# 19 Assessment of executive dysfunction

John R. Crawford and Julie D. Henry

## Abstract

Executive deficits typically have a much more profound effect on recovery and adjustment than the more circumscribed deficits that arise from posterior lesions. However, the behavioural features of executive dysfunction have proven hard to capture formally. In keeping with the emphasis in this book on the use of quantitative evidence to guide practice, this review focuses on the measurement properties of putative tests of executive dysfunction and on validity information (e.g. data comparing anterior lesion cases with controls or posterior cases are used to calculate effect sizes for commonly used tests). The tests reviewed range from long-standing clinical tests (e.g. verbal fluency and the Wisconsin Card Sorting Test) to more recent tests that are more explicitly derived from theory, such as the Cognitive Estimation Task, the Brixton and Hayling Tests, dual task methods, and the Behavioural Assessment of the Dysexecutive Syndrome (BADS). The issue of the ecological validity of tests is discussed as is the need to consider a patient's premorbid ability when assessing executive functioning. Finally, the rating scales and questionnaire methods of assessing executive problems and disability (e.g., the DEX, PRMQ and FrSBe) are briefly reviewed.

## Assessment of executive dysfunction

### Executive deficits: serious problems and seriously problematic

Executive deficits arising from damage to the prefrontal cortex and related structures typically have a much more profound effect on a client's prospects for successful adjustment and independent living than the more circumscribed deficits arising from posterior lesions. However, these deficits have proven difficult to quantify and as such can be regarded as the most problematic area in neuropsychological assessment (Crawford *et al.* 1998).

In a rehabilitation setting the reasons for conducting a comprehensive assessment of executive functioning are largely self-evident. Given the impact of executive problems on a client's quality of life, rehabilitation efforts are often targeted directly at the executive problems themselves; this cannot be done successfully without identifying the nature and quantifying the severity of such problems. In addition, the presence or absence of significant executive problems are important determinants of the approach taken to rehabilitation of other functions. Moreover, when attempting to arrive at a formulation of what may appear to be other, more specific difficulties, it is crucial to consider the extent to which they may be a reflection of a broader executive dysfunction.

For example, as Bradley, Kapur and Evans note (Chapter 11, this volume), a patient's everyday memory problems may largely stem from difficulties in self-organisation and initiation rather than represent a core memory deficit. An empirical demonstration of the need to consider executive deficits in this context was provided by Crawford *et al.*'s (2000b) study of executive dysfunction in Huntington's disease (HD). They reported that the HD sample was severely impaired on the California Verbal Learning Test (CVLT) (Delis *et al.* 1987). However, controlling for executive dysfunction (using a composite measure of performance on executive tasks) completely abolished the group differences on the CVLT; group membership (i.e. HD versus controls) accounted for 90 per cent of the variance in CVLT performance, this fell to 0.001 per cent when executive dysfunction was controlled for. Furthermore, this effect was specific to the executive composite as large group differences in memory remained when general intellectual ability was controlled for using WAIS-R IQ. Similar effects, although less dramatic, were reported by Crawford *et al.* (2000a) in their study of memory, executive functioning and general ability in normal ageing.

Clinical skills and experience are of greater importance in the assessment of executive problems than in any other area of neuropsychological assessment. However, in keeping with the overall emphasis of this book on the quantitative evidence-base for clinical practice, and in view of space constraints, this chapter will focus primarily on evaluating formal ability tests and rating scales of executive functioning in terms of their measurement properties and their validity.

One common means of assessing the validity of putative executive measures is to compare the performance of patients with frontal lesion against healthy controls and posterior cases; the presence of a frontal lesion is taken as a proxy for the presence of executive problems (although all would recognise that it is a very imperfect proxy, as is the use of posterior cases as a proxy for the absence of executive problems). In the present work particular emphasis will be placed on quantifying the effect sizes for such studies and those from other types of validity studies. It is increasingly recognised that there has been an overemphasis on significance tests and a consequent neglect of the magnitude of effects (American Psychological Association, 2001). However, despite strongly worded recommendations to report effect sizes, this is still rarely done in research papers and test manuals.

A simple and commonly used index of effect size is *r*, the (point-biserial) correlation between group membership and test performance. This effect size can readily be calculated from routinely reported summary statistics (the means and SDs of the groups being compared). Although not commonly used for reporting purposes, the square of this effect size gives an even more meaningful measure; it tells us the proportion of variance on the measure of interest that can be attributed to group differences.

Effect sizes are particularly useful when evaluating and comparing neuropsychological tests. A highly significant difference between controls and a patient sample (or between two patient samples) may nevertheless be associated with a modest effect size (particularly if the *N*s were large) and may still mean that the test will have limited utility when used in the individual case; i.e., the overlap in score distributions may still be very substantial. Similarly, when evaluating the sensitivity of two or more tests, the results of their individual significance tests is not very informative whereas expressing the group differences as effect sizes is immediately enlightening (for example, see the presentation of effect sizes for BADS subtests in Table 19.3; there were significant group differences on all these subtests but it can be seen that the magnitude of effects vary substantially).

