the testees’ observed test scores. This empirical result consequently highlights that personality assessment should additionally include the use of the genre of *experiment-based assessment of behavior* because otherwise relevant differences between testees might not be disclosed.

So far the application of such tests in practice seems justified, yet. Some applications have even

been communicated to the public in scientific papers:

- The test-batteries *Work Styles* and BAcO-D as well as *ILICA* are used for the selection of the heads of jailhouses (Kubinger, Holocher-Ertl & Frebort, 2006; Holocher-Ertl, Kubinger & Frebort, 2006).
- For the selection of candidates for justice-ships *Work-Styles* and *ILICA* are used (Frebort, Kubinger & Holocher-Ertl, 2006).
- The test-battery *Work Styles* is being field-tested in the selection of army officers (Melter, 2006).
- The selection of applicants for social workers’ college is based on BAcO-D and a modification of LEWITE (Litzenberger & Haiden, 2006).
- Applicants to colleges of higher education for marketing and sales are tested with BAcO-D (Ortner & Janous, 2006).
- The subtest “Coding symbols” from *Work Styles*, LAMBDA and again a modification of LEWITE are being used to counsel student applicants on the basis of a non-binding self-assessment for a qualifying examination (Frebort & Kubinger, 2008).

Discussion

We are confident that the genre of *experiment-based assessment of behavior* will become more and more useful and will contribute, in particular, to an improved professional personal selection practice. This is due to the fact that this technique essentially prevents faking as testees generally do not have to assess themselves; and also tests of this genre measure the personal style of a person, namely how he/she behaves in typical situations in quite typical ways, rather than the actual personality traits. But, of course, we are aware that the approach still needs to be elaborated to a certain degree in order to be promoted to a universal standard instrument. Many shortcomings have been mentioned so far. Some of them may be eliminated very easily in the future; others may prove to be quite a challenge. The fact is that a single and very small research group such as ours cannot do the job all on its own. Hence, this paper may encourage other research groups to contribute empirical evidence of this genre’s quality. Above all, the question of generalizability should be investigated thoroughly. By the way, analyses are also needed pertain-

ing to the dependencies of subsequent actions and reactions during specific test procedures, for instance in *ILICA*; it has not yet been systematically tested whether certain “negative” or sub-optimal decisions at the very beginning can be compensated in the resulting test-score by optimal decisions later on.

As concerns the Chinese context, there is, of course, deliberately research necessary whether the genre of *experiment-based assessment of behavior* entails any cultural difference in test-scores and, all above, in reasonableness.

