Neurol Sci (2000) 21:279-291

ORIGINAL

M. Laiacona • M.G. Inzaghi • A. De Tanti • E. Capitani

Wisconsin card sorting test: a new global score, with Italian norms, and its relationship with the Weigl sorting test

Received: 15 June 2000 / Accepted in revised form: 10 November 2000

Abstract The Wisconsin card sorting test and the Weigl test are two neuropsychological tools widely used in clinical practice to assess frontal lobe functions. In this study we present norms useful for Italian subjects aged from 15 to 85 years, with 5-17 years of education. Concerning the Wisconsin card sorting test, a new measure of global efficiency (global score) is proposed as well as norms for some well known qualitative aspects of the performance, i.e. perseverative responses, failure to maintain the set and non-perseverative errors. In setting normative values, we followed a statistical methodology (equivalent scores) employed in Italy for other neuropsychological tests, in order to favour the possibility of comparison among these tests. A correlation study between the global score of the Wisconsin card sorting test and the score on the Weigl test was carried out and it emerges that some cognitive aspects are not overlapping in these two measures.

Key words Wisconsin card sorting test • Weigl test • Italian norms

M. Laiacona (✉)
S. Maugeri Foundation, IRCCS
Rehabilitation Institute of Veruno
Via per Revislate 13, I-28010 Veruno (NO), Italy

M.G. Inzaghi • A. De Tanti
Rehabilitation Centre Villa Beretta
Costamasnaga (LC), Italy

E. Capitani
Clinic for Nervous Diseases
Milan University
San Paolo Hospital, Milan, Italy

Introduction

The Wisconsin card sorting test (WCST) [1, 2] is designed to assess the ability of reasoning and of shifting cognitive strategies and is based on previous studies on abstract thinking by Weigl [3]. The WCST is generally considered a useful tool for detecting frontal lobe dysfunction [4, 5] and for assessing the integrity of executive functions. In fact, it requires (i) a preliminary abstraction process, necessary to identify latent regularities within recurring experiences, (ii) the development of an appropriate problem solving strategy, and its maintenance across changing stimulus conditions, (iii) the ability to change a current rule when negative rewarding occurs, (iv) progressive learning based on self-instruction, (v) the capacity to memorise previously tested rules, and finally (vi) the avoidance of certain possible solutions because of their logical implausibility.

During the test, a subject is requested to match a series of cards according to their similarity with four stimulus cards. The subject must infer the similarity criterion from the examiner's feedback by making hypotheses and verifying their accuracy, taking into account also right or wrong answers. In addition the subject has to maintain this sorting principle ignoring other stimulus dimensions.

At a certain undisclosed point, and generally without warning, the examiner changes the criterion, and the subject has to decide if to follow or to modify the preceding criterion. The more efficient the procedure adopted, the fewer the number of response cards 'wasted' in trials: the more precise the hypotheses, the more adequate the strategy.

In clinical practice, different administration procedures have been suggested, some of which are summarised as follows:

Number of cards

- Two sets of 60 cards [1]
- One set of 48 cards [6]
- One set of 64 cards [7]

Order of cards within decks
- Random criterion [1]
- Experimental criterion [6]
- Fixed arrangement (no consecutive same-colour, same-shape, same-number cards allowed) [8]

Ambiguous cards
- Generally included
- Removed [9]

Shifting criterion
- After five consecutive correct responses [1]
- After six correct responses [9]
- After 10 cards [7]
- After a different number of correct responses, i.e. a varied number of consecutive correct sorts [2]

Warning
- Yes, Nelson [9]

Sorting categories and their sequence
- Only two fixed categories: colour, number [6]
- Six fixed sorting categories: colour, number, form, number, colour, form, [2]
- Colour, number and form sorting criteria in 24 different sequences [10]
- 9 sorting categories in 96 different sequences [1]
- Six 'cycles' each having three categories in standard order: colour, form, number [11]

Discontinuation
- When two consecutive categories are "achieved" or missed [6]
- When the subject achieves the six categories or when more than 64 cards have been administered for a single sorting category [12]
- When six categories are achieved, but dropping out subjects to whom all 256 cards have been administered [13]
- After sorting 64 cards [7]
- When all 120 cards are sorted or when the 9 categories are achieved [1]
- When 18 categories are achieved or when only three categories are achieved administering at most 15 cards for each [11]

The different administration procedures make a comparison between different studies almost impossible. After the publication of a standard set of WCST materials, instructions and normative values derived from North American subjects [14], the use of the WCST has become more appealing. Nevertheless, these normative values cannot be applied when different administration procedures are adopted (e.g. the short versions with 64 cards [15, 16]).

In the present study we adopted the material and the administration procedure (including the arrangement of the response cards and the configuration of the response and stimulus cards) of Robinson, Heaton, Lehmam and Stilson [17] and Heaton [14]. We did not adopt the short version as it underestimates the perseverations of responses based on previous criterion [18].

The WCST is widely used in clinical practice with psychiatric patients, head injured patients, with subjects suffering from cerebrovascular diseases and with patients suffering from dysexecutive syndrome [19]. Focal frontal lobe damage is not always present [20-22]. However, PET studies have documented selective activation of regional cerebral blood flow in the left dorso-lateral prefrontal cortex [23] and more generally, Heaton et al. [24] pointed out that executive functions impairment can also derive from differently located cortical lesions if the latter are detrimental to the efficiency of complex neural networks that include the frontal lobe. Other authors have underlined the relevance of some extra frontal brain regions for the efficiency of control functions, with special reference to the hippocampus [25, 26], but see also [27] for a critical comment.

Patients with dysexecutive syndrome, can present the following types of pathological behaviour:
- Perseverations of the preceding criterion, due to inability to modify a strategy according to the examiner's reply
- Failure to maintain the set, due to inability to maintain the correct sorting principle
- Responses that do not match any sorting criteria, as the patients perform misleading associations
- Impulsiveness

These features do not always cluster together and their severity may vary from one patient to another.

The need for Italian norms is strongly suggested by the demographic and generational differences that are likely to occur with the North American reference sample used by Heaton [14]. Moreover, we felt it useful to adopt a methodology [28] used for the standardisation of most neuropsychological Italian tools that would allow a direct and reliable comparison of the performances given on different tests.

