HUMAN PSYCHOPHARMACOLOGY, VOL. 6, 91–98 (1991)

REVIEW

Performance Tests in Human Psychopharmacology (2): Content Validity, Criterion Validity, and Face Validity

A. C. PARROTT
Department of Psychology, Polytechnic of East London, London, E15 4LZ, UK*

The four traditional aspects of validity comprise: criterion, content, face and construct. Criterion validity is assessed by calculating the degree of correlation between test score and external criterion (e.g. driving ability of flying skill). While this form of evidence is rare, it has been demonstrated in a few psychopharmacology studies, especially with psychomotor tasks (e.g. target tracking). Content validity is an index of the appropriateness of test selection. Many test batteries cover a wide range of psychological functions, and thus display a degree of content validity. However, the functions being assessed often differ, reflecting the absence of agreement over what 'human performance' should comprise. Most test batteries also contain a preponderance of simple tasks, and content validity could be improved through the inclusion of 'higher cognitive' tasks. The difficulties of assessing performance using a single test are also critically examined. Face validity is not a true index of validity, but a reflection of whether the test looks right. It can often be misleading, although it does have positive functions (e.g. for subject motivation). Sensitivity to psychoactive drug effect is also a prerequisite for tests in this area, although it is not in itself an index of validity. The construct validity of human performance tests, together with an overview of test meaning and interpretation, are examined in the final paper of this series.

KEY WORDS—Psychometrics, validity, performance, test, assessment, psychopharmacology, psychoactive drug.

BASIC PSYCHOMETRIC PRINCIPLES

Psychometric principles have been incorporated into many forms of psychological assessment, but in human psychopharmacology performance research they have often been ignored. Cull and Trimble (1987, p. 141) stated that the automated performance tests used in human psychopharmacology 'Have not been well evaluated as regards their psychometric properties, particularly with respect to reliability and validity.' Hindmarch (1989) stated that human performance tests had often not been adequately validated. Wittenborn (1987, p. 74) concluded that the behavioural concepts behind them were often 'obscure', 'incompletely defined', or 'poorly articulated'. Further details from these reviews were presented earlier (Parrott, 1990). Also reported was a brief analysis of the frequency with which psychometric evidence was quoted in human psychopharmacology research. Neither task reliability nor validity were documented in any paper within the review sample. Furthermore, those few papers that did mention validity, generally only stated that it had not been investigated; none discussed how it might be established. Overall, there has been little constructive debate, within human psychopharmacology, on the problems of establishing the validity of performance tests. General principles of psychometrics are described in: Anastasi (1982), Cronbach (1984), or Kline (1986). Recent papers on validity include: Anastasi (1986), Angoff (1987), and Cronbach (1987).

VALIDITY: INTRODUCTION

> A test is valid if it measures what it claims to measure (Kline, 1986, p. 3).

Validity is not a static property, intrinsic to a test, but a reflection of how it is being used. A test may be valid for one purpose, but invalid for another. Validity is essentially a psychological judgement based upon 'the relationship between performance on the test and other independent observable facts about the behaviour' (Anastasi, 1982, p. 131). There are numerous ways in which this might be investigated. In 1954 the American Psychological

* Formerly North East London Polytechnic.

0885-6222/91/020091-08$05.00

Accepted 19 September 1990

Association discussed the numerous types of validity evidence then being used, and published its influential classification into three main types: content, criterion (concurrent and predictive) and construct. This classification system has been widely used ever since, while 'face validity' is generally also assessed as a fourth (non-technical) aspect (Anastasi, 1982). Several authorities have recently suggested a pre-eminence for construct validity (Anastasi, 1986; Angoff, 1987; Cronbach, 1987); this is discussed later.

FACE VALIDITY

> Face validity is not a validity in the technical sense; if refers not to what the test actually measures, but to what it appears superficially to measure (Anastasi, 1982, p. 137).