References

- Baldinger, D. (2006). Der Einfluss sozial erwünschten Verhaltens auf das Ergebnis Objektiver Persönlichkeitstests [*The effect of socially desirable behavior on the test scores of objective personality questionnaires*]. Unpubl. master thesis, university of Vienna.
- Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.
- Bruner, J. S. & Postman, L. (1947). Emotional selectivity in perception and reaction. *Journal of Personality*, 16, 69–77.
- Cattell, R. B. & Warburton, F. W. (1967). *Objective personality and motivation tests*. Urbana, IL.: University of Illinois Press.
- Chen, Z. G. (1983). Item Analysis of Eysenck Personality Questionnaire Tested in Beijing-Districts (in Chinese). *Acta Psychologica Sinica*, 2, 211–218.
- Chen, C. S., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students (in Chinese). *Psychological Science*, 6, 170–175.
- Cheung, F. M. (2004). Use of Western- and indigenously-developed personality tests in Asia. *Applied Psychology: An International Review*, 53, 173–191.
- Cheung, F. M. (2009). The cultural perspective of personality assessment. In J. N. Butcher (Ed.). *Oxford Handbook of Personality Assessment* (pp. 44–56). New York: Oxford University Press.
- Cheung, F. M., Cheung, S. F., Wada, S., & Zhang, J. X. (2003). Indigenous measures of personality assessment in Asian countries: A review. *Psychological Assessment*, 15, 280–289.
- Cheung, F. M., Fan, W. Q., Cheung, S. F., & Leung, K. (2008). Standardization of the Cross-cultural [Chinese] Personality Assessment Inventory for adolescents in Hong Kong: A combined emic-etic approach to personality assessment (in Chinese). *Acta Psychologica Sinica* (心理学报), 40, 839–852.
- Cronbach, L. J. (1970). *Essentials of psychological testing*. New York: Harper & Row.
- Dai, X. Y. & Wu, Y. Q. (2005). A Study on NEO-PI-R Used in 16–20 Years Old People (in Chinese). *Chinese Journal of Clinical Psychology*, 13, 14–18.
- Dichert, S., Ones, D. S., Viswesvaran, C. & Deller, J. (2006). Response distortion in personality measurement: born to deceive, yet capable of providing valid self assessments? *Psychology Science*, 48, 209–225.
- Frebort, M. & Kubinger, K. D. (2008). *Qualitätsansprüche an ein Self-Assessment zur Studienwahlberatung: Der Wiener Ansatz* [Quality claims on self-assessments for student advisory services]. In H. Schuler & B. Hell (eds.), *Studierendenauswahl und Studienentscheidung* [Selection of student applicants and students’ decision for courses] (pp. 95–101). Göttingen: Hogrefe.
- Frebort, M., Kubinger, K. D. & Holocher-Ertl, S. (2006). Der diagnostische Prozess zur Auswahl von Richteramtsanwärtern gemäß DIN 33430 [The diagnostic process for the selection of judge aspirants according to the German industry norm 33430]. In K. Westhoff (ed.), *Nutzen der DIN 33430* [The use of the German industry norm 33430] (pp. 53–65). Lengerich: Pabst.
- Frebort, M., Kubinger, K. D. & Litzenberger, M. (in prep.). Iterative optimizing test-scores’ discriminant function for presonal selection – an empirical example.
- Hamid, P. N., Lai, J. C. L., & Cheng, S. T. (2001). Response bias and public and private self-consciousness in Chinese. *Social Behavior and Personality*, 29, 733–742.
- Herle, M. (1999). Zur Reliabilität von psychologischen Testergebnissen in Simulationsverfahren am Beispiel von *ILICA* und *MAILBOX’90* [On the reliability of computer simulations’ test-scores exemplified by *ILICA* and *MAILBOX’90*]. Unpubl. master thesis, university of Vienna.
- Hofmann, K. & Kubinger, K. D. (2001). Herkömmliche Persönlichkeitsfragebogen und Objektive Persönlichkeitstests im „Wettstreit“ um Unverfälschbarkeit [Personality questionnaires and objective personality tests in contest: More or less fakeable?]. *Report Psychologie*, 26, 298–304.
- Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3. Aufl.). Odessa: Psychological Assessment Resources.
- Holocher-Ertl, S., Kubinger, K. D. & Frebort, M. (2006). Objektive Persönlichkeitstests in der Personalauswahl: Justizanstalt-sleiter [Objective personality tests for employee selection: heads of jailhouses]. In T.M. Ortner, R. Proyer & K.D. Kubinger (eds.), *Theorie und Praxis Objektiver Persönlichkeitstests* [Theory and practice of objective personality tests] (pp. 222–233). Bern: Huber.
- Jäger, R. S. (2003). Arbeitsprobe [Work demonstrations]. In K.D. Kubinger & R.S. Jäger (eds.), *Schlüsselbegriffe der Psychologischen Diagnostik* [Keynotes to Psychological Assessment] (pp. 42–43). Weinheim: PVU.
- Krapf, B. (2003). *Validierungsstudie des Rigiditätstest nach Siebenhandl et al* [A validation study on the Rigidity Test from Siebenhandl et al.]. Unpubl. master thesis, university of Vienna.
- Kubinger, K. D. (1997). Zur Renaissance der objektiven Persönlichkeitstests sensu R.B. Cattell [Renaissance of Objective Personality tests sensu R.B. Cattell]. In H. Mandl (Ed.), *Bericht über den 40. Kongreß der Deutschen Gesellschaft für Psychologie in München 1996* [Proceedings of the 40th congress of the German society of psychology in Munich 1996] (p. 755–761). Göttingen: Hogrefe.
- Kubinger, K. D. (2002). On faking personality inventories. *Psychologische Beiträge* [later: *Psychology Science*], 44, 10–16.
- Kubinger, K. D. (2006). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens* [Psychological Assessment – Theory and Application]. Göttingen: Hogrefe.
- Kubinger, K. D. & Draxler, C. (2006). Testtheoretische Probleme

