HUMAN PSYCHOPHARMACOLOGY, VOL. 6, 1–9 (1991)

REVIEW

Performance Tests in Human Psychopharmacology (1): Test Reliability and Standardization

A. C. PARROTT
Department of Psychology, Polytechnic of East London, (formerly: North East London Polytechnic), London E15 4LZ, UK

Reliability, validity and standardization are well-established principles in most areas of psychometrics, but are rarely mentioned in performance assessment within human psychopharmacology. Test reliability and standardization are examined here, while validity is covered in two succeeding articles. The undocumented reliability of human psychopharmacology performance tests makes the interpretation of findings difficult and ambiguous. Reliability, or consistency within and between test sessions is, however, easy to calculate. Several procedures for calculating reliability are described, with test–retest reliability recommended as the most appropriate single summary measure. Poor test standardization is another major problem, with many research groups using different tests. The extent of this problem is examined, and a set of standardization requirements proposed. They could comprise the basis for test manuals. The problems of undocumented reliability and unstandardized tests have been recognized within the field for many years, yet no moves have been made to remedy the situation. One simple solution would be for psychopharmacology journals to accept only documented tests, and to request test–retest reliability data.

KEY WORDS—Psychometrics, reliability, assessment, test, standardization, performance, psychopharmacology.

BASIC PSYCHOMETRIC PRINCIPLES

> 'Psychometric testing sums up performance in numbers. If a thing exists it exists in some amount. If it exists in some amount it can be measured' (Cronbach, 1984, p. 41).

The application of psychometric principles is routine in many areas of psychology (e.g. personality questionnaire, clinical rating scale), but in human psychopharmacology it is rare. Cull and Trimble (1987, p. 141) have commented: 'Many of the automated performance tests used in the psychopharmacological research discussed here have not been well evaluated as regards their psychometric properties, particularly with respect to reliability and validity'. Others have also noted the absence of psychometric principles in this area. Wittenborn (1987) concluded that: 'The behavioural concepts served by the various tests are obscure, incompletely defined and poorly articulated.' Furthermore, because of non-standardized testing practices, differences in tests between research groups, and the failure to establish the meaning of different tests: 'It is not realistic to draw a usefully standard body of information concerning drug-related psychomotor impairments from published current literature.' In a similar vein Hindmarch (1989) has stated: 'The exponential growth of the subject area has resulted in a proliferation of tests and assessment systems purporting to be accurate measuresof psychoactive drug effects in man. In many instances the 'tests' owe more to the, often whimsical, ingenuity of their progenitors than to any serious attempt to provide a proper psychometric. Such tests have no foundation in either a methodological or theoretical framework.'

The aim here is to critically examine the 'methodological and theoretical framework' of the tests used in human psychopharmacology, and to put forward practical recommendations for test reliability and standardization (present article), and test validity (succeeding articles).

CURRENT SITUATION

Complaints about the poor psychometric analysis of tests have been made repeatedly, yet none of the above reviews documented the scope of the problem. This reflects the broad nature of this area, making a comprehensive survey logistically difficult. In order to estimate the frequency with which psychometric data are being presented, a sample of human psychopharmacology papers involving performance tasks was analysed. Since each research group tends to use the same test battery, only one study from any research laboratory (the

0885–6222/91/010001–09$05.00

Accepted 20 August 1990

first encountered during the literature search) was included. Thirty-eight papers, from sixteen journals, involving 115 test versions, and all the major drug types, formed the resulting database. The publication dates ranged from 1972 to 1988. None of these studies documented either reliability or validity. One report mentioned test reliability, noting that it should be investigated. Since each group also tended to use their own battery, very few tests were used by more than one centre (although this was often difficult to assess, since test descriptions were generally very brief). While task validity was not documented in any publication, it was mentioned in several ($n = 9$); generally as a brief note that test validity had not been established. Most of the papers where these problems were acknowledged were *not* mainstream psychopharmacology journals. This perhaps identifies the core of the problem: that psychopharmacology researchers, and the journals in which they publish (edited by peers/colleagues), have not been concerned with documenting test reliability or validity.