Assessment of the abstract reasoning ability of brain-damaged patients has been carried out by a number of authors also with alternative procedures, many of which are based on Weigl's original work [3]. A version of this task has often been employed with patients affected by left hemisphere focal damage and aphasia: the patient is given 12 objects that differ according to 5 criteria (shape, size, colour, thickness and the suit displayed on their surface); the request is to discover the sorting criteria and, in the case of failure, to reproduce the sorting carried out by the examiner. The Weigl test has the advantage of being of short administration, relatively free from attentional load and better suited for severely impaired patients. However, to our knowledge, the Weigl test and WCST have not been cross-validated, and we do not know to what extent they are tapping the same psychological abilities. As abstract reasoning ability has often been assessed in the past not only by the WCST but also by means of the Weigl test, we decided to collect the performances on this latter test with the same control sample used for WCST in order to measure their correlation. We further decided to recalculate norms also for the Weigl test. The Weigl test is still widely used [29], but normative values appropriate for subjects under age 40 years are not available. This is a hindrance, as many patients presenting a possible dysexecutive syndrome (e.g. head-injured patients) are under age 40 years.

Materials and methods

The investigation was carried out on 205 subjects, including 100 men and 105 women (Table 1). Their ages ranged from 15 to 85 years (mean, 46.5 years) and education from 5 to 17 years (mean, 11.4 years). Among the subjects, 19 were patients admitted to the Valduce Hospital (Costa Masnaga) for other than neurological or psychiatric illnesses (e.g. orthopaedic problems), and 186 were randomly recruited from the healthy population of the same area (e.g. relatives of patients attending the hospital). The average hospital stay of the 19 patients was 20 days (SD = 1.4; range 18-21). None of them presented overt psychiatric or psychological disturbances. As far as could be ascertained, all were free from other conditions potentially detrimental to cognitive performance, such as alcohol abuse or the use of drugs known to affect the central nervous system (CNS). We were interested in the study of the normal population, defined as subjects neither selected for nor distinguished by any medical or metabolic pathological condition [26]. The inclusion criteria were not too selective, in order to avoid the sampling of a "hyper-normal" group. Probably, we included a few subjects affected by mild hypertension and diabetes with a satisfactory drug treatment. Table 1 shows the composition of the study group according to demographic variables.

The Wisconsin card sorting test

The standard WCST material was used [14]; it includes four stimulus cards and two identical decks of 64 fixed-sequence response cards. Each card presents a certain number of figures of the same form and colour. Four stimulus cards were placed in front of each subject displaying, respectively: one red triangle, two green stars, three yellow crosses, and four blue circles (from left to right). Each subject was given the two decks of 64 response cards and was instructed to match each consecutive card with one of the stimulus cards. Three possible sorting categories were assumed: form (crosses, circles, triangles or stars), colour (red, blue, yellow or green), and number (one, two, three or four). The subjects were notified only whether the responses were right or wrong, without mention of the underlying sorting criterion (known only to the examiner, in sequence: colour, form, number, colour, form, number). Once a subject had made 10 consecutive correct matches, the sorting criterion was changed, without warning. The test was discontinued when a subject achieved all six sorting categories or when all 128 cards were sorted. We examined the following quantitative measures.

Global score. This represents an overall index of the WCST performance, though has not previously been used as such to our knowledge. It estimates how many cards the subject actually used in excess of the minimum necessary to achieve the six categories (or the possibly lowest number of categories effectively detected). The global score is computed by subtracting from the total number of administered trials the number of categories completed multiplied by ten (as ten is the number of correct matches required for each category). The formula is:

Global score = [n° of trials - (n° of achieved categories x 10)]

The global score ranges from a worst of 128 to a theoretical best of 0, i.e. the lower the score, the better the performance. It allows capturing, in a single measure, the combined information of the four measures suggested by Heaton [14]: number of categories completed, number of trials administered, percent conceptual level responses and total number of errors. Different authors have adopted several other scoring procedures (e.g. Heaton et al. [24]), but we think that this global score is useful in identifying dysexecutive patients in a concise and informative way.

Table 1 Distribution of the experimental sample according to age and education level. Values are numbers of subjects

	Age, years						
	15-29	30-39	40-49	50-59	60-69	70-85	Total
Educational level							
5-7 years							
men	0	0	0	8	5	2	15
women	0	0	6	7	5	4	22
8-12 years							
men	13	8	10	4	6	1	42
women	10	7	7	6	7	4	41
13-16 years							
men	4	4	1	6	4	2	21
women	8	4	4	3	4	0	23
17-24 years							
men	5	5	4	2	5	1	22
women	3	4	3	4	4	1	19
Total							
men	22	17	15	20	20	6	100
women	21	15	20	20	20	9	105

Perseverations. This measure quantifies the perseverative behaviour. In this study we adopted the scoring and recording procedures suggested by Heaton [14], who considered whether or not the response card matches the 'perseverated-to' principle. In some instances the response might be both correct and perseverative and, in that condition, it is possible that the judgement 'right' given by the examiner reinforces the perseverative behaviour presented by the subject. Several reports [30] emphasised the occurrence of inaccuracies in the detection of perseverative responses and advised that less-trained examiners be supplied with further instructions [30-33].

Non-perseverative errors. According to Heaton [14], incorrect responses that do not match the perseverated-to principle are scored as non-perseverative errors. These errors may indicate lack of strategy in the search for the correct matching and a tendency to give chance and sometimes-bizarre responses that do not match any of the sorting criteria (colour, form, number).

Failure to maintain the set. A failure to maintain the set occurs whenever an incorrect response follows a consecutive series of correct matches. Whereas Heaton [14] suggested that the correct sequence should be composed of at least five responses, in our normative study we reduced the number of correct matches to four (ambiguous or not), to make the test more sensitive to detecting the failure to maintain the set (Fig. 1, first column, trial 7; second column, trial 58; third column, trial 31).

Weigl sorting test

This task is designed to assess the ability to categorise stimuli according to some perceptual features, and to shift to a different criterion on a test repetition. Twelve geometrical shapes were displayed in front of each subject for the whole length of the test. The stimuli differed as regards form, colour, size, thickness, and on each of them a suit (hearts, clubs, diamonds or spades) was depicted. The subject was asked to sort several times the same set of stimuli according to one of the five similarity criteria to be discovered. After the discovery of each criterion, and the related sorting, the examiner scrambled the stimuli, and the subject shifted to a new sorting criterion, and so on. Material descriptions, administration and scoring instructions are reported in previous papers [29, 34, 35].