- bei Objektiven Persönlichkeitstests – Ideenskizze für die Nutzung der Item-Response-Theorie [Psychometric problems of objective personality tests – towards the use of item response theory]. In T.M. Ortner, R. Proyer & K.D. Kubinger (eds.). *Theorie und Praxis Objektiver Persönlichkeitstests* [Theory and practice of objective personality tests] (pp. 88–101). Bern: Huber.
- Kubinger, K. D. & Ebenhöf, J. (1996). *Arbeitshaltungen - Kurze Testbatterie: Anspruchsniveau, Frustrationstoleranz, Leistungsmotivation, Impulsivität/Reflexivität* [Work Styles – A short test-battery: Level of aspiration, achievement motivation, frustration tolerance, impulsiveness/reflexiveness] Test: Software and Manual. Frankfurt: Swets.
- Kubinger, K. D. & Ebenhöf, J. (2002). *Arbeitshaltungen - Kurze Testbatterie: Anspruchsniveau, Frustrationstoleranz, Leistungsmotivation, Impulsivität/Reflexivität* [Work Style – A short test-battery: Level of aspiration, achievement motivation, frustration tolerance, impulsiveness/reflexiveness] – Version 26.00. Test: Software and Manual (writer of this manual: T. Karner & M. Sommer). Mödling: Wiener Testsystem/ Schuhfried.
- Kubinger, K. D., Holocher-Ertl, S. & Frebort, M. (2006). Der diagnostische Prozess zur Auswahl von Justizanstaltsleitern gemäß DIN 33430 [The diagnostic process at the selection of heads of jailhouses according to the German industry norm 33430]. In K. Westhoff (ed.), *Nutzen der DIN 33430* [The use of the German industry norm 33430] (pp. 66–75). Lengerich: Pabst.
- Kubinger, K. D. & Litzenger, M. (2006). On the shortcomings of a mono-methodical approach to personality assessment. *Individual Differences Research*, 4, 139–158.
- Litzenger, M. & Haiden, D. (2006). Evaluation einer sequentiellen Auswahlstrategie mittels Objektiver Persönlichkeitstests und Assessment Center [Evaluation of a sequential selection strategy using objective personality tests and an assessment center]. In T.M. Ortner, R. Proyer & K.D. Kubinger (eds.). *Theorie und Praxis Objektiver Persönlichkeitstests* [Theory and practice of objective personality tests] (pp. 234–243). Bern: Huber.
- Luchins, A. S. & Luchins, E. H. (1959). *Rigidity of Behavior - A Variational Approach to the Effect of Einstellung*. Eugene (OR): University of Oregon Books.
- Maierhofer, P. (2008). Zur Optimierung psychologisch-diagnostischer Strategien in der Begutachtung von Sexualstraftätern [On optimizing diagnostic strategies for the assessment of sexual offenders]. Unpubl. Master thesis, University of Vienna.
- Melter, A. (2006). Objektive Persönlichkeitstests in der Militärpsychologie [Objective personality tests in military psychology]. In T.M. Ortner, R. Proyer & K.D. Kubinger (eds.). *Theorie und Praxis Objektiver Persönlichkeitstests* [Theory and practice of objective personality tests] (pp. 254–263). Bern: Huber.
- Möseneder, D. & Ebenhöf, J. (1996). *ILICA - Ein Simulationstest zur Erfassung des Entscheidungs-verhaltens* [A computer simulation for decision-making assessment]. Test: Software and Manual. Frankfurt/M.: Swets Test Services.
- Niesner, G. (2008). Eine Machbarkeitsstudie: Ein Objektiver Persönlichkeitstest für Kinder [A study of feasibility: an objective personality test for children]. Unpubl. Master thesis, university of Vienna.
- Ortner, T. M. & Janous, G. (2006). Zur Nützlichkeit des Einsatzes eines Objektiven Persönlichkeitstests zur Auswahl von Fachhochschulstudierenden [The use of objective personality tests for the selection of applicants to colleges of higher education]. In T.M. Ortner, R. Proyer & K.D. Kubinger (eds.). *Theorie und Praxis Objektiver Persönlichkeitstests* [Theory and practice of objective personality tests] (pp. 244–253). Bern: Huber.
- Ortner, T. M. & Kubinger, K. D. (in prep.). First psychometric analyses on objective performance tests assessing a person's stress resistance.
- Ortner, T. M., Kubinger, K. D., Schrott, A., Radinger, R. & Litzenger, M. (2006). *Belastbarkeits-Assessment: computerisierte Objektive Persönlichkeits-Testbatterie – Deutsch (BAO-D)* [Stress resistance assessment: computerized objective test-battery – German version]. Test: Software and Manual. Frankfurt/M.: Harcourt Test Services.
- Ortner, T. M., Proyer, R. T. & Kubinger, K. D. (eds.) (2006). *Theorie und Praxis Objektiver Persönlichkeitstests* [Theory and practice of objective personality tests]. Bern: Huber.
- Proyer, R. & Häusler, J. (2008). *Multimethodische Objektive Interessenbatterie (MOI)* [Multi-methodical Objective Test-Battery of Interests]. Test: Software and Manual. Mödling: Viennese Testsystem/ Schuhfried.
- Seiwald, B. B. (2002). Replicability and generalizability of Kubinger's results: Some more studies on faking personality inventories. *Psychologische Beiträge* [later: *Psychology Science*], 44, 17–23.
- Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.
- Song, W. Z. (1985). Analysis of Results of MMPI of Normal Chinese Subjects (in Chinese). *Acta Psychologica Sinica* (心理学报), 4, 346–355.
- Viswesvaran, C. & Ones, D. S. (1999). Meta-analyses of fability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59, 197–210.
- Wagner-Menghin, M. M. (2004). *Lexikon-Wissen-Test (LEWITE)* [Lexicon Knowledge Test]. Software und Manual. Mödling: Dr.G.Schuhfried GmbH.
- Wang, D. F. & Cui, H. (2003). (in Chinese:) Processes and Preliminary Results in the Construction of the Chinese Personality Scales (QZPS). *Acta Psychologica Sinica* (心理学报), 35, 127–136.
- Yang, J., Bagby, R. M., & Ryder, A. G. (2000). Response style and the revised NEO Personality Inventory: Validity scales and spousal ratings in a Chinese psychiatric sample. *Assessment*, 7, 389–402.
- Zhang, H. C. (1988). Psychological measurement in China. *International Journal of Psychology*, 28, 75–95.
- Zhang, J. X., Song, W. Z., & Cheung, F. M. (1999). The MMPI-2 and its standardization in China and Hong Kong. *Chinese Mental Health Journal*, 13, 20–23.
- Zhang, J. X., & Zhou, M. J. (2006). Searching for a personality structure of Chinese: A theoretical hypothesis of a six factor model of personality traits (in Chinese). *Advances in Psychological Science* (心理科学进展), 14, 574–585.
- Zhou, M. J. & Zhang, J. X. (2008). The Use and Development of Personality Testing in China. *Testing International*, 19, 5–7.