The two broad reasons for assessing performance in human psychopharmacology are practical and theoretical. On the practical side, it is necessary to determine the level of drug-induced side-effects upon everyday skills, such as car driving and operating machines (Hindmarch, 1980; Cull and Trimble, 1987; Parrott, 1987; Wesnes *et al.*, 1987; Wittenborn, 1987). The theoretical rationale is concerned with the neurochemical basis of psychological functioning (Warburton, 1975; Ashton, 1987; Stahl *et al.*, 1987). For both these objectives, meaningful and consistent data are required. Test assessments must therefore be psychometrically sound. The basic psychometric principles of reliability and test standardization are examined here. Detailed coverage of these topics can be found in: Anastasi (1982), Cronbach (1984), Kline (1986) or Jones and Appelbaum (1989).

RELIABILITY

> 'Test reliability covers several aspects of consistency ... it indicates the extent to which differences in test scores are attributable to true differences in the characteristic under consideration, or to chance errors' (Anastasi, 1982).

Reliability is indicated by the consistency of two independent sets of scores. Identical sets of scores produce perfect correlation ($r = +1{\cdot}00$); the lower the correlation the less reliable the test until zero correlation ($r = 0{\cdot}00$) is reached. Divergence from perfect correlation can be attributed to the operation of other factors, collectively termed the 'error variance' (Anastasi, 1982; Cronbach, 1984). The different procedures for computing reliability each quantify different aspects of the error variance. For instance, test–retest reliability samples time variance (variation between test sessions). Internal consistency samples content variance (test item differences). Alternate form reliability samples further aspects of content variance (test item variation as before, but also imperfect matching between test forms). Inter-rater reliability samples variance due to differences in scoring techniques. Inter test-presenter reliability samples variation in methods of test presentation (Anastasi, 1982). In addition to these, all measures of reliability are affected by the normal fluctuation in individual performance over time, e.g. changes in interest, motivation, skill, distraction, or fatigue. These can affect performance independently of the test instrument, and comprise further sources of error variance.

The reliability coefficient is dependent upon a number of physical characteristics of the task: duration, frequency/regularity/probability of stimulus events, number of responses required, and response simplicity. For instance, the longer the test, the higher the reliability estimate (Bittner *et al.*, 1986). With tests of the same overall duration, greater reliability will follow from the test requiring more responses. Thus 4 minutes of testing may generate 200 responses on a symbol-coding task, but only one response on a vigilance task. Tasks generating regular and simple responses will also tend to produce greater reliability.

TEST–RETEST RELIABILITY

> 'The measurement of test-retest reliability is essentially simple. The scores from a set of subjects tested on two occasions are correlated' (Kline, 1986).

The procedure for calculating test–retest reliability is described in Appendix 1. A test with high reliability will produce similar scores on different occasions. Obviously this is important for all assessment devices (Anastasi, 1982; Cronbach, 1984), but it is particularly crucial in repeated-measures analysis of variance (ANOVA) designs, where stable inter-trial correlation and stable variance are technical requirements (Edwards, 1985). In the non-psychopharmacological literature, test–

retest reliability figures are sometimes given for tests similar to those used in psychopharmacology research. For instance, Bittner *et al.* (1986) collected test–retest reliability data on 140 tests in their: 'Performance Tests for Environmental Research (PETER)' program. Perceptual and simple cognitive information processing tasks predominated, although motor tasks and arcade games simulations were included. Many tests demonstrated high test–retest reliability (greater than +0·80): logical reasoning, letter cancellation, code substitution, four-choice reaction time, while many other tests had lower reliability (Table 1). Secondarily derived parameters such as Stroop colour/B&W word difference, Sternberg information slope, and choice reaction time information slope, were particularly unreliable (Table 1). It should be noted that these tasks were rewritten for the local (PETER) computer system, and often differ considerably from the original tests. Also some aspects of the data from this study were surprising, and possibly atypical (e.g. low reliability for auditory digit-span, tracking, simple reaction time). This emphasizes that reliability has to be calculated for each test, and applies only for the situation under which it was given. The duration of the test–retest interval may also affect reliability. Bittner *et al.* (1986) tested subjects on a daily basis, a procedure mimicking many psychopharmacology trials. However, in other studies subjects are tested once per week, or even less frequently, and test–retest reliability estimates under these conditions would be lower.