Statistical methods

The choice of the statistical methods was prompted by the need to obtain norms that could be directly confronted with the already available norms of a wide set of other neuropsychological tests. This possibility is granted by the equivalent scores (ES) procedure [28].

The influence of age, education and gender was evaluated through a covariance linear model [28, 36]. Several models of dependence were analysed and we adopted the transformation of the concomitant variables, which proved most effective in reducing the residual variance. The effect of each variable was studied partialling out the effect held in common with the other variables. In that way we built a linear model through which it was possible to calculate the expected score of a given subject taking into account age, education and gender. Taking this model as the basis, from the raw score of our subjects we calculated an adjusted score, by adding or subtracting the contribution of the significant concomitant variables. Adjusted scores were then ranked, and by means of a non-parametric procedure we set tolerance limits [37]. In this way we found both the *outer* and the *inner* one-sided tolerance limits. Above the outer tolerance limit we found *at least* 95% of the normal population (with 95% confidence): when a score was below the

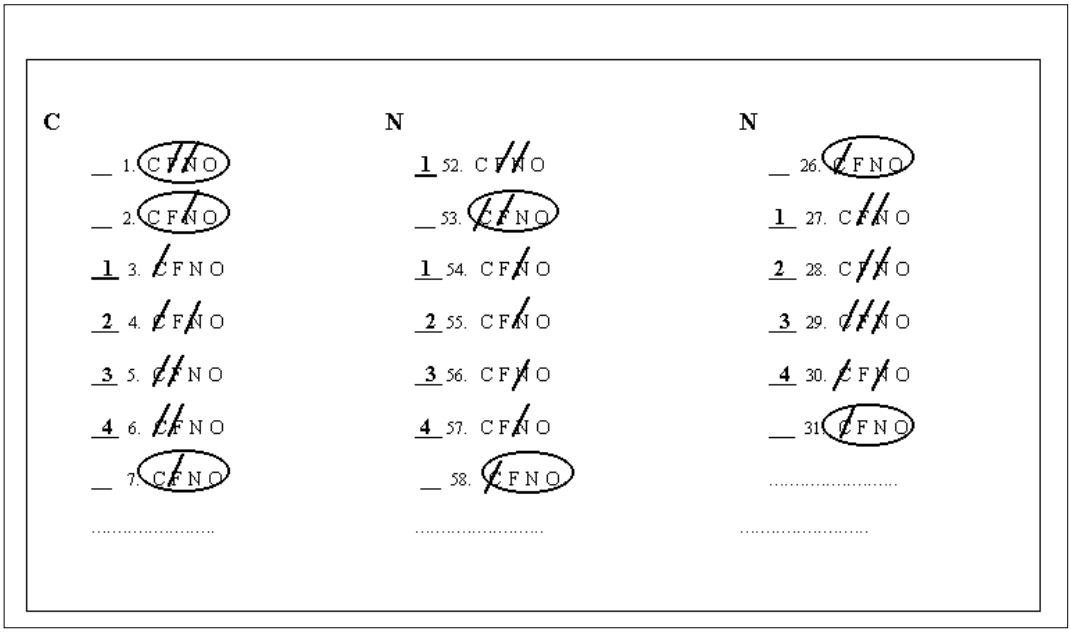

Fig. 1 Different examples of failure to maintain the set after a sequence of four correct responses (in the first column the presence of one unambiguous correct match should have suggested the correct criterion, in the second column a set loss occurred after a sequence of unambiguous correct responses and in the third column we can see that the overall analysis of the four ambiguous correct matches should have suggested the correct criterion). *C*, colour; *F*, form; *N*, number

outer tolerance limit, we declared a subject "not normal" with 95% confidence. Above the inner tolerance limit we found *at most* 95% of the population (with 95% confidence): when a score was above the inner tolerance limit, we declared a subject 'normal' with 95% confidence. When a score fell between the outer and inner tolerance limits, no inferentially controlled judgement was possible.

To avoid errors due to the fixed scale limits of the test scores, no adjustment was made to scores corresponding to the scale ends. The adjusted scores were classified into five categories (equivalent scores) endowed with an ordinal relationship: 0 = scores lower than the outer 5% tolerance limits; 4 = scores higher than the median value of the sample; 1, 2 and 3 = intermediate scores between the central value and the pathology threshold. A wider explanation of the ES scoring system has been given by Capitani [36] and Capitani and Laiacona [28].

To illustrate the raw score adjustment, let us consider the case of a 20-year-old male, brain-injured patient with 8 years of schooling. The global score achieved on WCST was 78. The adjusted score became 78 + 7.3 = 85.3. The corresponding equivalent score was 1. The number of perseverative responses was 19. The adjusted score was 19 + 6.4 = 25.4 and the ES = 2. The number of non-perseverative errors was 19. The corresponding adjusted score was 19 + 1.7 = 20.7 and the ES = 2. The patient failed to maintain the set 4 times. No adjustment was necessary and the corresponding ES = 0. On the Weigl test, the patient achieved a score of 9. The corresponding adjusted score was 9-0.3 = 8.7 and the ES = 1. In this patient we observed a perfect agreement between the global score of the WCST and the score achieved on the Weigl test.

Results

The Wisconsin card sorting test

Table 2 shows the outcome of global score analysis. The means for the different age groups are reported at the top: the score was higher as age increased, indicating a worsening in the performance of elders, and higher education was associated with a better performance (lower global score). Both age and education were found to significantly influence the scores ($p<0.0001$), whereas gender was not influential. The linear model which proved to be most effective in reducing the residual variance was:

$$y = 41.21 + 0.53 \times (\text{age} - 46.48) - 1.97 \times (\text{education} - 11.44)$$

This accounts for 25.3% of the variance. Table 2 also shows a correction grid with the points to add to, or subtract from the raw scores in order to obtain adjusted scores. For the combinations not reported in Table 2, either an interpolation between the reported adjustments has to be made, or the proper adjustment has to be directly calculated using the linear model above and reversing the sign of the parameters. Values equal to or higher than 90.6 (outer non-parametric tolerance limit) indicate a pathological performance; values equal to or lower than 81.9 (inner non-parametric tolerance limit) indicate a normal performance; intermediate scores (from 90.5 to 82.0) mean that performance is borderline. In the same table the values delimiting the equivalent scores, the number of the subjects of the sample comprised within each equivalent score (density) and the cumulative frequency of subjects comprised from 0 to 1, 2, 3 and 4 equivalent scores are reported.