当代客观化人格测验的技术——基于实验的行为评估： 维也纳研究小组开发的多种计算机化测验介绍

Klaus D. Kubinger

(Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Austria)

摘 要 目前多数人格测验(特别是在中国使用的人格测验)基本上都是人格问卷,基于实验的行为评估类客观化人格测验应用很少;而后者近来在德语圈国家中则有复苏的迹象。因此,本文综述了此类客观测验相对于人格问卷来说所具有的特点和优势,如,被试很难在这类客观化人格测验中作伪。本文介绍了维也纳研究小组所做的几个测验,并讨论了这些测验的心理测量学性质和缺点。最后,还列举了这些测验的实际应用。

关键词 客观化人格测验;心理测验;人格问卷;防伪性;计算机化测验

分类号 B841.7

沉痛悼念葛小佳先生

国际著名心理学家和行为遗传学家、美国明尼苏达大学资深教授、中国科学院心理研究所客座研究员、《心理学报》编委葛小佳先生,因病医治无效,于2009年当地时间8月26日下午4时40分,在美国明尼苏达辞世,享年55岁。

葛小佳先生1954年10月24日生于北京。1977年考入四川大学历史系。1981年就读厦门大学台湾研究院,获硕士学位,毕业留校任教。1985年赴美国求学于爱荷华州立大学,1990年获社会学博士学位,就职于该校农村家庭精神健康中心。1995至2006年执教加利福尼亚大学戴维斯分校,先后任人类与社区发展系副教授和教授。2006年受聘明尼苏达大学儿童发展研究中心资深教授。

葛小佳先生致力于研究儿童和青少年如何以及为什么出现行为和情绪问题,这些问题会给他们的生活带来何种不良后果。研究工作强调时间、生理、环境和遗传-环境相互作用在青少年心理发展过程中的重要性。在国际发展心理学顶级刊物上发表很多重要论文,并因在这一领域的开创性工作而获得广泛的学术声誉。

葛小佳先生曾任《发展心理学》(Developmental Psychology)、《国际行为发展》(International Journal of Behavioral Development)和《家庭心理学》(Journal of Family Psychology)等多家有影响力的学术期刊编委,美国青少年研究协会项目评审专家及全国双年会联合主席,儿童发展研究协会评论小组主席,美国精神卫生研究所(NIMH)资助的家庭联合研究会III和IV项目(Family Research Consortium III and IV funded by NIMH)的领导者,美国国立卫生研究院(NIH)科学评审中心研究部门的成员,以及一些私立基金会(如McArthur基金会和William T. Grant基金会)的特邀评论者。

葛小佳先生始终关心中国心理学的发展,将国外先进的理念和方法介绍到中国。2005年应邀任中国科学院心理研究所客座研究员,建立青少年情绪与行为问题研究中心,建设中国青少年双生子库,开展行为遗传学研究,培养青年学者及研究生。2006年任《心理学报》海外编委。2008年作为特邀编辑,约请一批国际知名学者为《心理学报》编纂“行为遗传学专辑”,为中国行为遗传学的发展作出了重要贡献。

葛小佳先生千古!

《心理学报》编委会