Table 1. Test–retest reliability coefficents normalized for a three-minute administration period (from Bittner *et al.*, 1986)

Assessment measure	Reliability coefficent
Stroop: time to name colour words	+0·97
Logical reasoning time	+0·93
Arithmetic vertical addition	+0·90
Letter search	+0·87
Aiming hand–eye coordination	+0·87
Code substitution	+0·84
Perceptual speed: number comparison	+0·84
Four-choice reaction time	+0·80
Sternberg item recognition (set 4)	+0·80
Manakin test mental rotation	+0·79
Minnesota manual dexterity: turning	+0·64
Air combat manoeuvring: Atari simulation	+0·63
Letter classification memory: LTM name retrieval	+0·55
Target tracking accuracy: two-dimensional	+0·52
Memory: free recall	+0·52
Stroop: colour/word naming difference	+0·47
Choice reaction time: information slope	+0·41
Navigational plotting accuracy	+0·40
Sternberg item recognition: information slope	+0·11

Psychopharmacological research is concerned with drug effects. Thus while Table 1 presented baseline scores, the crucial statistic of interest comprises the reliability of the pre/post-drug or placebo/drug difference scores. These do not seem to have been reported, even though only tests demonstrating reliable drug effects can be considered useful. The reliability of drug effect scores should therefore be reported. There are, however, difficulties in its calculation. Reliability estimates are: 'Affected by the *range* of individual differences in the group' (Anastasi, 1982). It is therefore not appropriate to use scores from a single drug condition, since that drug effect should be consistent (a given dose of drug should have a narrow range of effect across all subjects; i.e. within-drug variance should be low). Scores from a range of sedative and alerting drugs would therefore need to be used in order to maximize the 'drug' variance. This will then maximize the 'drug/error' variance ratio, or reliability estimate.

INTERNAL CONSISTENCY RELIABILITY

> 'From a single test presentation ... it is possible to arrive at a measure of reliability by various split-half procedures' (Anastasi, 1982, p. 113).

With split-half reliability, the subscore for half the test (e.g. the odd-numbered stimuli), is correlated with the subscore for the other half (e.g. even-numbered items). The resulting correlation indicates the reliability of a half test, so the Spearman–Brown formula needs to be applied (Anastasi, 1982; Cronbach, 1984). These procedures are summarized in Appendix 2. Other measures of internal consistency include the Kuder–Richardson 20, and Cronbach's coefficient alpha. They are particularly appropriate for computerized test procedures; Kline (1986, ch. 5) describes them more fully.

Internal consistency, as with the other types of

reliability estimate, is rarely reported in human psychopharmacology. However, the available evidence suggests that it should generally be high. This is because many of the performance tests used in this research area are simple and repetitive, conditions which maximize response consistency. Parrott (1982) reported split-half reliability of $r = +0{\cdot}97$ with the critical flicker fusion test; this was based on a meta-analysis of five psychopharmacology studies involving single tests from 101 subjects. As stated earlier, internal consistency samples the content variance, i.e. test item differences. Since most performance tests use extremely similar stimulus items (e.g. coding, subtraction, Stroop, choice reaction time, logical reasoning, digit span, mental rotation, etc.), internal reliability should be very high. Anastasi (1982, p. 121) also suggests that this type of reliability estimate is not appropriate for speeded tests: 'Single trial reliability coefficients are inapplicable to speeded tests ... reliability coefficients found by these methods will be spuriously high.' She recommends test–retest reliability, or alternate form reliability, when appropriate (Anastasi, 1982).

Internal consistency is a requisite for tests of narrow, well-defined skills (e.g. symbol copying, typing speed, subtraction). It may be less desirable when the attribute being assessed is broad (e.g. overall cognitive ability or 'intelligence'; the whole personality; psychomotor skill in general). Cattell suggests that, with tests covering a wide topic area, the underlying psychological factors being assessed are largely independent, so that inter-item correlation should *not* be high (Cattell and Kline, 1977). This can be illustrated by the low correlations generally found between 'intelligence' and 'creativity' tests, also by the low correlations frequently noted amongst different personality tests (e.g. extraversion and neuroticism).

