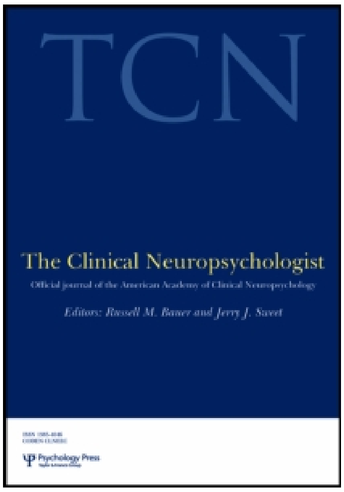

The Clinical Neuropsychologist
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713721659

Normative data and psychometric properties for qualitative and quantitative scoring criteria of the Five-point Test

Simone Goebel [a]; Regina Fischer [a]; Roman Ferstl [b]; Hubertus Maximilian Mehdorn [a]
[a] Department of Neurosurgery, University Hospital Schleswig-Holstein, Kiel [b] Institute for Psychology, University of Kiel, Kiel, Germany

First Published on: 16 October 2008

The Clinical Neuropsychologist, 2008, 1–16, iFirst
http://www.psypress.com/tcn
ISSN: 1385-4046 print/1744-4144 online
DOI: 10.1080/13854040802389185

Psychology Press
Taylor & Francis Group

NORMATIVE DATA AND PSYCHOMETRIC PROPERTIES FOR QUALITATIVE AND QUANTITATIVE SCORING CRITERIA OF THE FIVE-POINT TEST

Simone Goebel[1], Regina Fischer[1], Roman Ferstl[2], and Hubertus Maximilian Mehdorn[1]

[1]*Department of Neurosurgery, University Hospital Schleswig-Holstein, Kiel and* [2]*Institute for Psychology, University of Kiel, Kiel, Germany*

The Five-point Test (Regard, Strauss, & Knapp, 1982) was introduced for the measurement of figural fluency as part of the examination of executive functions. Until now, no differentiated norms exist. We present normative data for adults aged 18–80 (n = 280) for the number of unique designs (productivity), the percent of perseverations (flexibility), the percent of rotated (strategic) designs, and the number of rule breakings. As age and education were correlated with test performance, norms were stratified by these two variables. Test–retest reliability and inter-rater reliability were calculated. Moreover, convergent and divergent validity as well as factorial validity were assessed through intercorrelations and correlations with other neuropsychological tests. All together, the Five-point Test proved to be reliable and valid.

Keywords: Five-point Test; Design fluency; Normative data; Executive functioning; Reliability; Validity; Qualitative scoring.

INTRODUCTION

The technique of item generation is widely used in clinical and experimental neuropsychology for the assessment of executive functions. Typically, subjects are asked to produce as many unique items as possible within an allotted time period by following given rules while avoiding repetitions. Requested answers may be designs (e.g., Regard et al., 1982), words (e.g., Benton & Hamsher, 1989), or gestures (e.g., Jason, 1985). The number of correct responses as an indicator for the ability to initiate and sustain mental productivity as part of executive functioning is traditionally considered the most important score. However, in addition to this quantitative information, fluency measures yield considerable qualitative (or process-oriented) information as well. Abwender and colleagues (Abwender, Swan, Bowerman, & Connolly, 2001) stress that, in the framework of "process" approaches to neuropsychological testing, the examiner is not only interested in how well a subject performs, but also in *how* the examinee performs a task (see Kaplan, 1988). The main qualitative aspects of executive functions contributing to quantitative test

Address correspondence to: Dr. Simone Goebel, Department of Neurosurgery, University Hospital Schleswig-Holstein, Schittenhelmstr. 10, D-24105 Kiel, Germany. E-mail: goebels@nch.uni-kiel.de

Accepted for publication: August 5, 2008. First published online: Month/day/year.

results in fluency tasks are flexibility, rule-breaking behavior, and strategic performance. Flexibility and the capacity to shift are part of the ability of self-regulation (Lezak, Howieson, & Loring, 2004). Patients with frontal lobe lesions often exhibit deficits here in terms of increased perseverative errors (e.g., Salloway, 1994). Moreover, non-perseverative errors (rule breakings) can be observed in patients with difficulties in executive functions (Lezak et al., 2004). Finally, the utilization of adequate strategies can enhance performance. Lezak (1995) highlights that through an effective strategy each item no longer necessitates a unique solution; rather a series of items can be produced from one idea. Hereby, the use of production strategies can effectively maximize productivity while concurrently minimizing repetitious responses. Accordingly, Ruff (1988) documented that fewer production strategies are associated with more perseverative errors, and Ross and colleagues (Ross, Foard, Hiott, & Vincent, 2003) as well as Abwender et al. (2001) found positive correlations between the number of unique items and strategy indices. Ruff (1988) argued that the qualitative analysis of production strategies can differentiate whether a patient's impaired performance is due to an initiation deficit or deficient planning abilities: Whereas the former tend to produce very few items overall, the latter produce more designs at the expense of more perseverative errors and lower strategy scores.

Patients impaired in fluency measures are individuals with lesions of the frontal lobes (e.g., Baldo, Shimamura, Delis, Kramer, & Kaplan, 2001; Lee et al., 1997; Stuss et al., 1998; Tucha, Smely, & Lange, 1999), traumatic brain injury (e.g., Ruff, Evans, & Marshall, 1986), Parkinson's disease (PD) (e.g., Williams-Gray, Foltynie, Brayne, Robbins, & Barker, 2007), and cerebellar lesions (Gottwald, Wilde, Mihajlovic, & Mehdorn, 2004) as well as miscellaneous other patient groups (e.g., Harter, Hart, & Harter, 1999; Lezak et al., 2004).