A test has face validity when it looks appropriate. An example of a performance test with face validity is the Mackworth clock test. It was designed to be a laboratory analogue for the military task of radar scanning (Mackworth, 1957). The clock face has visual similarity with the face of the radar screen, while the sweep of the dial around the clock resembles the sweep of the radar beam. Other aspects of the task are less veridical: the speed of the sweep of the arm, the nature of the stimulus target, the nature of noise events and the frequency of targets. Another test with face validity is the car-racing game (Kennedy *et al.*, 1982). Here, steering wheel control, the use of accelerator/brake, and visual input, all loosely resemble aspects of high-speed car-racing. However, the resemblance is only superficial, with the real-world task of high-speed racing differing in many fundamental ways from the video game. It should, however, be noted that well-designed driving simulators exist; it is just that arcade games are generally based on superficial resemblances. Face validity is therefore not an index of true validity. Conversely, a task without face validity may be valid. For instance, auditory vigilance may be a valid index of driving ability, since vigilance is an important factor in everyday driving (Herbert and Jaynes, 1964), yet auditory vigilance has no face validity as a measure of car driving.

Face validity can be misleading. Researchers may believe their tasks to be valid because they feel and look good. But this can engender unjustified feelings of satisfaction, and a failure to attempt to establish true validity. Many non-experts (possibly some experts?), fail to realize that face validity is not an index of true validity. Kline has called face validity 'A trivial aspect of the test' (1986, p. 152). Yet he also acknowledges that it can be useful. Subjects should feel content doing the task, otherwise they may abreact against the testing situation. For instance, military subjects performing the Mackworth clock test should perceive its likeness to radar scanning. Face validity may also be important to generate appropriate motivation (rather than high motivation). Thus the Mackworth subjects might be expected to perform the clock test in a similar way to the radar task. Task duration is also relevant here. When the laboratory task is given for a period similar to the real-world task, face validity will be present, but again not true validity. Face validity is also important for the client paying for the study. If findings emerge from tests which look appropriate, then results will appear more meaningful, and recommendations will have a higher likelihood of being accepted. The converse of the above is that, when using tasks with *low* face validity, it may be necessary to explain to subjects/clients that, contrary to appearances, the tasks are meaningful and predictive.

Overall, face validity should be built into each task at the design stage. It will aid subjects to be appropriately motivated when undergoing testing. However, face validity is not proper validity in *any* sense. Some good-looking tests are insensitive and lack validity.

CONTENT VALIDITY

> Content validation determines whether the test covers a representative sample of the behaviour domain (Anastasi, 1982, p. 131).

Content validity reflects the degree to which a particular area of interest is being adequately covered. It is comparatively straightforward to assess in narrow, well-designed areas (e.g. typing skills), but is more difficult to assess in 'broad' topics (e.g. human performance in general). In order to investigate the content validity of tests in this area, current understandings on 'human performance' need to be considered. This is covered later, in the section on construct validity, but several aspects of content validity are examined here: the content validity of test batteries; factor analysis as an aid to task classification; the overrepresentation of simple types of task in batteries; the limitations of using single

tasks; and lastly the content validity of task simulation and real-world performance assessment.

Task batteries are widely used in human psychopharmacology. Indeed, this area has been at the forefront of modern test battery development. The main problem in designing the battery is deciding which psychological functions to assess. Wesnes *et al.* (1987). p. 80) suggested four: attention (the selection of information from the environment), cognition (the processing of this information), memory (the entry of this information into storage) and behavioural response (the skilled physical coordination in responding to information). Hockey and Hamilton (1983, pp. 348–349) suggested five areas: alertness, selectivity, speed, accuracy, and short-term memory. Cull and Trimble (1985) also proposed five: attention and sensory processing, mental speed, central cognitive processing ability, memory and perceptuo-motor performance. Parrott (1986) used six: sensory reception and attention, arousal and alertness, simple information processing, complex information processing and cognition, memory storage, and simple psychomotor. Holding's (1989) analysis of human skills led to four perceptual-motor groupings: simple/perceptual (e.g. vigilance), simple/motor (e.g. tapping), complex/perceptual (e.g. air traffic control) and complex/motor (e.g. aircraft piloting), while verbal–intellectual skills required further tests (Holding, 1989). Other taxonomies have been suggested (see the tasks listed within Fleishman, 1975; Hindmarch, 1980; Cull and Trimble, 1987). While these taxonomies display broad general agreement, they still retain much variation. Furthermore, when the position of an individual test is investigated there is often disagreement on the psychological function(s) being assessed. For instance with digit symbol substitution, Hindmarch (1980) lists it as a sensory processing task, Parrott (1986) used it as a simple information-processing measure, while Cull and Trimble (1987, p. 152) categorized it as a psychomotor task. Although these authors would probably agree on the nature of symbol coding as a task, their variation in task labels does illustrate the difficulty of assigning any one task to a simple category.