The implications for human performance testing are that any assessment attempting a broad coverage will require subtests which will not necessarily correlate highly together. Most broad areas are not covered by a single test, but by a test battery. For instance, the recently developed British Ability Scales comprise a battery of separate diagnostic tasks, which can either be used independently, or combined to produce an overall 'IQ' index (Elliott *et al.*, 1978). In personality assessment the 16PF questionnaire comprises a number of subscales which provide both local-factor and overall-profile information (Cattell and Kline, 1977). Similarly in human performance assessment, a battery of tests covering different functions is commonly used (Cull and Trimble, 1987; Hindmarch, 1980; Johnson and Chernik, 1982; Parrott, 1986; Wesnes *et al.*, 1987; Wittenborn, 1987). In these batteries each test or subscale should display high internal consistency, while inter-test correlations should ideally be lower. This has been demonstrated for the British Ability Scales and the 16PF, but does need to be investigated for human performance test batteries.

ALTERNATE FORM RELIABILITY

> 'The same persons can be tested with one form on the first occasion and with another, comparable form on the second. The correlation between the sets of scores obtained on the two forms represents the alternate-form reliability' (Anastasi, 1982).

Many performance tests, particularly those involving repetitive information processing, have been prepared in sets of matched forms: letter cancellation, code substitution, mental arithmetic, logical reasoning, and many memory tests. Both pencil-and-paper tests and computerized tests are often produced in parallel or multiple versions. Alternate form reliability should then be investigated, but again it rarely seems to be presented.

A related aspect of particular relevance to human psychopharmacology is the use of different versions of an apparently standard test. Thus all the tests noted above (letter cancellation, code substitution, mental arithmetic, logical reasoning, and others), are used in many psychopharmacology laboratories. Yet the actual tests differ between groups. Pencil-and-paper versions become retyped, shortened, or modified in other ways, while response and scoring systems may also differ. Computer versions become re-edited for new machines, leading to changed stimulus presentation and response production/measurement (or computer-clock systems). The cross-correlation of test versions having the same name should therefore be investigated. While many test versions may *appear* to be quite similar, tests masquerading under the same name may contain important differences, and their intercorrelation might be surprisingly low.

INTER-USER RELIABILITY

The calculation of inter-rater reliability is standard with subjective rating scales (Anastasi, 1982). Reliability estimates are, however, also required

when an identical test is given by different administrators or research groups. Potential sources of variation include:

(1) Instructions: these are rarely standardized, and can therefore vary in contents (speed/accuracy emphasis), or mode of presentation (written/verbal; given carefully/rushed).
(2) Experimenter effects: age, sex, personality and attitude are of potential importance.
(3) Subject effects: differences in age, sex, intelligence, task practice, interest/motivation and remuneration, may each affect performance.
(4) Data collection: manual timing accuracy can vary. Error scoring by hand invariably involves inaccuracies. The manual scoring of repetitive results sheets can be seen as a vigilance task, with its well-recognized suboptimal performance (Warm, 1984). Calibration may be necessary for automated tests (e.g. joystick centring in target tracking).
(5) Test program: variation in stimulus characteristics, task duration, and response requirements, are frequent. Many tests exist in different versions, with each research group tending to rewrite/reprogram tests for their own particular requirements (see above). Computerized test programs may vary in the definition of error types.
(6) Test situation: subjects may be tested individually or in groups, with tests presented singly or as part of a battery. In field situations, control over extraneous variables may be difficult: reaction time with hospital patients in bed (Herbert *et al.*, 1983); target tracking and critical flicker fusion with soldiers in foxholes (Seashore and Ivy, 1953); code substitution and letter cancellation aboard ships in stormy seas (Parrott and Jones, 1985).

Overall, there are numerous ways in which data collection can vary between studies. Therefore even with a test of established reliability, its reliability should always be calculated afresh under the particular conditions pertaining within that study.