Deficits in verbal fluency measures (e.g., Controlled Word Association Test; Benton & Hamsher, 1989) often accompany aphasic disabilities as well as lesions of the frontal lobes or the striatum (Stuss et al., 1998). Lesions of the left hemisphere have been reported to produce more severe deficits in verbal fluency than right-sided lesions (e.g., Baldo et al., 2001). Figural fluency tests are supposed to measure executive functions as a non-verbal analogue of verbal fluency measures. It has been proposed that tests of figural fluency are particularly sensitive for right-sided brain lesions (e.g., Baldo et al., 2001; Lee et al., 1997; Ruff, 1988; Ruff, Allen, Farrow, Niemann, & Wylie, 1994; but see Tucha et al., 1999).

The Design Fluency Test (Jones-Gotman & Milner, 1977) was the first measure of figural fluency that was accompanied by a multitude of theoretical and practical limitations (e.g., Lee et al., 1997). To overcome these problems Regard et al. (1982) developed the Five-point Test (FPT), which has been shown to be valid and reliable (e.g., Lezak, 1995; Spreen & Strauss, 1998). There are a number of adaptations of the FPT, of which the Ruff Figural Fluency Test (RFFT) is the most widespread one. The RFFT is longer and more complex than the FPT. Research has shown, however, that the use of multiple test conditions and trials in the RFFT might be redundant and unnecessary, and might penalize cognitively more impaired patients (Lee et al., 1997; Ruff, 1988; Spreen & Strauss, 1998).

Thus, compared to other figural fluency measures, advantages of the original FPT are its brevity and simplicity and therefore its adequacy for patients who are more severely impaired in basic cognitive functions. However, sufficient norms for its qualitative and quantitative test scores have not yet been provided.

In this study, the FPT was administered to a normative sample of healthy adults. Calculated test scores were productivity as well as the qualitative dimensions flexibility, strategic behavior, and rule breaking. It was predicted that young age and higher level of education should yield higher test performance. Furthermore, we anticipated that the strategy index would relate positively to productivity and negatively to perseverative errors. Scoring dimensions were evaluated for inter-rater and test–retest reliability. Convergent and divergent validity were assessed through intercorrelations and in relation to other neuropsychological measures.

METHOD

Participants

Participants were recruited through personal contacts, different courses at the local university or adult education center, sports clubs, music schools, or the employment center in three different German federal states. A self-reported history of medical and psychiatric problems was obtained from each participant. Any person with a history of neurological or psychiatric illness was excluded ($n = 17$). Participants did not receive any financial reward. All participants were living independently in the community. All participants aged 60+ completed the Mini Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) to control for global cognitive decline. Participants who solved less than 90% of the items (i.e., scored below 27 points) were excluded ($n = 2$). The final normative sample consists of 280 healthy adults (53.4% female) aged 18–80 ($M = 44.85$; $SD = 17.90$; skewness = .139; kurtosis = −1.260; Kolmogoroff-Smirnoff <.2). According to the procedure drawn from Tombaugh (2004), participants were classified as having lower (1–13 years) or higher (more than 13 years) education. This was mostly equivalent to people with or without university-entrance diploma (the German "*Abitur*"). Of the participants aged between 18 and 39 ($n = 118$), 35.6% ($n = 42$) were students, 47.5% ($n = 56$) working outside and 16.9% ($n = 20$) working inside the home. Of the participants aged between 40 and 59 ($n = 84$), 71.4% ($n = 60$) were working outside and 21.4% ($n = 18$) inside the home, and the remaining 7.2% ($n = 6$) had retired early. In the subgroup of participants aged 60 and older ($n = 78$), 26.9% ($n = 21$) were working outside and 23.1% ($n = 18$) inside the home, and 50% had retired. A total of 34 participants volunteered for a second test session for the establishment of test–retest reliability across a 4-week testing interval. Those were either working acquaintances or undergraduate students from psychology courses who received credit towards their courses. Age in this subgroup ranged from 19 to 67 years ($M = 29.4$, $SD = 11.1$) and 84% of these participants had received more than 13 years of education.

Measures

Five-point Test (FPT). The FPT was administered according to procedures drawn from Schnider (1997), using standard oral instruction and presented examples (Figure 1). Stimulus material consists of a page on which 35 identical squares are printed in seven rows and five columns, each square containing five symmetrically arranged dots. Participants were asked to draw as many different figures as possible in 3 minutes by connecting two or more dots with straight lines. The briefing that items should not be repeated was stressed.

Other neuropsychological measures. The following procedures were administered for the assessment of construct validity: lexical verbal fluency (LBS, German version of the FAS; Lezak, 1995), visuo-motor speed and the ability to shift (Trail-Making-Test (TMT) A and B; Reitan & Wolfson, 1985; Tombaugh, 2004), verbal short-term and working memory (Digit Span Forward and Backwards from the German WMS-R; Härting et al., 2000), and IQ via a German multiple choice vocabulary test, the Mehrwachwahl-Wortschatz-Intelligenztest MWT-B (Lehrl, 1999), a reliable and valid estimator of verbal IQ.