Table 3 shows the outcome obtained with perseverative responses. Perseverative responses increased with age and decreased with higher education. A marginal advantage of male subjects was found ($p = 0.0537$); age and education significantly influenced the performances ($p<0.0001$). The best linear model was:

$$y = 17.89 + 0.313 \times (\text{age} - 46.48) - 1.035 \times (\text{education} - 11.44) - 1.652 \text{ (if male) or } + 1.652 \text{ (if female)}$$

The model explains 32% of the variance. The correction grids indicate the adjustment to be added to or subtracted from the raw scores, separately for males and females. For combinations not reported, the procedure to calculate the adjustment is that indicated above. Scores equal to or higher than 42.7 point to a pathological performance; scores equal to or lower than 36.8 indicate a normal performance and scores between 36.9 and 42.6 (included) indicate a borderline outcome. The values delimiting the equivalent scores are also reported in Table 3.

Table 4 shows the results of non-perseverative error analysis. These errors increased with age and were more frequent in the less-educated subjects; age and education significantly influenced the performances ($p = 0.0003$ and $p = 0.0002$, respectively), whereas gender did not ($p <1$). The best linear model was:

$$y = 11.82 + 0.119 \times (\text{age} - 46.48) - 0.406 \times (\text{education} - 11.44)$$

It accounted for 15.1% of the variance. The correction grid indicates the adjustments for different age/education combinations (for the combinations not reported, see above). Scores equal to or higher than 30.0 fall within the pathology range; scores equal to or lower than 23.2 are in the normality range; intermediate scores (from 23.3 to 29.9) are borderline. The equivalent scores limits are reported in the same table.

Table 5 shows the results of the analysis of failure to maintain the set. As demographic variables never influenced the performance, the model reduces to the mean. The outer tolerance limit, above which a subject can be considered pathological with a controlled risk, is 4. A subject can be declared normal when failures are equal to or less than 1 (inner tolerance limit). 2 and 3 failures are borderline. The values delimiting the equivalent scores are reported in the same Table.

Table 2 Global score on WCST, determined as n° of administered trials - (n° of achieved categories x 10). **a** Mean values per age and education group over 205 subjects (standard deviations) are reported; general mean score: 41.2 (SD = 28.4). **b** Correction grid. **c** Equivalent score

a Means

		Age, years						
		15-29	30-39	40-49	50-59	60-69	70-85	Total
Education, years								
	M	–	–	–	47.8 (35.9)	77.6 (15.9)	4.0 (48.1)	57.2 (33.2)
5-7	F	–	–	69.3 (19.6)	51.0 (27.0)	75.4 (22.8)	79.5 (18.9)	66.7 (24.2)
	M	21.2 (13.4)	37.8 (25.3)	49.6 (32.3)	44.3 (28.6)	58.3 (33.3)	38.0 (0.0)	39.0 (27.8)
8-12	F	28.2 (23.1)	41.1 (14.3)	39.4 (27.8)	60.3 (34.7)	59.1 (37.2)	80.0 (25.9)	47.4 (30.7)
	M	33.5 (26.0)	36.8 (21.0)	17.0 (0.0)	34.0 (27.6)	33.8 (15.9)	57.5 (57.3)	35.8 (25.1)
13-16	F	30.5 (21.6)	42.5 (27.4)	18.3 (4.1)	25.3 (20.5)	31.8 (16.4)	–	30.0 (19.6)
	M	13.8 (3.5)	23.4 (18.0)	20.8 (9.4)	24.0 (2.8)	38.6 (28.3)	23.0 (0.0)	24.2 (17.5)
17-24	F	38.3 (20.6)	21.8 (22.8)	30.0 (16.1)	27.0 (16.6)	19.8 (9.7)	88.0 (0.0)	29.8 (21.5)
	M	21.7 (15.6)	33.3 (22.1)	39.7 (30.0)	40.6 (29.5)	53.3 (29.2)	44.0 (35.9)	37.8 (27.7)
Total	F	30.5 (21.3)	36.3 (21.1)	42.8 (27.5)	45.2 (29.2)	49.9 (32.7)	80.7 (19.8)	44.5 (28.9)

b Correction grid

	Age, years						
Education, years	20	30	40	50	60	70	80
5	1.3	-4.0	-9.3	-14.6	-19.9	-25.2	-30.5
8	7.3	2.0	-3.3	-8.6	-13.9	-19.2	-24.5
13	17.1	11.8	6.5	1.2	-4.1	-9.4	-14.7
17	25.0	19.7	14.4	9.1	3.8	-1.5	-6.8

c Equivalent scores

	Score interval	Density	Cumulative frequency
0	128 - 90.6	5	5
1	90.5 - 81.5	14	19
2	81.4 - 59.4	33	52
3	59.3 - 37.2	50	102
4	37.1 - 0	103	205

Table 3 Perseverative responses on WCST. **a** Mean values per age and education group over 205 subjects (standard deviations) are reported. General mean score: 17.9 (SD = 14.6). **b** Correction grid. **c** Equivalent scores

a means

		Age, years						
		15-29	30-39	40-49	50-59	60-69	70-85	Total
Education, years								
	M	–	–	–	18.4 (17.3)	40.2 (9.9)	20.0 (22.6)	25.9 (18.0)
5-7	F	–	–	29.2 (17.8)	24.7 (14.2)	36.0 (7.5)	47.3 (9.5)	32.6 (15.0)
	M	7.5 (5.0)	16.3 (14.4)	19.6 (13.4)	17.8 (8.5)	27.2 (20.2)	20.0 (0.0)	16.1 (13.5)
8-12	F	11.8 (9.3)	16.0 (6.0)	19.1 (13.8)	21.3 (9.9)	24.4 (16.4)	52.8 (20.2)	21.3 (16.3)
	M	17.3 (14.6)	11.0 (5.4)	6.0 (0.0)	13.5 (8.4)	14.0 (7.3)	19.5 (19.1)	14.0 (9.5)
13-16	F	11.6 (7.6)	17.3 (13.5)	9.0 (2.4)	14.3 (15.4)	15.3 (10.1)	–	13.1 (9.4)
	M	5.0 (2.0)	6.2 (4.0)	6.8 (2.5)	6.5 (0.7)	17.8 (12.4)	12.0 (0.0)	9.0 (7.8)
17-24	F	17.3 (9.6)	6.2 (3.2)	14.0 (10.1)	10.3 (7.1)	6.3 (1.9)	32.0 (0.0)	11.4 (8.7)
	M	8.7 (8.0)	12.1 (10.9)	15.3 (12.5)	15.6 (12.4)	25.5 (16.4)	18.5 (13.6)	15.6 (13.4)
Total	F	12.5 (8.5)	13.7 (8.9)	19.4 (14.6)	19.3 (12.5)	21.9 (15.1)	48.0 (15.2)	20.1 (15.4)