PROCEDURES FOR IMPROVING RELIABILITY

Many sources of unreliability would be reduced by the use of standardized tests (see below). Several further practical improvements can be suggested which would improve reliability.

(1) With pencil-and-paper tests, two scorers should independently score each test, without placing marks on the test sheet. Their scores should be separately entered onto the computer for automatic cross-checking.
(2) The timing of manually administered tests can be standardized by the use of a timer with an automatic start/stop signal.
(3) In automated tests error definition needs to be explicit; e.g. characteristics of omission, commission, perseveration or random error.
(4) The treatment of missed/repeated targets should be described. Also the time cutoff employed, and whether a 'long' response is included in the calculation of the mean, deleted or repeated, or counted as an error. Wesnes *et al.* (1987, p. 82–84) discuss the treatment of 'long' responses.
(5) In task design floor/ceiling effects should be avoided, since low error scores are unreliable. This can be accomplished by writing the task so that errors are frequent (by increasing speed/difficulty), or by avoiding errors completely (the wrong response remains on screen until correct, with the number of attempts being recorded).
(6) Pre-trial practice should be extensive, so that the initial steep part of the learning curve is avoided within the study. One way of presenting the final practice session is as a proper data-collection session (using additional placebos: later discarded). Parrott and Wesnes (1987) used this to ensure that learning about the testing procedures was finalized before the trial proper.
(7) Microprocessors provide more standardized stimulus/response conditions than pencil-and-paper methods of presentation (Cull and Trimble, 1987, p. 141), and should therefore lead to greater reliability through control of the error variance.

The calculation of reliability coefficients would not necessitate any change to most experimental designs. Internal consistency can be calculated for any test simply by recording the variance of each subject's data; Cronbach describes a simple procedure for this (Cronbach, 1984, pp. 167–171). Thus the reliability of trials can be calculated, as long as the database contains the variance of each subject's scores (and not just the mean score). Past trials could be reanalysed if they had these data. Test–retest reliability can be calculated for any

repeated-measures design involving pre-drug testing; the scores for the different pre-tests need to be correlated. Alternate form reliability can also be readily calculated. The only type of reliability requiring a modified experimental design would be the calculation of drug effect reliability, since drug conditions would need to be replicated.

TEST STANDARDIZATION

> 'Standardization: the stimulus situation is controlled, reproducible, applicable in a nearly uniform manner to everyone' (Cronbach, 1984, p. 538).

It takes little knowledge of psychometrics to recognize the problems of using unstandardized and inadequately documented tests. In personality assessment, clinical rating scales, ability testing, and many other fields of psychology, only standard versions of documented tests are accepted for general use. This obvious requirement is necessary for the basic scientific tenet of replicability. Findings from different research reports can be compared with ease only if they have used the same measurement device.

It is therefore remarkable that the tests used in human psychopharmacology comprise such an *ad hoc* collection of unstandardized and poorly documented procedures. The reasons for this state of affairs are complex, but include the following. Most important is the absence of test standardization in mainstream human performance research. This perhaps follows from its long history, with many current tests (e.g. letter cancellation, coding) coming from the era of Wundt. The ethos of these pioneers, for each laboratory to devise their own test, surprisingly remains to this day, yet Wundt and colleagues would be the first to criticize the present poorly organized scenario. There is also no predominant theory of human performance or performance testing, into which new tests can be accommodated (Parrott, unpublished). Moving more specifically to human psychopharmacology, the recent increase in drug research has coincided with an expansion in (competing) research groups. They have developed their own tests, and generally use only their own versions. Hindmarch started his (1980) review of performance testing with a lengthy list of some of the *ad hoc* measures then available. Few of them were then used by more than one research group. With the increase in microcomputers, writing performance tests has increased as a pastime for human psychopharmacology researchers, and many new tests could be added to Hindmarch's (1980) compendium.