Fleishman (1975) attempted to classify performance tests through factor analysis, with the belief that test selection should be 'an empirical rather than armchair question'. Numerous factors have been generated, but these factor patterns vary with the particular test battery. He has therefore concluded: 'The search for a single general taxonomy of human performance is not likely to be successful ... some invention is necessary' (Fleishman, 1975, p. 1147). Bittner *et al.* (1986) developed their test battery on psychometric principles. Forty-five tests, selected from a literature survey of 145 tasks, passed initial reliability and stability-of-variance criteria (see Table 1 in Parrott, 1990). Factor-analytic and domain information was used to identify similar test subsets. Then 'reliability–efficiency data' were employed to select the most sensitive test versions (Bittner *et al.*, 1986; p. 699). The resulting battery comprised five tests: logical reasoning (left hemisphere cognitive), pattern comparison (right hemisphere cognitive), code substitution (memory/perceptual), aiming (fine sensory-motor control) and spoke control (gross psychomotor). Despite the psychometric care taken in its design, the battery has numerous omissions: attention and vigilance, memory, arousal, simple psychomotor speed, complex psychomotor skill and several others. Also, test sensitivity to the effects of stressors (e.g. noise, fatigue, psychoactive drug), was not a criterion in the selection procedure. Bittner's battery is therefore neither stronger nor conceptually more sound than those where choice of test was based on theoretical and practical considerations.

Simple information processing and psychomotor tasks tend to be overrepresented in most test batteries. They are easy to learn and perform, and generate a volume of data in a short time period. In contrast, complex tasks, particularly those involving higher cognitive processing, are generally underrepresented. Logical reasoning is probably the most complex task commonly found in batteries (Baddeley, 1968; Bittner *et al.*, 1986), yet many other complex tasks have been successfully used in psychopharmacology studies: modified video games for air combat and slalom driving (Kennedy *et al.* 1982; Bittner *et al.*, 1986); navigational plotting (Wiker *et al.*, 1983); mental rotation (Tomlinson *et al.*, 1982); concept identification (Parrott *et al.*, 1982; Thompson and Trimble, 1982); Tower of Hanoi problem-solving (Morris *et al.*, 1987); and creativity (Warburton, 1987). The inclusion of complex tasks such as these would improve the content validity of most test batteries.

Many studies have assessed performance using a single task. This is valid as long as conclusions are offered only for that measure, but there is sometimes a tendency to generalize to performance in general. For instance, Wood *et al.* (1985) undertook a comparative study of motion-sickness prophyl-

axes for astronauts. They reported unimpaired target tracking, the sole performance task used, under scopolamine, leading to the conclusion that it did not affect performance: 'These drugs should produce no significant performance decrement in an operational situation' (Wood *et al.*, 1985, p. 315). While medium dose levels of scopolamine generally *do* leave simple psychomotor skills unimpaired, memory and sustained attention are generally impaired at these dose levels (Parrott, 1986, 1989). Thus the authors would have been correct if they had concluded that psychomotor control was not impaired, but were probably incorrect to attempt a broader generalization.