b Correction grid

		Age, males						Age, females						
Education, years	20	30	40	50	60	70	80	20	30	40	50	60	70	80
5	3.3	0.1	-3.0	-6.1	-9.2	-12.4	-15.5	0.0	-3.2	-6.3	-9.4	-12.5	-15.7	-18.8
8	6.4	3.2	0.1	-3.0	-6.1	-9.3	-12.4	3.1	-0.1	-3.2	-6.3	-9.4	-12.6	-15.7
13	11.6	8.4	5.3	2.2	-1.0	-4.1	-7.2	8.3	5.1	2.0	-1.1	-4.3	-7.4	-10.5
17	15.7	12.6	9.4	6.3	3.2	0.0	-3.1	12.4	9.3	6.1	3.0	-0.1	-3.3	-6.4

c Equivalent Scores

	Score interval	Density	Cumulative frequency
0	128 - 42.7	5	5
1	42.6 - 35.8	14	19
2	35.7 - 24.0	33	52
3	23.9 - 17.1	50	102
4	17.0 - 0	103	205

Table 4 Non-perseverative errors on WCST. **a** Mean values per age and education group over 205 subjects (standard deviations) are reported. General mean score: 11.8 (SD = 7.9). **b** Correction grid. **c** Equivalent scores

a Means

		Age, years											
		15-29		30-39		40-49		50-59		60-69		70-85		Total			
Education, years																	
	M	–	–	–		13.6	(9.6)	17.4	(3.2)	12.5	(14.8)	14.7	(8.3)				
5-7	F					20.8	(2.3)	12.3	(8.2)	20.4	(6.8)	16.3	(8.8)	17.2	(7.4)		
	M	7.9	(4.8)	12.1	(5.4)	13.8	(7.5)	14.3	(11.9)	18.5	(9.5)	12.0	(0.0)	12.3	(7.6)		
8-12	F	8.1	(5.9)	11.4	(6.7)	11.4	(8.4)	15.7	(9.6)	17.3	(8.7)	12.8	(7.0)	12.4	(7.9)		
	M	7.5	(4.1)	10.5	(3.3)	8.0	(0.0)	11.8	(12.0)	12.8	(6.8)	21.0	(21.2)	11.7	(9.1)		
13-16	F	8.8	(6.9)	12.5	(9.8)	5.3	(1.9)	5.0	(1.0)	6.8	(3.9)	–		8.0	(6.2)		
	M	4.0	(1.9)	8.6	(5.9)	7.5	(3.1)	7.0	(2.8)	11.6	(9.8)	9.0	(0.0)	7.9	(5.9)		
17-24	F	14.0	(5.6)	9.5	(11.0)	7.7	(1.5)	11.0	(7.4)	6.3	(2.6)	27.0	(0.0)	10.5	(7.6)		
	M	7.0	(4.3)	10.7	(5.1)	11.7	(6.9)	12.6	(10.0)	15.4	(7.9)	14.7	(12.6)	11.6	(7.9)		
Total	F	9.2	(6.3)	11.2	(8.2)	12.5	(7.8)	12.0	(8.2)	13.8	(8.7)	15.9	(8.3)	12.1	(7.9)		

b Correction grid

	Age, years						
Education, years	20	30	40	50	60	70	80
5	0.5	-0.7	-1.8	-3.0	-4.2	-5.4	-6.6
8	1.7	0.6	-0.6	-1.8	-3.0	-4.2	-5.4
13	3.8	2.6	1.4	0.2	-1.0	-2.2	-3.4
17	5.4	4.2	3.0	1.8	0.6	-0.5	-1.7

c Equivalent Scores

	Score interval	Density	Cumulative frequency
0	128 -30.0	5	5
1	29.9 -22.5	14	19
2	22.4 -15.7	33	52
3	15.6 -10.1	49	101
4	10.0 - 0	104	105

Table 5 Failure to maintain the set on WCST. **a** Mean values per age and education group over 205 subjects (standard deviations) are reported. General mean score: 0.4 (SD = 0.9). **b** Equivalent scores

a Means

		Age, years											
		15-29		30-39		40-49		50-59		60-69		70-85		Total	
Education, years															
	M			–		–		–		1.1	(2.1)	0.4	(0.5)	0.5 (0.7)	0.8 (1.6)
5-7	F			–		–		0.3	(0.8)	0.9	(1.9)	0.8	(0.8)	0.3 (0.5)	0.6 (1.2)
	M	0.1	(0.3)	0.3	(0.5)	0.6	(1.6)	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)	0.0 (0.0)	0.2 (0.8)
8-12	F	0.2	(0.6)	0.6	(0.8)	0.1	(0.4)	1.0	(1.3)	0.4	(0.8)	0.0	(0.0)	0.4 (0.8)	
	M	0.5	(0.6)	1.0	(1.4)	0.0	(0.0)	0.2	(0.4)	0.5	(0.6)	0.0	(0.0)	0.4 (0.7)	
13-16	F	0.3	(0.5)	1.0	(1.2)	0.3	(0.5)	0.0	(0.0)	0.3	(0.5)	–		0.3 (0.6)	
	M	0.4	(0.5)	0.8	(1.8)	0.3	(0.5)	0.5	(0.7)	0.4	(0.5)	0.0	(0.0)	0.5 (0.9)	
17-24	F	0.3	(0.6)	0.5	(1.0)	0.0	(0.0)	0.0	(0.0)	0.3	(0.5)	0.0	(0.0)	0.2 (0.5)	
	M	0.2	(0.4)	0.6	(1.2)	0.5	(1.3)	0.6	(1.4)	0.3	(0.5)	0.2	(0.4)	0.4 (1.0)	
Total	F	0.2	(0.5)	0.7	(0.9)	0.2	(0.5)	0.6	(1.3)	0.5	(0.7)	0.1	(0.3)	0.4 (0.8)	

b Equivalent scores

	Score interval	Density	Cumulative frequency
0	5 - 4	5	5
1	3 - 2	11	16
2	1 - 1	34	50
3-4	0	155	205

A word of caution is in order with regard to the interpretation of adjusted performance of the subjects older than 70. Besides the small frequencies of these subjects in our survey, this age group is often at risk of selection because general health and census could prevent from observing less efficient subjects. This could explain the counterintuitive improvement presented by elders in some tasks.