The tests to be reviewed were selected on the basis that they are either currently used widely in clinical practice or have actual or at least potential advantages over their more common counterparts; Table 19.1 summarizes some of the strengths and weaknesses of these tests. The review will commence with two clinical tests and then move on to consider tests stemming from theories of the executive system (see Burgess and Simons, Chapter 18 this volume) and those aimed at providing measures that possess superior ecological validity (i.e. relate to everyday problems); happily some tests are grounded in theory *and* exhibit this latter quality.

**Table 19.1** Summary of some strengths and weaknesses for selected tests of executive dysfunction

| Test | Strengths | Weaknesses |
|---|---|---|
| Wisconsin Card Sorting | Extensive research base<br>Good norms<br>Moderate sensitivity<br>Moderate ecological validity | Poor specificity<br>Potentially confusing for clients |
| Verbal fluency | Extensive research base<br>Good norms<br>High reliability<br>Quick and easy to administer and score<br>Moderate sensitivity<br>Normally distributed<br>Moderate ecological validity | Low specificity<br>Highly influenced by premorbid verbal IQ |
| Cognitive Estimation | Derived from theory | Poor sensitivity<br>Poor specificity<br>Poor ecological validity<br>Poor psychometric properties<br>Poor norms |
| Brixton Spatial | Derived from theory | Modest normative sample |
| Anticipation Test | Moderate sensitivity<br>Moderate specificity<br>Quick and easy to administer and score<br>Normally distributed | Coarse-grained scoring (Sten scores)<br>Limited research base as yet |
| Hayling Sentence | Derived from theory<br>Moderate sensitivity<br>Moderate specificity | Modest normative sample<br>Coarse-grained scoring (Sten scores)<br>Limited research base as yet |
| Behavioral Assessment of the Dysexecutive Syndrome | Derived from theory<br>Very high ecological validity<br>Moderate sensitivity (six elements) | Limited research base as yet<br>Low sensitivity (most subtests)<br>Specificity unknown |
| Dual task methods | Derived from theory<br>High ecological validity<br>Good specificity | Not yet fully standardized and normed<br>Potential problem with unreliability |

## An old warhorse: The Wisconsin Card Sorting Test (WCST)

The WCST (Grant and Berg 1948; Heaton 1981) has complex task demands but primarily measures concept formation, the ability to shift between these concepts, and the ability to utilise feedback to modify responses. Testees have to sort cards by the attributes (colour, shape and number of objects) they share with a set of stimulus cards. The rule to be applied is not specified and changes as the test progresses; testees are informed whether each card sort is wrong or right. A modified version of the Wisconsin was developed by Nelson (MCST) (Nelson 1976). The modifications were primarily aimed at reducing the confusion that testees can experience when performing the original version. Thus, in the MCST, cards sharing more than one attribute with a stimulus card are removed and testees are informed when the rule has changed.

Existing reviews (Mountain and Snow 1993; Reitan and Wolfson 1994; Parker and Crawford 1992) of the sensitivity and specificity of the WCST have concluded that the test has limited utility. For example, Mountain and Snow (1993) stated that

> The evidence that frontal patients perform more poorly than nonfrontal patients is weak. There is insubstantial evidence to conclude that the WCST is a measure of dorsolateral-frontal dysfunction. The clinical utility of the test as a measure of frontal-lobe dysfunction is not supported.
>
> (p. 108)

Studies published subsequent to these reviews have yielded results that reinforce their conclusions.

A particularly important study was conducted by Axelrod *et al.* (1996) using the WCST standardization data (356 healthy controls and 343 patients, including samples of patients with focal lesions). Axelrod *et al.* reported that the WCST achieved a modest degree of overall discrimination between patients and controls but did not discriminate between the different patient samples, i.e. the performance of frontal cases was not appreciably poorer than anterior cases; the effect size (*r*) for this latter comparison, calculated by the present authors, was 0.24. There has been much less evaluation of the MCST. In Nelson's (1976) original study a cut-off of 50 per cent perseverative errors had high specificity (but low sensitivity) for frontal lobe lesions. However, subsequent studies have reported patterns of results that mirror those found for the WCST. For example, van den Broek *et al.* (1993) found that although, overall, neurological patients could be differentiated fairly successfully from controls, the performance of anterior and posterior cases was indistinguishable.

Thus it would appear as if the WCST and its variants, although moderately impairment sensitive, have poor specificity for the presence of anterior lesions. Set against these disappointing results, there is some evidence that these tests have moderate ecological validity. Burgess *et al.* (1998) found that the MCST correlated significantly (0.37) with ratings made by the relatives of neurological patients ($N = 92$) on the Dysexecutive Questionnaire (DEX) (Wilson *et al.* 1996). There is also evidence that the WCST is a moderate predictor of the level of functional independence achieved following discharge from acute rehabilitation (Hanks *et al.* 1999); for example, the correlations between the WCST and measures of subsequent community integration and level of disability were −0.32 and −0.42 respectively.