Several groups have however recommended using a single task as an index of overall performance. Hockey and Hamilton (1983) discussed a series of studies where different stressors and drugs had been assessed using the Wilkinson continuous choice reaction time task. They suggested that each stressor/drug tended to produce a different performance pattern, which could then be subjected to an individual interpretation (Hockey and Hamilton, 1983, pp. 340–342). Hindmarch (1980) similarly proposed that discrete choice reaction time provided an index of overall sensory-motor integrity. Stimulus recognition time reflected the efficiency of input processing, while motor response time reflected the adequacy of psychomotor output systems. Sternberg (1969) developed the choice reaction time paradigm into a formal model of information-processing stages (see Construct validity section). Furthermore, these different information processing stages have been shown to be sensitive to particular drug effects (Frowein, 1981). While confirmatory studies are needed for these specific stage/drug effects, the findings suggest that a single task can be used to assess different aspects of information processing, but only when the paradigm has been developed on sound theoretical principles (Sternberg, 1969; Frowein, 1981; Sanders, 1983).

Task simulation and real-world task assessment are two approaches with the potential for high content validity. Both require that performance area of interest be defined beforehand. Then either a simulator is constructed, or a procedure for quantifying performance on the real task devised. For instance, Billings demonstrated significant impairments in flying light aircraft following both alcohol (Billings *et al.*, 1973), and secobarbital (Billings *et al.*, 1975). Seashore and Ivy (1953) investigated the effects of amphetamine upon army personnel undergoing a range of military tasks. Porges *et al.* (1981) tested 'hyperactive' young children under methylphenidate, while playing in their school classroom. The children's performance at identifying brightly coloured aliens as 'targets', on the VDU of a spacecraft toy, was recorded as an index of concentration. Comitas (1976) compared sugarcane cutting yields over the harvest season, in farm workers who used or did not use cannabis. Van Lunteren and Stanssen (1967; cited in Michon, 1973, p. 164) developed a model for control behaviour during cycling, which they then assessed while under alcohol. They concluded: 'The increasingly erratic path of the intoxicated driver is due to increasing swaying of the trunk rather than of the whole vehicle–driver system'. These sorts of study show high content validity when the overall problem is defined in narrow terms (e.g. how does alcohol affect cycling?), but show reduced generalizability for wider problems (e.g. how does alcohol affect performance in general?). Some of the above studies are described more fully in the next section, since real-world performance assessments can be used as external criteria in the investigation of criterion validity.

CRITERION VALIDITY

> With criterion validity, performance on the test is checked against a criterion, a direct and independent measure of what that test is designed to predict (Anastasi, 1982, p. 137).

When criterion and test are measured at the same time, concurrent criterion validity is being assessed, whereas if the criterion data are collected later, then predictive criterion validity is being assessed (Cronbach, 1984). This is the only type of validity with a clear statistical basis, namely the correlation between test score and criterion measure. There are, however, several problems in its calculation: the multitude of real-world types of performance; the difficulty of deciding upon a good external criterion measure; the problems of measuring performance in the real world; contamination of the criterion by knowledge of the test score; and the large sample sizes required in order to minimize random error.

The external criterion needs to be a measure of real-world performance. However few real-life events can be measured easily. Sanders (1987, p. 116) has stated: 'A major and well recognised problem is that it is usually impossible or impractical to assess drug effects on performance in everyday

life.' Despite these undoubted problems, several areas of real-life performance have been subjected to controlled measurement. They include: aircraft flying, car driving, military activities, industrial tasks and several other real-world scenarios (Parrott, 1987). These studies have provided some impressive data on task validity, and a selection of them are presented below. Some of the difficulties inherent in establishing criterion validity are also covered.