The Weigl sorting test

Means are shown in Table 6. The performance worsened in elders and in subjects with lower education: both age and education significantly influenced the outcome ($p = 0.002$ and $p > 0.0001$, respectively). Gender was not significant ($p = 0.218$). The best linear model was:

$$12.69 - 0.029 \times (\text{age} - 46.48) + 0.148 \times (\text{education} - 11.44)$$

The variance accounted for was 17.1%. The adjustments to be added to or subtracted from the raw scores according to the model are reported in the correction grid in some age/education combinations. For the combinations not reported, see above. Scores equal to or lower than 8.0 are pathological; scores equal to or higher than 9.6 are normal and scores from 8.1 to 9.5 are borderline. The table shows also the values delimiting the equivalent scores, the density of subjects comprised within the limit values of each Equivalent Score and the cumulative frequency of subjects observed in the Equivalent Score cells of 0 to 4.

Table 6 The Weigl sorting test. **a** Mean values per age and education group over 205 subjects (standard deviations) are reported. General mean score: 12.7 (SD = 2.3). **b** Correction grid. **c** Equivalent scores

a Means

		Age, years						
		15-29	30-39	40-49	50-59	60-69	70-85	Total
Education, years								
	M	–	–	–	12.4 (2.5)	11.8 (2.9)	11.0 (5.6)	12.0 (2.9)
5-7	F	–	–	9.2 (2.5)	12.0 (1.7)	11.8 (3.7)	9.0 (2.0)	10.6 (2.8)
	M	12.9 (1.9)	13.0 (2.3)	12.7 (1.9)	13.5 (1.7)	11.8 (1.8)	13.0 (0.0)	12.8 (1.9)
8-12	F	13.7 (1.7)	12.0 (2.3)	11.4 (1.0)	13.2 (2.1)	11.7 (2.4)	11.0 (1.8)	12.3 (2.1)
	M	13.8 (2.5)	14.3 (1.5)	15.0 (0.0)	12.3 (2.3)	13.0 (2.4)	11.0 (1.4)	13.1 (2.2)
13-16	F	14.6 (1.1)	13.8 (1.5)	11.8 (2.4)	13.3 (1.5)	12.0 (2.4)	–	13.3 (2.0)
	M	14.4 (1.3)	15.0 (0.0)	13.0 (2.4)	12.5 (0.7)	13.0 (2.1)	15.0 (0.0)	13.8 (1.7)
17-24	F	14.0 (1.7)	13.8 (1.5)	12.3 (2.5)	12.5 (2.9)	15.0 (0.0)	15.0 (0.0)	13.6 (2.0)
	M	13.4 (1.9)	13.9 (1.9)	12.9 (2.0)	12.6 (2.1)	12.4 (2.2)	12.0 (3.1)	13.0 (2.1)
Total	F	14.1 (1.5)	12.9 (2.0)	11.0 (2.2)	12.7 (2.0)	12.5 (2.7)	10.1 (2.6)	12.4 (2.4)

b Correction grid

	Age, years						
Education, years	20	30	40	50	60	70	80
5	0.2	0.5	0.8	1.1	1.3	1.6	1.9
8	-0.3	0.0	0.3	0.6	0.9	1.2	1.5
13	-1.0	-0.7	-0.4	-0.1	0.2	0.5	0.7
17	-1.6	-1.3	-1.0	-0.7	-0.4	-0.1	0.1

c Equivalent Scores

	Score interval	Density	Cumulative frequency
0	0 - 8.0	5	5
1	8.1 - 9.8	13	18
2	9.9 - 11.1	36	54
3	11.2 - 12.9	48	102
4	13.0 - 15.0	103	205

Correlation between WCST global score and Weigl test

As the two tests are supposed to be sensitive to the same psychological ability, we confronted the WCST global score with the Weigl test score. Figure 2 shows the bi-variate frequency distribution of WCST and Weigl test scores. We calculated the correlation between the original scores, and the agreement between the subject classifications according to the Equivalent Scores procedure. As both tests are influenced by demographic variables, the correlation was computed partialling out the influence of age and education on the performance. The r correlation value was 0.202, $p < 0.01$. In addition, we considered the Equivalent Score classification of each subject and looked at the agreement between the two tests (Table 7). After the transformation into Equivalent Scores, the tests could be directly compared; thus it was possible to check whether or not these two tests claiming to assess executive functions yield a comparable classification. The weighted Cohen's K [38] computed on the data of Table 7 yielded a value of 0.164 (95% Confidence interval: 0.032-0.296): this points to a moderate agreement between the two measures.

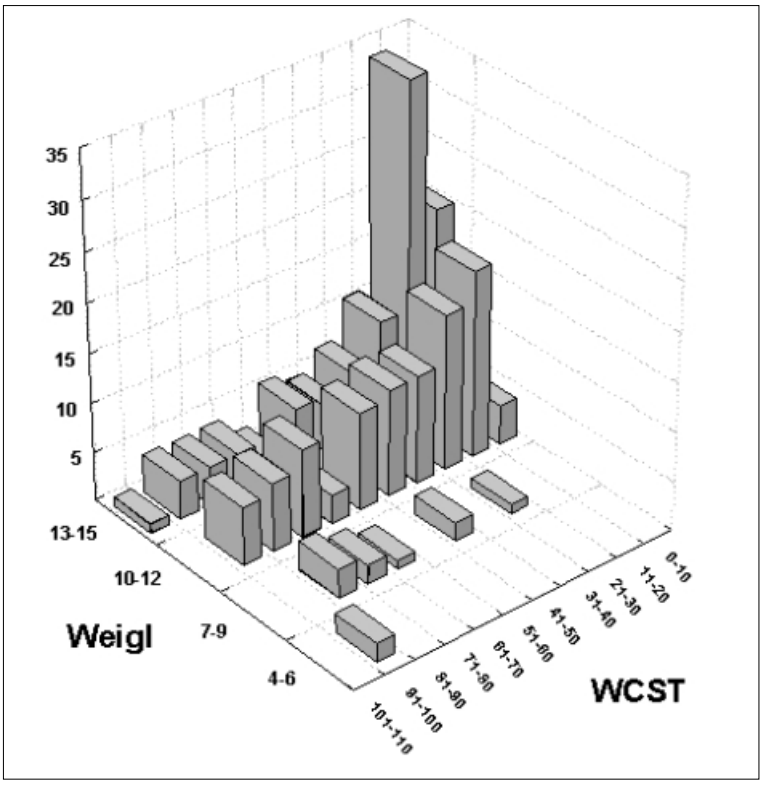

Fig. 2 Bi-variate frequency distribution of WCST and Weigl test scores. In the top right corner of the plot, where most observations cumulate, we find subjects performing well on both tests