Procedure

After their informed consent had been obtained, participants first underwent the health survey and, if applicable, the MMSE, followed by the FPT. After test administration all participants were appropriately debriefed. A total of 34 participants underwent additional testing for assessment of reliability and validity. The supplemental procedures were completed after the FPT during the first session. Tests were always administered in the order described in the section above. The second session was scheduled about 1 month (plus/minus 3 days), on average 32.7 days, after the first.

To provide a stringent examination of inter-rater reliability, four raters (one PhD-level psychologist and three senior psychology majors) scored 127 randomly chosen test protocols. Each rater had received training in the administration and scoring of several common neuropsychological tests as well as the FPT by the primary investigator. As inter-rater consistency seems to be best promoted by providing raters with opportunity for group discussion and practice exercises (Cone, 1999), 20 practice protocols were scored by each rater and results were compared and discussed. Afterwards, each rater scored all protocols independently.

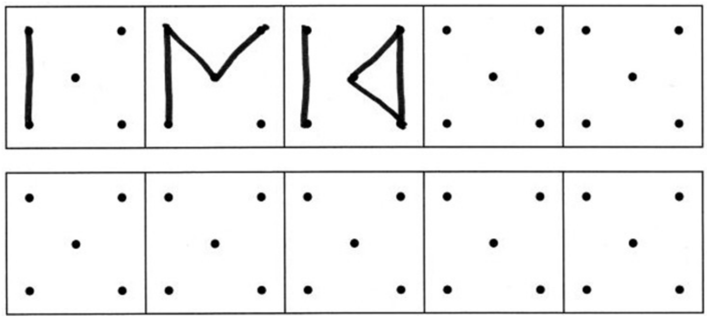

Figure 1 Test material and standardized examples for the Five-point Test (scaled down detail).

All calculations were done with SPSS 14.0. All tests were performed in a two-tailed fashion.

Scoring

Productivity. The number of unique designs is calculated by subtracting the number of perseverative designs and rule-breaking errors from the totality of produced designs.

Flexibility. The number of perseverative designs is divided by the number of unique designs and multiplied by 100 (percentage of perseverations).

Strategy score. A strategy point is given whenever the participant rotates and/or mirrors an item, so that it becomes different from the one directly before. The number of strategic items is divided by the number of unique designs and multiplied by 100 (strategy score). No strategy points are given if the rotated and/or mirrored item is itself an error.

Rule breaking. Non-perseverative errors (connecting dots from different squares, drawing curled lines or lines not connecting dots) are classified as rule breaking. We present a cut-off value valid for all participants, as rule breaking hardly ever occurred. For this, the number of rule breakings has to be counted.

RESULTS

Normative data

Correlations among demographic variables and FPT scores show that test scores are influenced by both age and education (Pearson's $r, p < .05$). Gender was slightly but significantly correlated with the strategy index (Table 1). However, after using partial correlations accounting for education, no significant correlation between gender and test performance remained ($r = -.082, p = .177$).

The relative effects of age, education, and gender on FPT scores were further explored by a regression analysis where each variable was entered separately. Regarding productivity, education and age accounted for 26.8% and 18.7% of variance. Gender did not contribute to variance ($R^2 = .000$). Regarding the percentage of perseverations, only age accounted for variance (9.1%), neither education ($R^2 = .027$) nor gender ($R^2 = .006$) contributed significantly.

Table 1 Correlations of age, education and gender with FPT performance and inter correlations of test scores

	Age	Education	Gender	FPT productivity	FPT% persever
Education	−0.387**				
Gender	−0.087	−0.165**			
PPT: productivity	−0.403**	0.509**	−0.001		
FPT: % persever	0.349**	−0.141**	−0.080	−0.169**	
FPT: % strategies	−0.308**	0.446**	−0.122*	0.492**	−0.256**

$* = p < .05$; $** = p < .01$

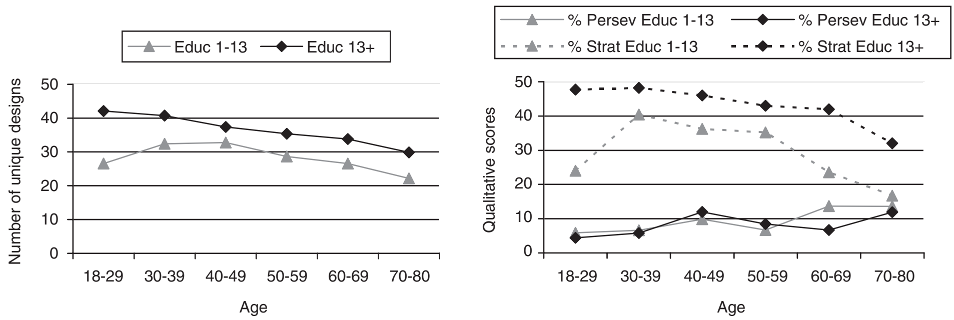

Figure 2 Performance on quantitative (left) and qualitative FPT scores as a function of six age groups and two education levels.

Table 2 Sample size, gender, and age for each normative group

Age	18–39		40–59		60–80	
Education	Normal 1–13	High 13+	Normal 1–13	High 13+	Normal 1–13	High 13+
N	32	86	47	37	57	21
Male/female	13/19	40/46	19/28	23/14	26/31	13/8
Age	18–39	18–36	40–59	41–59	60–80	60–80
	$M = 26.31$	$M = 26.84$	$M = 49.26$	$M = 49.72$	$M = 67.54$	$M = 67.10$
	$S = 8.08$	$S = 4.36$	$S = 5.38$	$S = 5.51$	$S = 5.57$	$S = 6.47$
	Md = 23.50	Md = 27.00	Md = 50.00	Md = 49.00	Md = 68.00	Md = 65.00

Finally, education and age accounted for 19.8% and 10.9% of variance of the strategy score whereas gender did not contribute significantly ($R^2 = .014$).