On initial inspection it might appear that several tests are in widespread use: simple reaction time, choice reaction time, target tracking, logical reasoning, critical flicker fusion, symbol coding, simple mathematics, letter cancellation, auditory vigilance, mental rotation, and others. However, these tests are generally identical only in name. Closer inspection reveals frequent differences between the many test versions. For instance, there are numerous visual reaction time apparatuses, which differ in: structure, size, nature of the stimulus (LED, other light-bulb stimulus, or computer-controlled VDU event), number/position and size/brightness of stimulus lights, number/size/position of response keys, type of response required (lift/move finger, hand movement as well, response key release/press/contact, etc.). Similar variation is apparent in the numerous target tracking tasks, and critical flicker fusion devices (Bobon and Holmberg, 1982). With pencil-and-paper tests, standard test versions are probably more widely used, yet all too often each laboratory develops its own test version. Stimulus sheets are retyped or abbreviated, with alterations in the number/size/clarity of the stimuli. Many groups develop their own tests from first principles, since it is comparatively straightforward to type letter cancellation sheets, devise a new version of the Stroop task, or program a mental rotation test. Even where a standard set of stimulus-and-response requirements exist, and remain unchanged, that task may be administered for a shorter duration, be given under different instructions, or scored using new rules.

Some standardized tests from the non-psychopharmacology area have become quite widely used in psychopharmacology research: logical reasoning (Baddeley, 1968); Wilkinson continuous reaction time (Wilkinson and Houghton, 1975). Similarly, some standardized tests have been developed within human psychopharmacology: rapid visual information processing (RVIP) (Wesnes and Warburton, 1983), Leeds psychomotor tester (Hindmarch, 1980). But even these 'standardized' tests have been changed or developed. For instance, the RVIP task comprises a more sensitive version of an earlier vigilance measure (Bakan, 1959), while the RVIP task has itself been developed into a variable-time constant-error version (Hasenfrantz *et al.*, 1989). Similarly, the Wilkinson continuous performance task exists in both portable four-choice,

and a static five-choice version. The configuration of the task has also evolved with developments in electronics and microprocessors. Many versions of the basic apparatus have been built by different technical services departments. Other tests have no standard bench mark version (target tracking, critical flicker fusion, coding, simple mathematics, letter cancellation, and many more), and their variety is even more marked.

Despite widespread recognition of the problem (Hindmarch, 1980, 1989; Cull and Trimble, 1987; Wittenborn, 1987), there have been few proposals on how to remedy the situation. This is surprising, since remedies are straightforward, although painful in the short term. Cronbach (1986, p. 124), in writing about test standards, has commented: 'No-one has laid down general rules about test quality. What is called for is information and evidence.' This summarizes the approach which should be followed here, since it would be misguided to attempt to establish a 'restricted list' of acceptable tests. The current diversity of tests, good and bad, should be encouraged; but each test should be properly documented. This documentation will then allow its quality to be assessed. Good tests (reliable/sensitive) will then become more widely used, while poorer tests will fall into disuse. Test documentation will also discourage the needless writing of new test versions, since many researchers are forced into writing tests because of the non-accessability of standard versions.

Test documentation should be available in a publication or manual. It might comprise an internal manual prepared by the researcher, a commercial publication, or a paper in a scientific journal. It would need to contain information along the following lines:

(1) Test instructions: for subjects. Either written instructions for pencil-and-paper tests, or transcriptions of the instructions on the VDU screen, in computer versions.
(2) Written instructions: for test administrators. Proper documentation and instructions for the administrators of computer tests. Similarly for pencil-and-paper tests.
(3) Test duration: recommended standard duration.
(4) Copies of all stimulus sheets.
(5) Response requirements.
(6) Practice sessions: 'Three 10-minute practice sessions, on separate days, before trial commencement', or similar.
(7) Error scoring: full description for pencil-and-paper tests. With automated tests, errors needs to be clearly defined, both in the test manual and computer program (e.g. characteristics of omission, commission, perseveration, or random error).
(8) Response time: timing systems, and the treatment of missed/repeated targets.
(9) Computer program (optional): full documentation, particularly where test development is encouraged.
(10) Reliability data: comprehensive list, not a selection of the high values. Test–retest reliability is the most important.
(11) Validity evidence (Parrott, unpublished).
(12) Test access: whether test is freely available, or to be purchased.
(13) Publication list: papers using that test.