Billings *et al.* (1975), investigated pilots flying a standard air-course in a light aircraft, and undertaking flight simulator tasks on the ground. Data on tracking accuracy and airspeed control were automatically recorded from both situations. Secobarbital impaired both real flying and flying-tracking on the simulator, in a dose-related manner. Thus the flying simulator was validated against the real-world flying. They also reported that the laboratory task displayed a greater proportion of explained-to-unexplained variance, indicating that the simulator was comparatively more sensitive. Several differences between the two tasks were, however, evident, including different learning functions, which led the authors to conclude: 'Extrapolation of simulator data to the flight environment must be approached with considerable caution' (Billings *et al.*, 1975, p. 304). Billings *et al.* (1973) had earlier demonstrated the sensitivity of the same flying task to the effects of alcohol. Experienced and novice pilots were required to fly the figure-of-eight course in the light aircraft, with a co-pilot to take over in case of emergency! Performance was significantly impaired at the lowest dose level (0·04 per cent blood alcohol concentration), while flying errors increased linearly in relation to alcohol level. At the highest dose (0·12 per cent BAC) the safety pilot had to take over in 53 per cent of the flights. In a related type of study Henry *et al.* (1974) validated two laboratory tracking tasks (multidimensional pursuit, and complex coordination), against performance on a modern Link flying simulator. The three tasks equated well on overall sensitivity. One limitation common to all the above studies is that none reported the correlation between performance levels from the different tasks. Overall, however, these tests of aircraft flying and flight simulation were very sensitive to psychoactive drugs (Parrott, 1987, p. 100). They therefore have great potential for the validation of laboratory performance tasks.

Seashore and Ivy (1953) investigated the effectiveness of CNS stimulants for combating sleepiness in UK Army personnel. In a series of well-designed placebo-controlled trials, portable versions of standard laboratory tasks: choice reaction time, critical flicker fusion and target tracking were 'inserted into the work cycle of the men'. Typical military tasks were assessed at the same time: night-time truck driving in blackout conditions, desert tank driving, prolonged guard duty and others. Several significant improvements in performance were noted under amphetamine, both on the laboratory tasks and the military assessments. Thus a broad tendency towards criterion validation was present. However, variation was also apparent, with different patterns of performance change across the parallel studies. Significant laboratory task improvements sometimes occurred, but sometimes did not; while variation was equally apparent with the criterion measures. The question remains: which tests were validated, and against which external criteria? This problem—of inconsistent statistical relationships—will invariably occur in studies which attempt to establish criterion validity across several tasks.

Several research groups have investigated drug effects upon car driving. Hansteen *et al.* (1976) investigated driving skills on a 1·1-mile closed course. In-car experts rated different aspects of driving, while external marshals counted the number of bollards hit. The laboratory task was target tracking. With both the real-world driving accuracy and target tracking accuracy, performance was best with placebo, impaired marginally by low-dose cannabis, impaired more by high-dose cannabis and impaired to the greatest extent by alcohol. Only with car driving speed was this pattern not found, since alcohol led to marginally faster course completion times than placebo. This laboratory tracking task was therefore validated against accuracy on the driving course. As with the aircraft flying studies described earlier, the correlation between laboratory and real-world performance was unfortunately not reported. In a similar type of study, Klonoff (1974) investigated dual-control car driving, both on a private closed course (slalom, reversing, emergency breaking, risk-judgements, etc), and on the streets of Vancouver (expert ratings for starting, stopping, lane changing, careless driving, etc.). The proportion of drivers significantly impaired was similar across the two driving scenarios: low-dose cannabis (33 per cent closed course, 42 per cent open street), high-dose cannabis (55 per cent closed course, 63 per cent open street). This study therefore validated

the closed course test, against a real-world driving situation.

De Geir *et al.* (1981) investigated dual-car driving through the streets of Utrecht, in two parallel groups of patients being seen by their physician for anxiety. They were either being prescribed diazepam (generally 5 g t.d.s.), or acted as controls. The diazepam patients had significantly higher 'insufficient' scores on several of the standardized rating scale items: cutting corners, poor observation at roundabouts, poor anticipation of events, and others. Subjects were also assessed on two laboratory performance tasks. On the low-attention task the time outside the target area was significantly greater in the diazepam group. There was also a significant correlation between 'time outside the target area' on the low-attention task, and 'insufficient items' on the driving test. Overall, therefore, one parameter from the low-attention task was validated against one of the driving assessment measures. It should, however, also be noted that in the high-attention task, false alarms were significantly higher in the control group (De Geir *et al.*, 1981). Does this mean that false alarms are an index of good driving, or that the finding should be ignored since it was not predicted? This illustrates an important problem with criterion validity. Positive correlation between performance test and external criterion will be widely quoted as evidence for the validity of that test. But how should the many non-significant or unexpected findings be explained? The only recourse will be to investigate the pattern of findings across many different studies, one of the processes subsumed within construct validation.