Table 7 Agreement between the Equivalent Scores classification of WCST and Weigl test. We report the cross tabulation of the equivalent scores classification of each test, which goes from 0 to 4

Weigl	WCST					
	0	1	2	3	4	Total
0	–	–	3	2	–	5
1	1	–	2	4	7	14
2	–	7	7	12	11	37
3	1	3	6	12	27	49
4	3	4	15	20	58	100
Total	5	14	33	50	103	205

Discussion

The first aim of this study was to offer normative values for two tests widely used in clinical practice to assess executive functions. The Equivalent Score standardisation methodology is widely used in Italy, making it possible to work with adjusted scores from any test whatsoever and make a direct comparison among different tools. However, a standardisation is valid only for the specific population on which it has been calculated. The norms reported in the original Wisconsin sorting test manual [14] were obtained from subjects living in the USA, while the existing Weigl test standardisation [29] was performed on Italian subjects over age 40 years. It is likely that for younger subjects the necessary adjustments cannot be calculated simply with an extension of the estimated regression functions. These considerations emphasise the need for of the present normative data as (i) they refer to Italian subjects aged 15 to 85 years, and (ii) they conform to the Equivalent Scores methodology.

Frontal lobe dysfunctions are present in many different clinical pathologies frequently observed, such as neoplasias and brain injury, and are prominent in the degenerative diseases affecting the frontal lobe [39]. Patients with frontal lobe injuries present attentional defects, react impulsively to stimuli, are context-dependent, do not plan their actions and do not shift strategies; perseveration, disinhibition, inertia and depression are other common behavioural aspects. Despite these clinical complaints, few tasks for 'frontal functions' are available for Italian subjects and often without the possibility

of a direct comparison with one another (visual attention, cancellation test and reversal learning [29]; cancellation test [40]; categorisation and recall of pictures and odd-man-out test [41]; verbal fluency [42]). In this study we collected normative data concerning two such tests.

A point deserving attention is the introduction of the global score in the WCST evaluation. At variance with the recording and scoring procedures suggested by Heaton [14], this measure yields a global judgement of the performance, satisfying the main query arising in clinical practice. In fact, according to Heaton [14], in order to achieve a similar judgement one has to consider a set of different measures, thus making a synthetic and handy evaluation more difficult. Besides this global judgement, the clinician interested in more qualitative aspects of the patient's performance can refer to the specific measures that quantify the perseverative behaviour (perseverative responses), the inability to plan a strategy (non-perseverative errors), or the inability to sustain attention and to suppress responses to irrelevant stimuli (failure to maintain the set). This further fractionation of the performances is more informative in terms of establishing anatomoclinical correlations [43] and, together with the global judgement, offers a possible outline to follow a patient's recovery and the efficacy of a treatment, and is useful for a more reliable prognosis.

A final remark concerns the agreement between the global score of the WCST and the Weigl test. Our results confirm that the two measures are not independent, but indicate only a moderate overlapping. Figure 2 shows that the dispersion of WCST is greater than that of Weigl test: it can be easily seen that, for a three-point interval in the top range of Weigl test, the WCST scores can have very different values. On the other hand, we found that 8.8% of the subjects were definitely normal on WCST (equivalent scores of 2 to 4), but had an equivalent score of 0 or 1 on the Weigl test. The same percentage of 8.8% was observed for the complementary dissociation. This suggests that, in fact, these tests may in part be sensitive to different psychological abilities. In particular Weigl test is much shorter and less demanding, and this could favour patients affected by a limited span of sustained attention. On the contrary Weigl test could be more exacting for visuoperceptual skills (for instance many normal subjects fail to detect the different width of the tokens), and the different categorisation criteria have a clear rank of perceptual saliency. Further, the WCST calls for a far greater memory component than Weigl test. Needless to say, the moderate correlation observed with normals does not necessarily mean that a comparable outcome would be found with pathological samples. In this case, it is possible that the similarities between our tests would prevail over the discrepancies, and that the agreement between a normal/pathological classification would be substantially greater. However, this is a matter for further empirical enquiry.

Acknowledgements We are grateful to F. Comazzi for the assistance with the drawing of Fig. 2. This research was supported by a Grant from the Fondazione Valduce. Rosemary Allpress revised the English text.

Sommario *Nella valutazione clinica delle funzioni frontali vengono spesso usati sia il Wisconsin card sorting test che il test di Weigl. In questo lavoro vengono presentati per entrambi i test i dati normativi validi per la popolazione italiana (età tra i 15 e gli 85 anni e con una scolarità almeno di 5 anni). Per quanto riguarda il Wisconsin card sorting test, si propone sia una nuova misura di efficienza globale (global score) che valori normativi per alcuni aspetti qualitativi delle prestazioni (risposte perseverative, difficoltà nel mantenere un criterio ed errori non perseverativi). Nella definizione dei dati normativi, si è seguito il sistema dei punteggi equivalenti usato in Italia per diversi altri test neuropsicologici, in modo da consentire un confronto reciproco tra le varie prove. Viene inoltre presentato uno studio sulla correlazione tra il Wisconsin card sorting test (global score) ed il test di Weigl dove risulta che le due prove non sono completamente sovrapponibili.*