The above evidence would allow researchers to assess the utility and quality of the test. The decision to use that test, or an alternative, could then be made on an informed basis. Publications would then reference that primary source for full documentation. Any test modifications (a procedure to be discouraged) could then be described within the individual article.

RECOMMENDATIONS

The main problem in human psychopharmacology is not recognition of the problem, but a willingness to undertake the steps necessary to sole it. Bureaucratic or restrictive systems might be advocated, but would be difficult to devise, almost certainly disputed, and would rarely be implemented. One simple solution would be for psychopharmacology journal editors to require articles using performance tests to have full documentation, or reference to a test manual. Similarly, the test–retest reliability coefficients attained in that study should be requested.

REFERENCES

Anastasi, A. (1982). *Psychological Testing*, 5th edn. Macmillan, New York.

Ashton, H. E. (1987). *Brain Systems, Disorders, and Psychotropic Drugs*. Oxford University Press, Oxford.

Baddeley, A. D. (1968). A three minute reasoning test based on grammatical transformation. *Psychonomic Science*, **10**, 341–342.

Bittner, A. C, Carter, R. C, Kennedy, R. S, Harbeson, M. M. and Krause, M. (1986). Performance evaluation tests for environmental research (PETER): evaluation

of 114 measures. *Perceptual and Motor Skills*, **63**, 683–708.

Bobon, D. P. and Holmberg, G. (1982). Critical flicker fusion in psychopharmacological research (symposium). *Pharmacopsychiatry*, **15**, (whole issue).

Cattell, R. B. and Kline, P. (1977). *The Scientific Basis of Personality and Motivation.* Academic Press, London.

Cronbach, L. J. (1984). *Essentials of Psychological Testing*, 4th edn. Harper and Row, New York.

Cull, C. and Trimble, M. R. (1987). Automated testing and psychopharmacology. In: *Human Psychopharmacology: Measures and Methods*, vol 1, (Hindmarch, I. and Stonier, P. D. (eds). John Wiley, Chichester.

Edwards, A. L. (1985). *Experimental Design in Psychological Research.* Harper & Row, New York.

Elliott, C. D. Murray, D. J. and Pearson, L. S. (1978). *British Ability Scales.* NFER, Windsor, UK.

Hasenfrantz, M, Michel, C, Nil, R. and Battig, K. (1989). Can smoking increase attention in rapid information processing during noise? Electrocortical, physiological and behavioural effects. *Psychopharmacology*, **98**, 75–80.

Herbert, A, Healey, T. E. J., Bourke, J. B, Fletcher, I. R. and Rose, J. M. (1983). Profile of recovery after general anaesthesia. *British Medical Journal*, **286**, 1539–1544.

Hindmarch, I. (1980). Psychomotor function and psychoactive drugs. *British Journal of Clinical Pharmacology*, **10**, 189–209.

Hindmarch, I. (1989). Psychometrics and psychopharmacology: editorial. *Human Psychopharmacology*, **4**, 79–80.

Hindmarch, I. and Bhatti, J. Z. (1987). Recovery of cognitive and psychomotor function following anaesthesia. In: *Aspects of Recovery from Anaesthesia*, Hindmarch, I., Jones, J. G. and Moss , E. (eds). John Wiley, Chichester.

Johnson, L. C. and Chernik, D. A. (1982). Sedative hypnotics and human performance. *Psychopharmacology*, **76**, 101–113.

Jones, L. V. and Appelbaum, M. I. (1989). Psychometric methods. *Annual Review of Psychology*, **40**, 23–43.

Kline, P. (1986). *A Handbook of Test Construction.* Methuen, London.

Parrott, A. C. (1982). Critical flicker dusion thresholds and their relationship to other measures of alertness. *Pharmacopsychiatry*, **15**, 39–43.

Parrott, A. C. (1986). The effects of transdermal scopolamine and four doses of oral scopolamine (0·15, 0·3, 0·6, 1·2 mg) upon psychological performance. *Psychopharmacology*, **89**, 347–354.