As performance was affected by age and education, norms were stratified by these two variables. In order to effect the ideal compromise between sample size and differentiation, scores were divided into three age and two education levels (Figure 2). The chosen age groups were (1) 18–39 years, (2) 40–59 years, and (3) 60–80 years. Age was normally distributed in all three age groups (Kolmogoroff-Smirnoff $p > .3$). Skewness was .249, −.07, or .575 respectively, and kurtosis values of age distribution constituted −.734, −.935, or −.603 respectively.

Multivariate analyses of variance appropriate for a 3 (Age) × 2 (Education) factorial design were performed on the three test scores. Productivity and strategy score decreased with increasing age and fewer years of education ($p < .002$). Percentage of perseverations increased with age ($p = .002$) but did not differ between participants with normal or high-level education ($p = .283$). Post-hoc tests (Scheffé) revealed significant differences between the three age groups for productivity ($p = .000$) and perseverations ($p < .01$). Regarding the strategy score, only age groups 2 (40–59) and 3 (60–80) differed significantly ($p = .000$) whereas performances of the two younger age groups were alike ($p = .395$). Therefore the six chosen norm groups were retained. Table 2 presents detailed information about the participants in each group. Tables 3–5 present the normative data for productivity, percentage of perseverations, and the strategy score.

Skewness for productivity and strategy score lay between −1 and 1 for all test groups (Kolmogoroff-Smirnoff $p > .3$). Thus, our norms cover the whole range of performance and are suitable for differentiating both in high and low proficiency levels. For perseveration ratio, this is true only for participants aged 60+. For younger participants, distribution of test scores is right skewed (skewness > 2, Kolmogoroff-Smirnoff $p < .3$). This indicates that performance shows ceiling effects and the FPT perseveration score is not suitable for differentiation within the high range of performance in participants aged younger than 60.

Our strategy score reflects the proportions of items generated by the rotational strategy. Of course, there are other strategies possible, for example the enumerative strategy where a line is added to the design systematically (e.g., Regard et al., 1982). Alternative strategies occurred in less than 5% of participants and thus only rarely in our sample. In only three participants (1.1%) did the number of items generated by an alternative strategy outnumber the number of items generated by the rotational strategy. However, none of these participants scored below the 25th percentile in the provided strategy score. Therefore no other measures of strategic functioning were included.

In only 3.6% of participants did rule breakings occur: 2.9% ($n = 8$) of participants made one error and 0.7% ($n = 2$) two errors. Both these were male, aged 60+ (respectively 62 and 76 years), and had received fewer than 13 years of education. There were no particularities regarding the medical or psychiatric history. MMSE scores were 30 and 27 respectively, which means that the older

Table 3 Percentiles for productivity (number of unique designs)

	Productivity (number of unique designs)					
	Age 19–39		Age 40–59		Age 60–80	
Percentile	Educ 1–13	Educ 13+	Educ 1–13	Educ 13+	Educ 1–13	Educ 13+
2	11	23	11	16	10	20
5	16	25	14	18	12	22
10	18	29	22	24	15	
16	20	32	24	26	16	23
20	22	33	25	27	17	24
25	23	34	26	29	19	26
30	24	35	27	31	20	27
40	25	37	28	35	23	28
50	28	41	31	36	24	29
60	29	44	32	37	26	31
70	31	47	33	40	30	32
75	33	49	35	42	31	35
80	35	51	37	45	32	42
84	37	52	38	47	33	46
90	39	57	40	49	34	48
95	46	58	42	55	37	51
98	47	60	44	56	41	53

Table 4 Percentiles for flexibility (percentage of per severations)

Flexibility (percentage of perseverations)						
	Age 19–39		Age 40–59		Age 60–80	
Percentile	Educ 1–13 (%)	Educ 13+ (%)	Educ 1–13 (%)	Educ 13+ (%)	Educ 1–13 (%)	Educ 13+ (%)
2	21.5	19.9	28.6	35.3	44.5	22.2
5	20.2	16.7	23.0	28.9	41.9	17.8
10	18.8	11.9	17.9	17.6	32.0	14.2
16	13.2	7.7	7.7	17.1	22.9	13.4
20	12.1	7.0	11.2	14.8	22.0	12.4
25	11.4	6.4	9.0	13.6	19.9	11.9
30	7.0	6.0	8.4	11.5	16.7	10.9
40	5.2	4.0	6.3	8.1	12.3	8.4
50	4.0	2.7	4.1	7.4	9.3	6.7
60	0.0	2.0	3.3	5.2	8.0	5.6
70		0.0	2.9	4.2	5.2	4.2
75			2.5	3.1	4.1	3.8
80			0.0	2.8	2.8	3.1
84				2.5	0.0	1.9
90				0.0		0.0

Table 5 Percentiles for the Strategy score (percentage of rotated and/or mirrored items)