## Another old warhorse: phonemic fluency

Phonemic fluency requires the generation of words by initial letters under time constraints (normally 60 seconds per each of three letters). This test is also known as the Controlled Oral Word Association Test (COWAT) (Benton and Hamsher 1976), the FAS test (because these are the three letters commonly used), or is simply referred to as verbal fluency. Large scale normative data are available, including extensive data for the elderly (Ivnik *et al.* 1996), the internal consistency and parallel form reliability of these tests are very high, as is their inter-rater and test-retest reliability (e.g. see Spreen and Strauss 1998). Furthermore, the test is quick and easy to administer and score.

Perret (1974) suggested that phonemic fluency was sensitive to executive dysfunction (and more sensitive than semantic fluency) because normally we retrieve words based on their meaning; the requirement to retrieve by initial letter is non-routine and also requires suppression of words that are semantically related to previously produced words. Evidence from dual-task studies in healthy participants indicates that phonemic fluency imposes significant executive demands. Martin *et al.* (1994), reported a cross-over interaction when studying the effects of secondary tasks designed to activate either temporal structures (a semantic decision task) or frontal structures (a motor sequencing task) on phonemic and semantic fluency. Phonemic fluency was severely disrupted by the sequencing task but much less so by the semantic decision task; the converse pattern was observed for semantic fluency.

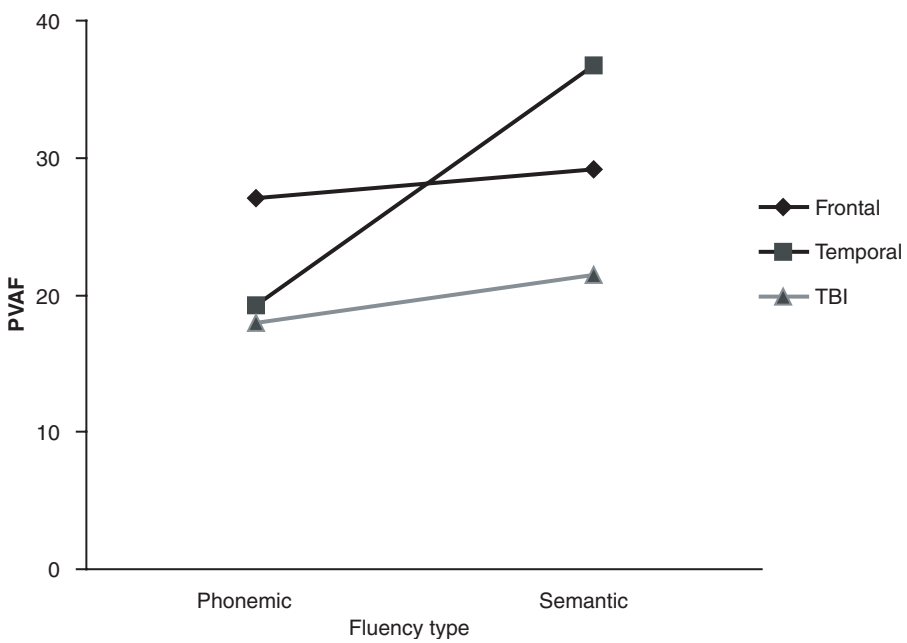
The evidence from focal lesion studies has provided further support for the position that phonemic fluency imposes significant executive demands, but the literature is full of contradictions (Reitan and Wolfson 1994). As sample sizes in these individual studies are often modest, many of these contradictions may simply reflect sampling error. In an attempt to clarify the literature, Henry and Crawford (2004) conducted a meta-analysis of focal lesion studies. When focal frontal lesion samples were compared to controls, a large mean effect size ($r = 0.52$) was obtained for phonemic fluency (the effect size for samples consisting exclusively of cases with *left* frontal lesions was even larger) but the effect

size for semantic fluency was of an equivalent magnitude. In posterior lesion cases the effect size for phonemic fluency was smaller than that observed for frontal samples but the semantic fluency effect size was substantially larger (i.e. there was a cross-over interaction between fluency type and lesion location). Therefore, if one takes the presence of a focal frontal lesion as an (imperfect) proxy for the presence of executive dysfunction, these results suggests that phonemic and semantic fluency impose comparable executive demands. However, semantic fluency is more sensitive to a compromised semantic system.

Henry and Crawford (2004) also obtained effect sizes for other cognitive measures in order to provide context for the fluency results. For focal frontal lesions, the effect sizes for psychomotor speed (Trails A), and for IQ were modest indicating that (a) impaired fluency performance was not simply a reflection of a general slowing in psychomotor speed, and (b) was disproportionate to the general level of cognitive impairment in the frontal samples. The effect size for the WCST was also markedly smaller than that obtained for phonemic fluency, indicating that fluency is the more sensitive of the two measures. Importantly, the above evidence for a differential deficit on phonemic fluency was not apparent in posterior cases; i.e., the effect sizes for these other measures and phonemic fluency were broadly comparable. These results reinforce the view that, like all tests of executive dysfunction, phonemic fluency tests must be interpreted in the context of a client's overall pattern of current performance (and in the context of their premorbid level of ability; see the later section on p. 000).

AU: Please update.