Parrott, A. C. (1987). Assessment of psychological performance in applied situations. In: *Human Psychopharmacology: Measures and Methods*, vol. 1, Hindmarch, I., and Stonier, P. D. (eds). John Wiley, Chichester.

Parrott, A. C. (unpublished). Performance test in human psychopharmacology (2): Content validity, criterion validity, and face validity. (3): Construct validity and test interpretation. *Human Psychopharmacology* (In press).

Parrott, A. C. and Jones, R. (1985). Effects of transdermal scopolamine upon psychological performance at sea. *European Journal of Clinical Pharmacology*, **28**, 419–423.

Parrott, A. C. and Wesnes, K. (1987). Promethazine, scopolamine and cinnarizine: comparative time course of psychological performance effects. *Psychopharmacology*, **92**, 513–519.

Seashore, R. H. and Ivy, A. C. (1953). The effects of analeptic drugs in relieving fatigue. *Psychological Monographs*, **67** (365), 1–16.

Stahl, S. M., Iversen, S. D. and Goodman, E. C. (1987). *Cognitive Neurochemistry.* Oxford University Press, Oxford.

Warburton, D. (1975). *Brain Behaviour and Drugs.* John Wiley, Chichester.

Warm, J. S. (1984). *Sustained Attention in Human Performance.* John Wiley, Chichester.

Wesnes, K. and Warburton, D. (1983). Effects of smoking on rapid information processing performance. *Neuropsychobiology*, **9**, 223–229.

Wesnes, K, Simpson, P. and Christmas, L. (1987). The assessment of human information processing abilities in psychopharmacology. In: *Human Psychopharmacology: Measures and Methods*, vol. 1, Hindmarch, I. and Stonier, P. D. (eds). John Wiley, Chichester.

Wilkinson, R. T. and Houghton, D. (1975). Portable four choice reaction time test with magnetic memory tape memory. *Behaviour Research, Methods and Instrumentation*, **7**, 441–446.

Wittenborn, J. R. (1987). Psychomotor tests in psychopharmacology. In: *Human Psychopharmacology: Measures and Methods*, vol 1, Hindmarch, I. and Stonier, P. D. (eds). John Wiley, Chichester.

APPENDIX 1: CALCULATION OF THE TEST–RETEST RELIABILITY COEFFICIENT

Mean logical reasoning test error scores, for the pre-drug administration data from two test sessions.

Subject Number	Mean error Scores	
	Session X	Session Y
1	6	11
2	5	7
3	13	11
4	6	3
5	2	1
6	12	14
7	3	6
8	10	11
9	8	9
10	10	7

The formula for calculating the correlation between sessions X and Y is as follows:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Where: X = scores on session X
Y = scores on session Y
N = number of subjects.

In the above example the correlation coefficient is as follows:

$$r = +0{\cdot}76$$

APPENDIX 2: CALCULATION OF THE INTERNAL CONSISTENCY, OR SPLIT-HALF RELIABILITY COEFFICIENT

The single test is divided into two halves, e.g. odd-numbered items and even-numbered items; alternatively, first-half scores, second-half scores. The mean performance value from each half is then calculated. The mean half-test scores from all subjects are then correlated, treating the two halves as two variables (as in Appendix 1 for test–retest reliability). This coefficient indicates the reliability of the half test. The Spearman–Brown formula then needs to be applied to indicate the reliability of the whole test.

$$\text{Spearman–Brown formula } r_{tt} = \frac{2r_{hh}}{1 + r_{hh}}$$

Where: r_{hh} = reliability coefficient for half test
r_{tt} = reliability coefficient for whole test

For example, if the half test reliability, $r_{hh} = +0{\cdot}80$, then the full test reliability will be, $r_{tt} = +0{\cdot}88$.

APPENDIX 3: CALCULATION OF THE ALTERNATE FORM RELIABILITY COEFFICIENT

The two alternate forms are given to all subjects under similar conditions on two occasions. The mean performance scores from the two tests are then treated as the two variables. The correlation between them is calculated as in Appendix 1 for test–retest reliability. This value indicates the alternate form reliability.