Strategy score (percentage of rotated and/or mirrored items)						
	Age 19–39		Age 40–59		Age 60–80	
Percentiles	Educ 1–13 (%)	Educ 13+ (%)	Educ 1–13 (%)	Educ 13+ (%)	Educ 1–13 (%)	Educ 13+ (%)
2		6.0		12.0		
5	0	12.0	0	14.5	0	0
10	4.5	24.0	10.1	18.5	2.7	11.6
16	6.5	31.5	14.9	26.0	5.1	14.5
20	12.0	35.5	17.3	29.4	5.8	23.5
25	16.7	39.4	25.6	32.4	7.0	29.5
30	20.5	40.0	27.9	34.3	8.3	34.9
40	23.1	43.0	32.1	37.0	14.3	39.2
50	30.4	48.1	35.4	47.5	17.8	40.3
60	35.9	53.2	38.5	49.0	20.2	41.9
70	39.2	59.2	46.2	55.1	28.0	46.5
75	42.4	61.3	48.1	56.8	30.7	49.3
80	45.1	63.3	51.5	59.3	36.0	54.3
84	45.8	65.8	55.2	61.6	37.1	55.4
90	48.1	67.6	60.9	67.1	47.7	59.4
95	54.1	70.6	62.3	68.9	53.1	64.7
98	59.5	71.3	63.5	70.9	61.5	65.4

participant only scarcely met our inclusion criterion. Both participants scored below the 16th percentile in the strategy score. For the older patient, the same is true for productivity and percentage of perseverations. Therefore we suggest the utilization of a simple cut-off score (Table 6). Whereas one error might be regarded as a borderline value, two or more rule breakings should be classified as impaired.

Reliability

Inter-rater reliability. To assess inter-rater reliability, Intraclass Correlation Coefficients (ICCs) of rater agreement were calculated. The ICC method considers raters as random effects, allowing the generalization from a subset of raters to a universe of raters (Shrout & Fleiss, 1979). Results were $r_{ICC} = .989$ for number of unique designs, $r_{ICC} = .988$ for percentage perseverations, $r_{ICC} = .991$ for the strategy score, and $r_{ICC} = .972$ for the number of rule breakings.

Test–retest reliability. None of the 34 participants made any rule-breaking errors at either time of testing, so this test score was excluded from analysis. The average performance of participants improved upon retesting for both number of unique designs and percentage of perseverations. In contrast, no significant differences were observed for time 1 to time 2 for strategic performance. Moreover, ICCs were used for the estimation of temporal stability (see Ross et al., 2003). ICCs are preferable to Pearson product–moment correlations when observations are not independent, as the latter might overestimate the relationship between variables (Shrout & Fleiss, 1979). Coefficients for unique designs and percentage of strategic items were satisfactory and highly significant. However, no significant correlation was found for the perseveration ratio. For details, see Table 7.

Table 6 Cut-off score for non-perseverative errors (rule-breakings)

Percentile	Number of errors	Suggested classification
<1	2	Impaired
4	1	Borderline
5–100	0	Unimpaired

Table 7 Mean and standard deviations for FPT scores for test and retest as well as comparisons and correlations between time points

	Mean (SD)			
	Time 1	Time 2	Paired *t*-test	ICC
Unique designs	39.13 (9.38)	44.30 (8.93)	$T = -4.436$ $p = .000$**	ICC = 0.838 $p = .000$**
% Perseverations	5.57 (6.02)	3.16 (4.51)	$T = 2.053$ $p = .049$*	ICC = 0.353 $p = .116$
% Strategies	47.56 (18.43)	52.00 (16.80)	$T = -1.635$ $p = .112$	ICC = 0.717 $p = .000$**

* = $p < .05$; ** = $p < .01$

Table 8 Inter correlations and relation to other neuropsychological measures for the test-retest subgroup of participants

	FPT: productivity	FPT: % persever	FPT: % strategies
FPT: % persever	−0.150		
FPT: % strategies	0.483**	−0.350*	
IQ (MWT-B)	0.472*	−0.133	0.049
LBS (German FAS)	0.090	−0.150	−0.049
TMT-A	−0.277	0.126	−0.050
TMT-B	−0.406*	0.252	−0.190
Digit Span Forward	−0.127	−0.296	−0.010
Digit Span Backwards	0.239	−0.524**	0.174

* = $p < .05$; ** = $p < .01$

Validity

In the test–retest subgroup of participants, productivity was significantly correlated with MWT-B IQ and TMT-B, measuring psychomotor speed and the ability to shift. Percentage of perseverations was inversely related to verbal working memory measured by Digit Span Backwards. Strategic behavior did not correlate significantly with any of the neuropsychological measures. Verbal fluency (German FAS), verbal short-term memory (Digit Span Forward), and psychomotor speed (TMT-A) were not correlated with any of the FPT scores. Table 8 presents detailed information about correlations with other neuropsychological measures, as well as intracorrelations between FPT scores in this subgroup. As test results might not be independent from general intelligence, partial correlations were calculated controlling for the MWT-B IQ. Significant correlations remained the same with the exception that FPT productivity then was additionally correlated with psychomotor speed (TMT-A; $r = -.345, p = .049$).

Principal factor analysis (PFA) with oblique rotation was used to establish factorial validity. Four factors with eigenvalues greater than 1.00 accounted for 57.43% of the variance. Correlations of variables with factors suggest that the first component reflects processes of productivity and speed, with TMT-B ($r = .902$), TMT-A ($r = .749$), and FPT productivity ($r = .635$) showing the greatest correlations. The second component reflects verbal working memory, with Digit Span Forward ($r = .812$) and Digit Span Backward ($r = .647$) loading high. The third factor consists of executive processes that are not dominated by a speed component, and contains the FPT strategy score ($r = .611$) as well as the FPT perseveration ratio ($r = .477$). Correlation of FPT productivity with this third factor is $r = .481$ and thus only somewhat lower than with the first factor. The final component reflects verbal components with verbal IQ (MWT-B; $r = .811$) and verbal productivity (German FAS; $r = .383$).