A meta-analysis of verbal fluency following traumatic brain injury revealed that the pattern of performance across phonemic and semantic fluency and the other cognitive measures referred to above was remarkably similar to that found in focal frontal cases. Figure 19.1 presents the effect sizes




AU: Henry and Crawford in press – are you able to update this?

**Figure 19.1** Effect sizes for phonemic and semantic fluency following focal frontal or temporal leisons or TBI; data are drawn from meta-analytic studies by Henry and Crawford (2004) and Henry and Crawaford (in press) and are presented as the percentage of variance accounted for (PVAF) by group membership, i.e. clinical versus control.

on phonemic and semantic fluency for TBI and focal frontal and temporal lesions; the effect sizes are presented as the percentage of variance in the fluency measures accounted for by group membership (i.e., clinical group versus controls). The extensive primary literature on verbal fluency tasks in other neurological and psychiatric disorders has also recently been subjected to meta-analysis, with results that are both of theoretical and practical interest (Henry *et al.* 2004, in press, Henry and Crawford in press a, b, c, d).

AU: Henry et al in press – are you able to update this?

AU: Henry and Crawford in press a, b, c and d – as above

There is some evidence that phonemic fluency has ecological validity as a measure of executive functioning. In the study by Burgess *et al.* (1998) referred to earlier, phonemic fluency was significantly correlated with ($r = 0.35$) with ratings of everyday executive problems as measured by the DEX. Similarly, in Hanks *et al.*'s (1999) study, phonemic fluency was a strong predictor of the level of functional independence achieved following discharge from acute rehabilitation; e.g. the correlation with the Disability Rating Scale was –0.52.

## Alternatives to conventional fluency

There are a number of alternatives to conventional phonemic fluency tests. These alternatives hold promise because, in principle at least, they impose greater demands on the executive system. Alternating fluency requires switching between retrieval by phonemic and semantic probes. Downes *et al.* (1993) have shown that in Parkinson's disease, a disorder associated with executive deficits, alternating fluency was differentially impaired relative to conventional, non-alternating fluency.

Warrington (2000) developed a verbal fluency task she entitled the Homophone Meaning Generation Test (HMGT). The task involves providing multiple definitions for a series of homophones (e.g., bear/bare). The task was standardized on 170 normals (Warrington 2000), has good reliability (alpha = 0.82), and yields scores that are normally distributed (Crawford and Warrington 2002). Warrington has argued that the HMGT imposes greater executive demands than conventional fluency tasks and existing set-shifting tasks (such as the WCST) because it requires constant switching between concepts. Anterior cases were significantly impaired on this task (effect size *r*, calculated by the present authors, was 0.44 when compared to healthy controls); posterior cases did not differ significantly from controls and, encouragingly, the corresponding effect size was very small ($r = 0.07$). An unusual and useful feature of this fluency task is that it is untimed; as a result, and unlike conventional verbal fluency tasks, the clinician can be confident that impaired performance is not down to a general decline in speed of processing.

Crawford *et al.* (1995) developed an excluded letter fluency task (ELF) that requires generation of words that do not contain a specified vowel (e.g., testees have to generate words that do not contain the letter *e*). In keeping with Perret's (1974) argument reviewed above, the test was designed to involve non-routine retrieval (no matter how the lexicon is organised, it is not organised by the absence of letters) and, in addition, to impose greater demands on self-monitoring. Crawford *et al.* (1995) found that head-injured participants committed many more errors on this task than controls; in contrast the error rates on conventional fluency tasks were equivalent across the groups. In comparison to phonemic fluency, ELF is less dependent on crystallised knowledge and is more sensitive to ageing and slowed information processing. Normative data from 399 healthy participants have been provided by Shores *et al.* (submitted); these data include normative data on change on retesting.

AU: Shores et al submitted –are you able to update this?

To the present author's knowledge, Zangwill (1966) was the first to identify the potential of ideational fluency tasks as a means of capturing executive problems. The Uses for Common Objects task (UCO) (Getzels and Jackson 1961) or 'alternate uses' task requires the generation of unusual uses for everyday objects (e.g. a brick). This task therefore attempts to capture the problems in self-initiation and diminished creativity that characterise many patients with executive dysfunction; in clinical

practice many patients are encountered who find it hard to move beyond the well-consolidated, conventional, uses for the stimulus objects.

Butler *et al.* (1993) compared frontal cases and controls on UCO and on phonemic fluency and reported that UCO was more sensitive; the UCO effect size (*r*), calculated by the present authors, was 0.53, versus 0.39 for phonemic fluency. Crawford *et al.* (1995) found that this task yielded the largest difference between head-injured patients and controls from among a battery that contained conventional fluency measures (i.e., phonemic and semantic fluency). Eslinger and Grattan (1993) reported that UCO performance was severely impaired in patients with frontal lesions (effect size *r*, computed by the present authors, was 0.76) whereas a posterior sample did not differ significantly from controls (effect size = 0.22).

Finally, drawing on evidence that naming verbs is disproportionately disrupted by frontal lesions, Piatt *et al.* (1999) suggest that action naming fluency is a potentially useful measure and have provided evidence consistent with the task imposing executive demands.