Many archetypal psychoactive drugs have been studied in real-world trials, and laboratory task investigations (e.g. diazepam, alcohol, cannabis, amphetamine). Meta-analyses of these data then allow the findings from real-world and laboratory study to be indirectly compared. For instance with reference to diazepam, O'Hanlon *et al.* (1982) assessed the degree of 'weaving' on 100 km of open-road driving at night, with lateral position measured by an electro-optical device. Lateral position variance (weaving) was significantly increased by diazepam, while mean speed was unchanged. De Geir *et al.* (1981) similarly reported a significant driving impairment under diazepam. Several other studies have demonstrated impaired driving under diazepam (summarized in Parrott, 1987). Similarly, laboratory performance tasks are generally impaired by diazepam (reviewed in Kleinknecht and Donaldson, 1975). Furthermore, Wittenborn (1987), noted significant impairments in each of nine studies investigating tracking accuracy under diazepam (eye tracking, saccadic eye movement or target tracking; Table 2 in Wittenborn, 1987). Overall, therefore, both laboratory tasks and real-world driving skills are impaired by diazepam. In contrast, neither car driving nor laboratory task performance are generally impaired by the 1·5 benzodiazepine clobazam. Rigal and Savelli (1975) found unimpaired performance during 8 h motorway driving under 20 mg clobazam, while EEG measures of alertness were also unchanged. Hindmarch and Gudgeon (1980) reported unimpaired closed-course car driving under clobazam, whereas significant decrements were evident in four out of five driving parameters under lorazepam. In the same study lorazepam produced significant decrements on four laboratory performance tasks, while clobazam produced a significant decrement on one task. This last finding was untypical, with clobazam generally leaving laboratory task performance unimpaired (Hindmarch and Parrott, 1979; Steiner-Chaskel and Lader, 1981). Thus, the evidence from real-world driving situations and laboratory performance tasks was broadly consistent, with sedation under diazepam, but unimpaired performance under clobazam. This might be conceptualized as indirect information for criterion validity, but is perhaps more correctly encompassed within construct validity (see later).

One further problem is the adequacy of the criterion measure, since it needs to be sensitive, reliable and conceptually sound. Anastasi (1982) has shown that 'supervisors' ratings' are often mis-used as an index in industry, being insensitive and biased. Hence an industrial task may show low correlation with the supervisors' ratings, not because it is a poor test, but because the real-world measure comprises a poor external criterion. The difficulty of selecting a good criterion measure for human performance is perhaps the biggest single obstacle to assessing criterion validity (Sanders, 1987). Also, should human performance be conceived as a unitary concept, and therefore should different real-world criteria be expected to correlate positively? if not, then how can any one laboratory task be valid as a measure for performance in general? These problems are raised again in the later section: 'Information processing or performance?'

There are also practical problems in establishing the statistical relationship between the test and criterion. Both measures need to be stable (Anas-

tasi, 1982). This means that random error should be minimized through the use of large samples. Anastasi has stated: 'The application of criterion-related validation is not technically feasible ... with subject samples smaller than 40–50 cases' (Anastasi, 1982; p. 143; also Anastasi, 1986, pp. 10–12). Within human psychopharmacology performance trials rarely exceed a quarter or half that number, thus making the investigation of criterion validity somewhat dubious. A second problem is in statistically comparing tests differing in reliability, since the more reliable test will appear more valid. One solution is to equate tasks for reliability by manipulating task duration (Bittner *et al.*, 1986); once equated for reliability then each can be compared with the external criterion. Lastly, the criterion should be measured separately from the performance tests, with all sources of test/criterion contamination removed. For instance, different teams of experimenters should be employed for the laboratory and field tasks (Stonier *et al.*, 1982). Thes statistical difficulties may be summarized as follows. Low correlation between test and criterion may reflect inappropriate choice of criterion, unreliable criterion, small sample size or genuinely invalid performance test.