References

1. Berg EA (1948) A simple objective test for measuring flexibility in thinking. J Gen Psychol 39:15-22
2. Grant DA, Berg EA (1948) A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card sorting problem. J Exp Psychol 34:404-411
3. Weigl E (1941) On the psychology of so-called processes of abstraction. J Abnorm Soc Psychol 36:3-33
4. Bornstein RA (1986) Contribution of various neuropsychological measures to detection of frontal lobe impairment. Int J Clin Neuropsychol 8:18-22
5. Milner B (1963) Effects of different brain lesions on card sorting. Arch Neurol 9:90-100
6. Gormezano I, Grant DA (1958) Progressive ambiguity in the attainment of concepts on the Wisconsin card sorting test. J Exp Psychol 55:621-627
7. Teuber HL, Batterbsby WS, Bender MB (1951) Performance of complex visual tasks after cerebral lesions. J Nerv Ment Dis 114:414-429
8. Drewe EA (1974) The effect of type and area of brain lesions on Wisconsin card sorting test performance. Cortex 10:159-170
9. Nelson HE (1976) A modified card sorting test sensitive to frontal lobe defects. Cortex 12:313-324
10. Grant DA, Curran JF (1952) Relative difficulty of number, form and color concepts of Weigl-type problems using unsystematic number cards. J Exp Psychol 43:408-413
11. Tarter RE (1973) An analysis of cognitive deficits in chronic alcoholics. J Abnorm Psychol 77:71-75
12. Fey ET (1951) The performance of young schizophrenics and young normals on the Wisconsin card sorting test. J Cons Psychol 15:311-319
13. Ross BM, Rupel JW, Grant DA (1952) Effects of personal, impersonal and physical stress upon cognitive behavior in a card sorting problem. J Abnorm Soc Psychol 47:546-551
14. Heaton KR (1981) A manual for the Wisconsin card sorting test. P.E.A., Odessa, Fl
15. Axelrod BN, Paolo AM, Abraham E (1997) Do normative data from the full WCST extend to the abbreviated WCST? Assessment 4:41-46

16. De Zubicaray G, Ashton R (1996) Nelson's (1976) modified card sorting test: A review. Clin Neuropsychologist 10:245-254
17. Robinson AL, Heaton RK, Lehmam RA, Stilson D (1980) The utility of Wisconsin card sorting test in detecting and localizing frontal brain lesions. J Cons Clin Psychol 48:605-614
18. Robinson LJ, Kester DB, Saykin AJ, Kaplan EF, Gur RC (1991) Comparison of two short forms of the Wisconsin card sorting test. Arch Clin Neuropsychol 6:27-33
19. Shallice T (1982) Specific impairments in planning. In: Broadbent DE, Weis Krantz L (eds) The neuropsychology of cognitive function. The Royal Society, London, pp 199-209
20. Anderson SW, Damasio H, Jones RD, Tranel D (1991) Wisconsin card sorting test performance as a measure of frontal lobe damage. J Clin Exp Neuropsychol 13:909-922
21. Mountain MA, Snow WG (1993) Wisconsin card sorting test as a measure of frontal pathology: A review. Clin Neuropsychologist 7:108-118
22. Axelrod BN, Goldman RS, Heaton RK, Curtiss G, Thompson LL, Chelune GJ, Kay GG (1996) Discriminability of the Wisconsin card sorting test using the standardization sample. J Clin Exp Neuropsychol 18:338-342
23. Rezai K, Andreasen NC, Alliger R, Cohen G, Swayze II, V, O'Leary DS (1993) The neuropsychology of the prefrontal cortex. Arch Neurol 50:636-642
24 Heaton KR, Chelune GJ, Talley Kay GG, Curtiss G (1993) Wisconsin card sorting test manual revised and expanded. P.E.A., Odessa, Fl
25 Corcoran R, Upton D (1995) The role of right temporal lobe in card sorting: A case study. Cortex 31:405-409
26 Corcoran R, Upton D (1996) A role for hippocampus in card sorting? Cortex 32:188-189
27 De Zubicaray G, Ashton R (1996) 'A role for the hippocampus in card sorting?' A cautionary note: A comment to Corcoran and Upton. Cortex 32:187-188
28. Capitani E, Laiacona M (1997) Composite neuropsychological batteries and normative values. Standardisation based on equivalent scores, with a review of published data. J Clin Exp Neuropsychol 19:795-809
29. Spinnler H, Tognoni G (eds) (1987) Standardizzazione e taratura italiana di test neuropsicologici. Ital J Neurol Sci 6[Suppl 8]
30. Axelrod BN, Greve KW, Goldman RS (1991) Comparison of four Wisconsin card sorting test scoring guides with novice raters. Source Assessment 1:115-121
31. Flashman LA, Horner MD, Freides D (1991) Note on scoring perseveration on the Wisconsin card sorting test. Clin Neuropsychologist 5:190-194
32. Axelrod BN, Goldman RS, Woodard JL (1992) Interrater reliability in scoring the Wisconsin card sorting test. Clin Neuropsychologist 6:143-155
33. Berry S (1996) Diagrammatic procedure for scoring the Wisconsin card sorting test. Clin Neuropsychologist 10:117-121
34. De Renzi E, Faglioni P, Savoiardo M, Vignolo LA (1966) The influence of aphasia and the hemisphere side of the cerebral lesion on abstract thinking. Cortex 2:399-420
35. Basso A, Capitani E, Luzzatti C, Spinnler H, Zanobio E (1985) Different basic components in the performance of Broca's and Wernicke's aphasics on the colour-figure matching test. Neuropsychologia 23:51-59
36. Capitani E (1997) Normative values and neuropsychological assessment. Common problems in clinical practice and research. Neuropsychol Rehab 7:295-309
37. Wilks SS (1941) Determination of sample size for setting tolerance limits. Ann Math Stat 12:91-96
38. Mehta C, Patel N (1996) StatXact for Windows. Cytel Software, Cambridge
39. Kertesz A, Munoz DG (Eds) (1998) Pick's disease and Pick complex. Wiley-Liss, New York
40. Della Sala S, Laiacona M, Spinnler H, Ubezio MC (1992) A cancellation test:its reliability in assessing attentional deficits in Alzheimer's disease. Psychol Med 22:885-901
41. Pomati S, Farina E, Magni E, Laiacona M, Mariani C (1996) Normative data for two neuropsychological tests sensitive to frontal dysfunction. Ital J Neurol Sci 17:201-209
42. Capitani E, Laiacona M, Basso A (1998) Phonetically cued word-fluency, gender differences and aging: a reappraisal. Cortex 34:779-783
43. Stuss DT, Levine B, Alexander MP, Hong J, Palumbo C, Hamer L, Murphy KJ, Izukawa (2000) Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. Neuropsychologia 38:388-402