## The importance of premorbid ability when assessing executive deficits

The vast majority of tests of executive function have moderate to high correlations with general intellectual ability in the healthy population (Obonsawin *et al.* 2002), where exceptions to this rule occur it usually reflects poor measurement properties of the tasks (such as ceiling effects). This has important theoretical implications, e.g. Duncan *et al.* (1995) have argued that executive functions and fluid intelligence are essentially synonymous, but it also has very practical implications for assessment. A client's performance on executive tasks must be interpreted in the context of their general level of *premorbid* ability; an average score on a particular executive task can represent a significant decline from a previously higher level in a patient of above average premorbid ability.

Verbal fluency tests provide a clear example of the need to consider this factor as they are highly correlated with verbal IQ in the general healthy population (Crawford *et al.* 1993). Furthermore, Borkowski *et al.* (1967) found that brain-damaged patients with high Verbal IQs outperformed healthy controls of below average IQ on verbal fluency tests.

A number of formal methods of estimating premorbid ability are available, of which the most common is the National Adult Reading Test (NART) (Nelson and Willison 1991). The NART is an oral reading test consisting of 50 words that violate grapheme-phoneme correspondence rules (e.g., *chord*). NART performance correlates highly with IQ (Crawford *et al.* 2001), and is robust in the face of many neurological and psychiatric disorders (Crawford 2004; O'Carroll 1995). The available evidence suggests that NART performance is relatively unaffected by focal frontal lesions (Bright *et al.* 2002; Crawford and Warrington 2002) or closed head injury (Watt and O'Carroll 1999).

The NART can be used to provide a general estimate of premorbid IQ. However, NART-based regression equations have also been developed specifically to provide comparison standard for a patient's performance on executive tasks. For example, Crawford *et al.* (1992) built a regression equation which can be used to estimate premorbid performance on phonemic fluency and a similar equation is available for the homophone fluency test (HMGT) referred to earlier (Crawford and Warrington 2002). In clinical practice the NART estimated premorbid fluency score is compared to the actual fluency score obtained on testing; a large (and statistically significant) discrepancy in favour of the former is taken as an indication of acquired impairment. Support for the utility of this approach is provided by results from a hierarchical discriminant function analysis in which inclusion of the NART as an index of premorbid ability significantly improved the ability of the HMGT to differentiate between frontal cases and healthy controls; the effect size (*r*) when the NART was included was 0.53 versus 0.44 for the HMGT alone.

## The Cognitive Estimation Task

The Cognitive Estimation Task (CET) (Shallice and Evans 1978) requires selection of an appropriate plan for generating an approximation and monitoring of the result prior to production. In its original form it consists of 15 questions (e.g., 'How fast do race horses gallop?') that either do not have precise answers or, if they do, would go beyond the knowledge of most individuals. A revised 10-item version, for which norms are available, has also been employed (e.g. see O'Carroll, 1994 #2562). Shallice and Evans (1978) reported that a sample of patients with anterior lesions performed significantly more poorly on the CET (an effect size for this difference was not reported and insufficient information is provided to permit its calculation from other statistics).

AU: O'Carroll 1994 – should this say O'Carroll et al? Also, what does #2562 denote

Subsequent evaluation of the measurement properties and validity of this test have been generally very disappointing. O'Caroll *et al.* (1994) reported that the reliability of the test was low (Cronbach's alpha = 0.40) in a sample of 150 healthy controls. Furthermore, although the CET yields a global score, O'Carroll *et al.* extracted five factors from 10 items; i.e., the scale is factorially impure. The available evidence suggests that the test's sensitivity to the presence of executive problems is also poor. Taylor and O'Carroll (1995) found that a sample of patients with anterior lesions did not differ significantly from a posterior sample. Moreover, these authors reported that the performance of a wide variety of neurological samples (including the anterior sample) did not differ significantly from controls; the one exception was a sample of patients with Korsakoff's syndrome. Crawford *et al.* (2000b) found that CET performance was significantly impaired in Huntington's disease (a condition associated with executive problems), but significantly less so than WAIS-R IQ (a measure relatively insensitive to executive problems). The CET was one of the few putative measures of executive function that failed to correlate significantly with rated everyday executive problems in the study by Burgess *et al.* (1998) referred to earlier. In conclusion, although the CET may occasionally yield clinically useful qualitative information, it cannot be recommended.

## Brixton Spatial Anticipation Test

Conceptually the Brixton Spatial Anticipation Test (Burgess and Shallice 1997) has some similarities with the WCST. It requires testees to discover the rules underlying the placement (apparent movement) of a blue circle among a grid of unfilled circles; after a given pattern is established the rule changes.It has been argued (Burgess and Shallice 1997) that the Brixton has a number of practical advantages over other measures of set-shifting; namely that it is less time-consuming and less stressful for patients, and yields scores that are normally distributed in the general population. The normative sample is relatively modest ($N = 121$) and the reliability (internal consistency) of the test is only moderate (0.62). However, the validation data are impressive. Burgess and Shallice (1997) reported a highly significant difference between a sample of cases with frontal lesions and healthy controls; the effect size for this comparison, calculated by the present authors, was large ($r = 0.50$). Furthermore, frontal cases were significantly more impaired on the Brixton than cases with posterior lesions; the effect size for this comparison was moderate ($r = 0.34$). Finally, posterior cases did not differ significantly from healthy controls and the corresponding effect size was small ($r = 0.16$); i.e. posterior cases had relatively little difficulty with the task. This provisional evidence of the sensitivity and specificity of the Brixton for anterior lesions stands in contrast to the results obtained for other set-shifting tests; i.e., see Axelrod *et al.*'s (1996) study of the WCST referred to earlier.