Within some areas of psychometrics a new test is correlated with an established test covering the same area (Kline, 1986, p. 4). Often this is undertaken when both the original and new test are difficult to validate properly. Intelligence testing is a prime example of this. The tests have a family resemblance, and generally correlate highly together, but correlations with true external criteria are lower, and more difficult to establish. Essentially this procedure validates the new test as 'another IQ test', but it does not validate it as a true test of 'intelligence'. Within human performance many new tests are quite similar to established tests. This makes them feel and look correct, so they have face validity for the researchers who develop them. However, high correlation with an established test does not indicate true criterion validity, but rather inter-test consistency. Nevertheless, the degree of correlation between the new and old tests should be investigated, since it provides information on convergent/discriminant validation (see: Factor analysis and task discrimination section).

The above studies have shown that it is possible to measure real-world performance under well-controlled conditions (Seashore and Ivy, 1953; Billings *et al.*, 1973; Henry *et al.*, 1974;Billings *et al.*, 1975; Rigal and Savelli, 1975; Hansteen *et al.*, 1976; Hindmarch and Gudgeon, 1980; De Geir *et al.*, 1981). There are, however weaknesses with this body of research. Firstly, they generally failed to investigate the degree of correlation between laboratory task and real-world measure. Secondly, while displaying great ingenuity in measuring the real-world task, less attention has generally been paid to the comparatively easier process of measuring laboratory task performance. Had the above studies each assessed their subjects on a well-structured performance test battery, then far more information would have been generated. Thirdly, they were too small. Studies using 40–50 subjects per condition would allow stronger statistical control, and produce more generalizable and meaningful findings.

DRUG SENSITIVITY AND VALIDITY

> It can be argued that ... if a test designed for use in drug-evaluation studies is shown to be sensitive to both beneficial and detrimental drug effects, then it is a valid measure (Cull and Trimble, 1987, p. 141).

Hindmarch and Bhatti (1987, p. 122) similarly stated: 'The best method for checking the validity of a test would be to run laboratory studies looking at changes in performance following different doses of, for example, a sedative–hypnotic.' This approach probably reflects the views of many human psychopharmacology researchers, since drug sensitivity is an obvious requirement for any test being used in human psychopharmacology. It can also be used to gauge the comparative sensitivity of different tasks. Thus even crude measures should show performance decrements following a large dose of a sedative drug. Whereas only a sensitive test will show measurable performance change following low-dose conditions (Clyde, 1981; Wesnes and Warburton, 1984; Parrott, 1986); dose-related performance change (Wesnes and Warburton, 1984; Hindmarch and Bhatti, 1987; Parrott and Winder, 1989), or subtle differences between drug/stressor conditions (Frowein, 1981; Stonier *et al.*, 1982; Hockey and Hamilton, 1983; Hasenfrantz *et al.*, 1989).

The presence of drug sensitivity is, however, not a true index of test validity. This is because the external criterion is defined by drug presence/absence, and not by performance level. Suppose a 'sedative' drug showed a significant impairment on a 'memory' task. Would that test be validated

as a memory index or a sedation index? In terms of criterion validity the answer is neither. It has not been validated as a memory test, nor as a measure of sedation, but as a measure of drug presence. That is because the external criterion is the presence/absence of the drug. Interpretative constructs are necessary to extend this conclusion to imply performance change, a process subsumed under construct validation. Drug sensitivity is therefore a *necessary* condition for any test being used in human psychopharmacology, but it is not *sufficient*. It is a crucial first stage, but further evidence is required to explain how, and in what ways, that particular measure comprises a performance test.

REFERENCES

References are listed in the final article in this series (Performance tests in human psychopharmacology (3): construct validity and test interpretation), appearing in the next issue of *Human Psychopharmacology*.