DISCUSSION

We have presented normative data for the Five-point Test and examined the psychometric properties of its qualitative and quantitative scores. For determination of normative groups, age, sex, and level of education were taken into consideration. Whereas gender had no influence on performance, performance increased with decreasing age and increasing education. Similar results have been reported for other measures of executive functions (e.g., Gladsjo et al., 1999; Tombaugh, 2004; van der Elst, van Boxtel, van Breukelen, & Jolles, 2006; Zalonis et al., 2007). Therefore differentiated norms were provided. In the individual subgroups, sample sizes ranged from 21 to 86. Therefore, sample sizes did not always meet modern-day test development standards. However, our sample is considerably large relative to other neuropsychological test samples, especially for the assessment of executive functions (Lezak et al., 2004). In addition, strength of our sample lies in the high number of relatively young, high-performance participants.

Most authors have not described gender differences for figural fluency measures, therefore our results are consistent with previous data (e.g., Harter et al., 1999; Regard et al., 1982; Ross et al., 2003; Ruff, Light, & Evans, 1987). Regarding age, decreasing performance with increasing age has been described regularly for figural fluency measures (e.g., Lezak et al., 2004), so our results correspond to this. Recent functional-imaging studies document region-specific changes in prefrontal function with age. Especially the right dorsolateral prefrontal cortex and anterior regions show a decline in functionality (for a review, see Rajah & D'Esposito, 2005). This might be held responsible for the decrease in executive and metacognitive functions frequently observed in elderly participants. Previous research yielded inconsistent results on the question whether intelligence and/or level of education have an impact on performance in figural fluency tasks (Evans, Ruff, & Gualtieri, 1985; Lee et al., 1997; Lezak et al., 2004; Ruff et al., 1987). In our study, higher education was strongly linked with better performance. This might be due to more experience and routine in execution of cognitive tasks in participants with higher-level education. Another contributing factor might be that cognitive processes constituting fluid intelligence seem to be highly linked to functionality of the dorsolateral prefrontal cortex (e.g., Duncan et al., 2000) which plays a central role for executive functions (e.g., Funahashi, 2006).

Productivity reflects the quantitative performance of subjects in the FPT. In this study adequate normative data for this parameter were provided for the first time. It proved itself to be suitable for differentiated assessment both in high and low ranges of performance. In the opinion of Abwender et al. (2001), the over-reliance on quantitative assessment has led to a lack of certainty regarding the nature of fluency measures and interpretation of poor performance. We offer three qualitative or process-oriented test scores with good psychometric properties to overcome this deficit. As rule breakings hardly ever occurred, only a simple cut-off score was defined. As predicted, the strategy score was related positively to unique design output, whereas perseveration ratio was correlated negatively with productivity. Therefore it could be hypothesized that our process-oriented test scores offer important information for interpretation of a patient's impaired quantitative test result.

The FPT perseveration ratio cannot differentiate within the upper range of performance due to ceiling effects in patients younger than 60. However, the FPT perseveration index is suitable for clinical application aiming to detect impairments and potentially offer an explanation for reduced productivity. Similar to the perseveration index of the RFFT (Ross et al., 2003), temporal stability was noticeably lower compared to the other test scores. In our study, ICC for assessment of test–retest reliability did not reach statistical significance. This cannot be attributed to restricted range during either or both administrations.

As predicted, the strategy score was related positively to unique design output and negatively to perseveration ratio. This supports the assumption of Lezak (1995) who emphasized that an effective strategy can improve both qualitative and quantitative performance in neuropsychological tests. Strategic behavior is part of metacognitive functions (e.g., Shimamura, 2000). Metacognition refers to monitoring and control of one's own cognitive processes, thereby influencing basic cognitive processes. The ability to quickly initiate and adequately utilize an effective strategy might enhance performance on a multitude of test procedures (e.g., Alexander, Stuss, & Fansabedian, 2003). Thus, strategic performance might be an important element regarding the interpretation of impaired test scores.

Our strategy score seems to be suitable for this due to its good psychometric properties. Furthermore, it is able to differentiate within both ends of the performance range. However, it is not yet clear whether good strategic abilities in the FPT are associated with good strategic abilities in other cognitive tests. This should be established in future studies. Moreover, more research on the utility of strategy scores in general would be desirable.

Ross et al. (2003) reported a variety of additional strategic measures (see Regard et al., 1982). In their study the total number of strategic clusters exhibited the strongest correlation with the unique design score. We limited our strategy score to a very basic assessment of strategy use with the aim of generating an economic and reliable measurement. As even our simple strategy score correlates with productivity and flexibility in the predicted way, it appears to be valid. However, this should be established in future studies using clinical samples.