One limitation of the Brixton is the use of Sten scores as a metric to express performance. Although Sten scores have the advantage of simplicity, they are coarse-grained (the difference between Sten scores correspond to 0.5 of an *SD*) and thus potentially meaningful differences between raw

scores are obscured. However, normative data in the form of *T* scores, based on an enlarged normative sample ($N = 222$), have recently been developed (Crawford *et al.* in preparation); this study also includes a method of testing for changes in Brixton scores on retesting.

AU: Crawford et al in preparation – are you able to update this?

## Hayling Sentence Completion Test

The Hayling Sentence Completion Test (Burgess and Shallice 1997) is primarily aimed at detecting difficulties in suppressing pre-potent responses and consists of two parts. In the first, the subject has to complete sentences with the pre-potent response e.g. providing the word 'ship' when presented with the sentence 'The Captain went down with the sinking …'. In the second part the subject has to suppress the pre-potent response and complete the sentences with an unrelated word. The test yields four indices, all of which are expressed as Sten scores derived from the same healthy sample used to norm the Brixton ($N = 121$). The indices are: completion latency for the pre-potent responses (Hayling 1), latency of completion in the suppression condition (Hayling 2), number of errors in the suppression condition (Hayling 2 errors), and an overall score. The reliabilities of the test are generally very high in impaired groups (0.72 to 0.93) and the test has moderate to high temporal stability (0.62 to 0.76) in normals (Burgess and Shallice 1997).

The validity of the Hayling has been assessed by comparing the performance of frontal lesion cases with controls and cases with posterior lesions. Expressing these group differences as effect sizes (Table 19.2) is revealing: when comparing anterior cases against controls there is only a very modest difference between the effect size on Hayling 1 (0.39) and Hayling 2 (0.41). In other words latencies in providing the *pre-potent* responses were just as effective in differentiating healthy and anterior cases; this suggests that an overall slowing of response, rather than a problem with inhibition, may lie underlie the effect in the anterior cases. However, the suppression condition comes in to its own when the frontal and posterior cases are compared. The effect size for errors under suppression (0.37) is markedly larger than the effect size for basic initiation (0.23); therefore the anterior cases had disproportionately greater difficulties than posterior cases in inhibiting the pre-potent response. It can also be seen that, on all indices, the effect sizes for posterior cases versus controls are small, thereby providing encouraging evidence of the Hayling's specificity for anterior lesions.

## Behavioural Assessment of the Dysexecutive Syndrome (BADS)

Many existing formal neuropsychological tests fail to detect important core components of executive dysfunction, such as problems in initiation and self-organisation, because they are highly structured. As Shallice and Burgess (1991) note:

> The patient typically has a single explicit problem to tackle at any one time … the trials tend to be very short … task initiation is strongly prompted by the examiner and what constitutes successful trial completion is clearly characterised.
>
> (pp. 727–728)

**Table 19.2** Effect sizes[a] (anterior cases versus controls and versus posteriors) for the Hayling Test

| Hayling Index | Anteriors vs. controls | Anteriors vs. posteriors | Posteriors vs. controls |
|---|---|---|---|
| Hayling 1 (Time 1) | 0.39 | 0.23 | 0.17 |
| Hayling 2 (Time 2) | 0.41 | 0.32 | 0.04 |
| Hayling 2 errors | 0.42 | 0.37 | 0.03 |
| Overall score | 0.48 | 0.41 | 0.06 |

[a]Effect sizes calculated by present authors using data presented in Burgess and Shallice (1997).

The BADS test battery was developed in an attempt to address some of these limitations. It consists of six subtests: the Zoo Map test (a planning task); the Modified Six Elements Test, which is a simplified version of the Six Elements Test developed by Shallice and Burgess (1991) and taps planning/self-directed organisation; the Temporal Judgement Test, which is akin to the CET reviewed above, and requires the application of intelligent guesswork and error checking; the Rule Shift test, which measures set shifting and the ability to inhibit previously established responses; the Action Program test, a novel practical problem solving task; and the Key Search test, which taps self-directed organisation.

The norms for the BADS are derived from a sample of 216 healthy individuals aged between 16 and 87 years. Scores on the individual subtests are categorised on a four-point scale (0 to 4); these are summed and converted to yield an overall Profile score which is expressed on an IQ metric (mean 100, SD 15). With regard to the measurement properties of the test, the inter-rater reliability of the subtests are uniformly excellent, with coefficients ranging from 0.90 to 1.00 (the majority being 0.98 or above). The test-retest reliabilities of most subtests are moderate in magnitude (0.64, 0.67 and 0.71) but the remainder are low ranging from –0.08 to 0.39. Although these latter coefficients look alarming, two factors must be borne in mind. First, these coefficients were obtained from healthy participants, and given that they would exhibit ceiling effects (and hence limited variability), this would attenuate the coefficients; the coefficients would be substantially higher in an impaired sample. Second, difficulties in coping with novelty is a central feature of patients with executive dysfunction and the BADS seeks to capture this. Thus, the task demands on a second testing are very different from those on first exposure. As a result, and as is the case for many other putative tests of executive dysfunction, it is not realistic to expect the BADS to exhibit high test-retest reliability (Crawford 2004).

Internal consistency data for the BADS are not reported in the test manual but Cronbach's alpha for the Profile score can be calculated from other information that is provided (i.e., the means and SDs of the raw Profile score and the means and SDs of the subtest scores contributing to the Profile score). Alpha in the patient sample recruited for validation purposes was moderate (0.70).

The BADS manual reports the results of significance tests comparing controls to the patient sample but it is perhaps more informative to express the differences between these samples as effect sizes. These were calculated by the present authors and are presented in Table 19.3. It can be seen that, with one noteable exception, the effect sizes for the individual subtests are small to moderate in magnitude. However, it must be stressed that, unlike most of the other effect sizes reported in this chapter, these effect sizes are derived from comparing healthy controls to a *general* sample of neurological patients rather than a sample of patients with frontal lesions. Therefore, many of the patients in this sample would not be expected to exhibit significant executive problems and it follows that large effect sizes would also not be expected.

**Table 19.3** Effect sizes (*r* for patients versus controls) for BADS subtests and Profile score and correlations with DEX ratings of everyday executive problems (equivalent results for WAIS-R IQ also provided for context)

| BADS test | Effect size (*r*)[a] | *r* with DEX[b] |
|---|---|---|
| Action program | 0.25 | –0.37 |
| Key search | 0.12 | –0.31 |
| Six elements | 0.53 | –0.40 |
| Rule shift cards | 0.21 | –0.45 |
| Zoo map | 0.15 | –0.46 |
| Temporal judgement | 0.24 | –0.40 |
| Profile score | 0.38 | –0.62 |

[a]Effect sizes calculated by present authors using data presented in Wilson *et al*. (1996).
[b]Data from Wilson *et al*. (1996).

The effect size for the overall Profile score is larger than all but one subtest and is moderate- to large in magnitude (0.38). It is to be expected that a composite measure would have a larger effect size than its components. In this particular example, the results are consistent with a fractionation of the executive system (see Burgess and Simons, Chapter 18 this volume). In other words, the small effect sizes for individual subtests may arise because only a proportion of cases exhibited deficits on any one subtest and these were often not the same cases who exhibited deficits on the other subtests. Having said that, it is also clear that the source of much of the effect for the Profile score stems form the inclusion of the Modified Six Elements Test; this subtest has by far the largest effect size (0.53). The practical implications of these results are that the full BADS should be administered when feasible and that performance on the Six Elements test should be weighted highly when arriving at a formulation.

The major strength of the BADS lies with its demonstrated relationship to everyday executive problems. In the BADS patient validation sample all subtests correlated significantly with executive problems rated by relatives (these correlations are reproduced in the second column of Table 19.3). It can also be seen that the overall Profile score has a very strong (negative) correlation (−0.62) with everyday problems; this correlation is substantially larger than that obtained for a measure of general ability (WAIS-R IQ) in the same sample (−0.42) thereby providing evidence of specificity.

## Dual tasks in the assessment of executive dysfunction

All major theoretical models of the executive system stress its role in the coordination of activity (see Burgess and Simon, Chapter 18, this volume). Baddeley and colleagues (e.g. Baddeley *et al.* 1997) have lain particularly strong emphasis on this aspect of the executive system and have explored the potential of using dual tasks to capture executive deficits. In Baddeley *et al.* (1997) patients with frontal lesions were assigned to one of two groups on the basis of whether they exhibited a dysexecutive syndrome (as assessed independently by two clinicians from a review of the medical notes). Dual task performance was assessed by combining digit span with a concurrent paper-and-pencil tracking task; the dependent variable was an index that compared single-task performance on these two tasks with performance on the tasks under the dual-task conditions.

On traditional clinical tests of frontal function (phonemic fluency and the WCST) the majority of both frontal groups were in the impaired range. However, the dysexecutive group was not significantly more impaired on these tasks than the non-executive group. In contrast, they did exhibit a significantly larger dual-task decrement than the non-executive group and the effect size ($r$) for this difference (calculated by the present authors from the reported $t$ value for this comparison) was substantial (0.58). This effect size is particularly impressive as it is based on a comparison of groups both of which had frontal lesions as opposed to the other effect sizes reported (which are based on comparing frontal groups with either controls or posterior lesion samples). Further important evidence of the ecological validity of dual task decrements and their relevance to rehabilitation planning has been provided by Alderman's (1996) study of severely head-injured patients. A large dual-task decrement was associated with a poor response to behavioural intervention.