Inter-rater reliability was over 0.95 for each test score. Judged by Cicchetti and Sparrow's (1981) criteria for evaluating clinical significance, inter-rater reliability coefficients are excellent. Inter-rater reliability was noticeable higher than in other figural fluency measures (e.g., Berning, Weed, & Aloia, 1998; Carter, Shore, Harnadek, & Kubu, 1998; Harter et al., 1999; Ross et al., 2003). This might be due to the more structured and standardized test material compared to the Design Fluency Test due to the less-complex test material compared to the RFFT. One might argue that the identification of rotational strategies in the FPT requires considerably fewer vigilance and spatial skills than in the RFFT, where in Parts IV and V the arrangements of the five dots are not concentric. Furthermore, for the strategy score only rotated and/or mirrored items were included, which leads to very simple scoring demands compared to other measures of strategic behavior (e.g., Harter et al., 1999; Ross et al., 2003). In addition we defined only four test scores, further reducing complexity of scoring. Another contributing factor might be the fact that our raters were experienced in administration and scoring of neuropsychological tests and had received intensive training. This constitutes a difference to Berning et al. (1998), who employed college freshmen after very little preparation. In conclusion, the present data suggest that all FPT indices possess excellent interscorer agreement, which is valid even for individuals with less than graduate-level training after sufficient preparation.

Consistent with previous research, the average performance of participants improved on testing (e.g., Basso, Bernstein, & Lang, 1999; Harter et al., 1999; Ross et al., 2003; Ruff, 1988). Similar results have been reported from other measures assessing executive functions (e.g., Greve et al., 2002). Objections against calculating test–retest reliability from the first and the second test examination can be raised, as temporal stability cannot be assessed here without practice effects (e.g., Beglinger et al., 2005). Besides, the time interval of only about 1 month between test sessions was relatively small (see for example Basso et al., 1999, or Ruff, Light, & Evans, 1987). This might enhance practice effects (e.g., Demakis, 1999). These are, however, inevitable in a multitude of cognitive test procedures, which results in well-known difficulties in the development of adequate parallel test versions (see Lezak et al., 2004). To overcome the named disadvantages further research is recommended for a more appropriate establishment of test–retest reliability. Despite these handicaps, temporal stability was satisfactory for productivity and strategic behavior.

Convergent and divergent validity were assessed through intracorrelations of FPT scores as well as correlations with other neuropsychological measures. Intraclass correlations were modest, suggesting that none of the three main test dimensions is redundant. Similar to Ruff et al. (1987) for the RFFT, we found that figural fluency did not correlate with verbal fluency (see Abwender et al., 2001).

With regard to other neuropsychological measures, correlations were maximal moderate. This indicates that the FPT contributes relevantly to other measures in a neuropsychological test battery. This was conformed by a factorial analysis, which revealed that the FPT is not only a measure of productivity and speed but also an independent measure of cognitive functioning.

In conclusion, differentiated norms were provided for the FPT for the first time. Besides the quantitative dimension productivity, the more process-oriented dimensions of perseverative and strategic behavior as well as rule breakings are included in our norms. Analysis of reliability and validity were promising. However, to establish the clinical validation as most important attribute of a clinical instrument, future studies with patient populations suffering from known cerebral impairments are recommended.

ACKNOWLEDGEMENTS

The authors would like to thank Lea Kaup and Nele Schmidt for serving as additional raters for this study.

REFERENCES

Abwender, D. A., Swan, J. G., Bowerman, J. T., & Connolly, S. W. (2001). Qualitative aspects of verbal fluency output: Review and comparison of several scoring methods. *Assessment*, *8*, 323–336.

Alexander, M. P., Stuss, D. T., & Fansabedian, N. (2003). California Verbal Learning Test: Performance by patients with focal frontal and non-frontal lesions. *Brain*, *126*, 1493–1503.

Baldo, J. V., Shimamura, A. P., Delis, D. C., Kramer, J., & Kaplan, E. (2001). Verbal and design fluency in patients with frontal lobe lesions. *Journal of the International Neuropsychological Society*, *7*, 586–596.

Basso, J. V., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across 12 months. *The Clinical Neuropsychologist*, *13*, 283–292.

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., et al. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology*, *20*, 517–529.

Benton, A. L., & Hamsher, K. D. (1989). *Multilingual Aphasia Examination*. Iowa City: AJA Associates.

Berning, L. C., Weed, N. C., & Aloia, M. S. (1998). Inter-rater reliability of the Ruff Figural Fluency Test. *Assessment*, *5*, 181–186.

Carter, S. L., Shore, D., Harnadek, M. C. S., & Kubu, C. S. (1998). Normative data and inter-rater reliability of the Design Fluency Test. *The Clinical Neuropsychologist*, *12*, 531–534.

Cicchetti, D. V., & Sparrows, S. S. (1981). Developing criteria for establishing the inter-rater reliability of specific items in a given inventory: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*, 127–137.

Cone, J. D. (1999). Observational assessment: Measurement development and research issues. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (pp. 183–223). New York: Wiley.

Demakis, G. J. (1999). Serial malingering on verbal and nonverbal fluency and memory measures: An analogue investigation. *Archives of Clinical Neuropsychology*, *14*, 401–410.

Duncan, J., Owen, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., et al. (2000). A neural basis for general intelligence. *Science*, *289*, 457–460.

Evans, R. W., Ruff, R. M., & Gualtieri, C. T. (1985). Verbal and figural fluency in bright children. *Perceptual and Motor Skills*, *61*, 699–709.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-Mental State (a practical method for grading the state of patients for the clinician). *Journal of Psychiatric Research*, *12*, 189–198.

Funahashi, S. (2006). Prefrontal cortex and working memory processes. *Neuroscience*, *139*, 251–261.

Gladsjo, J. A., Schuman, C. C., Evans, J. D., Peavy, G. M., Miller, S. W., & Heaton, R. K. (1999). Norms for letter and category fluency: Demographic corrections for age, education, and ethnicity. *Assessment*, *6*, 147–178.

Gottwald, B., Wilde, B., Mihajlovic, Z., & Mehdorn, H. M. (2004). Evidence for distinct cognitive deficits after focal cerebellar lesions. *Journal of Neurology, Neurosurgery and Psychiatry*, *75*, 1524–1531.

Greve, K. W., Love, J. M., Sherwin, W., Mathias, C. W., Houston, R. J., & Brennan, A. (2002). Temporal stability of the Wisconsin Card Sorting Test in a chronic traumatic brain injury sample. *Assessment*, 9, 271–277.

Harter, S. L., Hart, C. C., & Harter, G. W. (1999). Expanded scoring criteria for the Design Fluency Test: Reliability and validity in neuropsychological and college samples. *Archives of Clinical Neuropsychology*, *14*, 419–432.

Härting, C., Markowitsch, H. J., Neufeld, H., Calabrese, P., Deisinger, K., & Kessler, J. (Eds.). (2000). *Wechsler Gedächtnistest - Revidierte Fassung* (*WMS-R*). Bern: Huber.

Jason, G. W. (1985). Gesture fluency after focal cortical lesions. *Neuropsychologia*, *23*, 463–481.

Jones-Gotman, M., & Milner, B. (1977). Design fluency: The invention of nonsense drawings after focal cortical lesions. *Neuropsychologia*, *15*, 653–664.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B. T. Bryant (Eds.), *Clinical neuropsychology and brain function* (pp. 129–167). Washington, DC: American Psychological Association.

Lee, G. P., Strauss, E., Loring, D. W., McCloskey, L., Haworth, J. M., & Lehman, R. A. W. (1997). Sensitivity of figural fluency on the Five-point Test to focal neurological dysfunction. *The Clinical Neuropsychologist*, *11*, 59–68.

Lehrl, S. (1999). *Mehrfachwahl-Wortschatztest MWT-B*. Balingen: Spitta.

Lezak, M. D. (1995). *Neuropsychological assessment*. New York: Oxford University Press.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment*. New York: Oxford University Press.

Rajah, M. N., & D'Esposito, M. (2005). Region-specific changes in prefrontal function with age: A review of PET and fMRI studies on working and episodic memory. *Brain*, *128*, 1964–1983.

Regard, M., Strauss, E., & Knapp, P. (1982). Children's production on verbal and non-verbal fluency tasks. *Perceptual and Motor Skills*, *55*, 839–844.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Therapy and clinical interpretation*. Tucson, AZ: Neuropsychological Press.

Ross, T. P., Foard, E. L., Hiott, F. B., & Vincent, A. (2003). The reliability of production strategy scores for the Ruff Figural Fluency Test. *Archives of Clinical Neuropsychology*, *18*, 879–891.

Ruff, R. M. (1988). *Ruff Figural Fluency Test professional manual*. Odessa, FL: Psychological Assessment Resources Inc.

Ruff, R. M., Allen, C. C., Farrow, C. E., Niemann, H., & Wylie, T. (1994). Figural fluency: Differential impairment in patients with left versus right frontal lobe lesions. *Archives of Clinical Neuropsychology*, *9*, 41–55.

Ruff, R. M., Evans, R., & Marshall, L. F. (1986). Impaired verbal and figural fluency after head injury. *Archives of Clinical Neuropsychology*, *1*, 87–101.

Ruff, R. M., Light, R. H., & Evans, R. W. (1987). The Ruff Figural Fluency Test: A normative study with adults. *Developmental Neuropsychology*, *3*, 37–51.

Salloway, S. P. (1994). Diagnosis and treatment of patients with "frontal lobe" syndrome. *Journal of Neuropsychiatry and Clinical Neuroscience*, *6*, 388–398.

Schnider, A. (1997). *Verhaltensneurologie: Die neurologische Seite der Neuropsychologie*. Stuttgart: Thieme.

Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition*, *9*, 313–323.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests*. New York: Oxford University Press.

Stuss, D. T., Alexander, M. P., Hamer, L., Palumbo, C., Dempster, R., Binns, M., et al. (1998). The effects of focal anterior and posterior brain lesions on verbal fluency. *Journal of the International Neuropsychological Society*, *4*, 265–278.

Tombaugh, T. N. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, *19*, 203–214.

Tucha, O. W., Smely, C. W., & Lange, K. W. (1999). Verbal and figural fluency in patients with mass lesions of the left or right frontal lobes. *Journal of Clinical and Experimental Neuropsychology*, *21*, 229–236.

Van der Elst, W., van Boxtel, M. P. J., van Breukelen, G. J. P., & Jolles, J. (2006). The Stroop Color-Word Test: Influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment*, *13*, 62–79.

Williams-Gray, C. H., Foltynie, T., Brayne, C. E. G., Robbins, T. W., & Barker, R. A. (2007). Evolution of cognitive dysfunction in an incident Parkinson's disease cohort. *Brain*, *130*, 1787–1798.

Zalonis, I., Kararizou, E., Triantafyllou, N. I., Kapaki, E., Papageorgiou, S., Sgouropoulos, P., et al. (2007). A normative study of the Trail Making Test A and B in Greek adults. *The Clinical Neuropsychologist*, *22*, 842–850.