A potential problem with the use of dual tasks in individual assessment is that the key variable is a difference score (i.e. the dual task decrement). Difference scores have lower reliability than the components from which they are derived, particularly when, as is liable to be the case in the present context, the two components are highly correlated (Crawford 1996). However, the results reported above demonstrate that the effects are sufficiently large to overcome attenuation due to measurement error; furthermore, given the unreliability of differences, it can be concluded that the 'true' effect is considerably larger even than that obtained. In conclusion, dual tasks have great potential to capture what is a core executive process but their routine use in clinical practice awaits development of fully standardised tests and accompanying large-scale normative data.

## Disability rating scales

Disability rating scales play a crucial role in quantifying the impact of cognitive deficits on everyday functioning and, particularly in the case of executive deficits, in identifying difficulties not captured by formal ability tests. As ratings can be carried out by patients and their relatives, they are also useful in identifying and quantifying diminished insight.

A promising rating instrument for assessing executive dysfunction is the Dysexecutive Questionnaire (DEX) (Wilson *et al.* 1996) which comes bundled with the BADS test reviewed earlier. The questionnaire consists of 20 items and comes in self-rating and proxy rating versions. The underlying structure of the instrument remains to be clarified; a principal components analysis reported in the test manual obtained a three factor solution consisting of factors labeled as *Behaviour*, *Cognition* and *Emotion*. Burgess *et al.* (1998) reported a five factor solution (*Inhibition*, *Intentionality*, *Executive Memory*, *Positive Affect*, and *Negative Affect*) in a neurological sample and provided some evidence that these factors related differently to formal tests of executive ability. Chan (2001) also obtained a five factor solution in a healthy sample: these factors had many similarities, but were by no means identical, to those obtained by Burgess *et al.*

The authors of the DEX view it primarily as a qualitative instrument, but it clearly has the potential to also yield quantitative information; currently the only healthy normative data consist of the mean and SD of members of the BADS standardisation sample, however, a patient's score can be compared against percentiles from the patient sample. The strength of the DEX stems from the previously reviewed evidence that it correlates with formal measures of executive functioning; i.e. the presence of these sizeable correlations simultaneously provides evidence of convergent validity for both the formal tests and the DEX. Further evidence of convergent validity can be found in Chan's (2001) study of the DEX and other executive tasks.

A recently developed rating instrument that has many impressive and useful features is the Frontal Systems Behavior Rating Scale (FrSBe) (Grace and Malloy 2001). It consists of 46 items that yield a total score and score on three subscales *Apathy*, *Disinhibition*, and *Executive Dysfunction*. There are self-rating and proxy (i.e. family member) rating versions and it is also available in Before (i.e. pre-injury) and After (post-injury) formats. The normative sample is impressive consisting of 436 persons with an age range of 18 to 95; ratings on the scales are converted to *T* scores and are stratified by gender, age and education.

The reliabilities of this instrument are generally high (Cronbach's alpha ranged from 0.72 to 0.95). The available validation data are also positive. A factor analysis of the FrSBe and found support for the allocation of items to the three subscales (Grace and Malloy 2001). Grace *et al.* (1999) compared scores on the proxy rated versions for samples of healthy controls and patients with either frontal or non-frontal lesions. There were highly significant differences between the Before and After ratings in the frontal sample. In addition, the frontal cases were scored significantly higher than both normal controls and non-frontal cases. In summary, this instrument has the capacity to yield much clinically useful information and has good normative data and sound measurement properties.

Finally, rating scales that assess other aspects of cognition and behaviour can also be useful in assessing patients with executive problems. For example, and as noted, executive dysfunction can produce memory difficulties, particularly when the everyday memory task imposes heavy strategic/organisational demands or involves prospective memory. A number of memory self and proxy rating scales are available including the Everyday Memory Questionnaire (Sunderland *et al.* 1988), and the Cognitive Failures Questionnaire (Broadbent *et al.* 1982); the coverage of the latter instrument falls midway between that of memory rating scales and dysexecutive rating scales.

A recently developed questionnaire for assessing memory problems that, because of its systematic coverage of retrospective and prospective memory, may have particular relevance to assessing

patients with executive problems is the Prospective and Retrospective Memory Questionnaire (PRMQ) (Smith *et al.* 2000). This scale, which consists of 16 items, has high reliability (Cronbach's alpha = 0.89), comes in self- and proxy-rating versions, has normative data (expressed as *T* scores) from a sample of 555 healthy controls aged from 17 to 94, and has a latent structure that is consistent with the allocation of items to the Prospective and Retrospective subscales (Crawford *et al.* 2003).

## Conclusion

A comprehensive assessment of executive dysfunction is fundamental in planning any neurorehabilitation attempt. This area of assessment is very challenging; only recently have formal tests become available that combine adequate psychometric properties, ecological validity and a sound theoretical basis. Although the emphasis in this chapter has been on formal ability tests and disability rating scales, clinical skills and experience are crucial in and in integrating these diverse sources of information to achieve a formulation of a client's difficulties and to draw out their implications.