The Clinical Neuropsychologist, 23: 501–509, 2009
http://www.psypress.com/tcn
ISSN: 1385-4046 print/1744-4144 online
DOI: 10.1080/13854040802279675

Psychology Press
Taylor & Francis Group

RELIABILITY OF THE FIVE-POINT TEST

Alberto Luis Fernandez[1,2], Matías A. Moroni[3], Juan Manuel Carranza[3], Nicolás Fabbro[3], and Brian K. Lebowitz[4,5]

[1]*Department of Neuropsychology, Catholic University of Córdoba, Argentina,* [2]*Department of Psychometric Techniques, National University of Córdoba, Argentina,* [3]*National University of Córdoba, Córdoba, Argentina,* [4]*Department of Neurology, Stony Brook University Medical Center, Stony Brook, NY, USA, and* [5]*Department of Psychiatry, Harvard Medical School, Boston, MA, USA*

The purpose of this study was to assess the internal consistency and stability of the Five-Point Test (FPT), developed by Regard, Strauss, and Knapp (1982). A test of non-verbal fluency, the FPT is an executive measure that may be particularly useful in the evaluation of individuals with presumed frontal lobe damage. In the internal consistency study, 209 healthy participants were administered the FPT. A split-half analysis revealed a correlation of .80 for unique designs, and .48 for perseverative errors. In the stability study, 142 healthy participants were administered the FPT on two occasions with a mean interval of 37.8 days. Across the two administration periods, a test–retest correlation of .78 for unique designs and .51 for perseverative errors was found. When the mean performances were compared across administration periods, significant differences were found for unique designs, but not for percentage of perseverative errors. Taken together, the results of the two studies suggest that the internal consistency and stability coefficients of the FPT are acceptable for unique designs but low for the percentage of perseverative errors.

Keywords: Five-Point Test; Reliability; Split-half reliability; Test–retest reliability; Figural fluency.

INTRODUCTION

The Five-Point Test (FPT) is a non-verbal measure of executive functioning that was designed to assess design or figural fluency. Specifically, the test measures the ability of an individual to produce unique geometric designs or figures within a given period of time. Developed by Regard, Strauss, and Knapp (1982), the test consists of a sheet of paper with 40 five-dot matrices. Subjects are asked to produce as many different figures as possible by connecting the dots within each rectangle. The scoring system involves the amount of unique designs produced within 3 minutes, as well as the percentage of perseverative errors [(perseverative errors/total unique designs) $\times$ 100]. In this way, the test provides information regarding an individual's figural fluency capacity, as well as information related to perseverative behavior.

Address correspondence to: Alberto Luis Fernandez, Cortex – Neuroterapias, Chacabuco 1296, 1° D, 5000, Córdoba, Argentina. E-mail: neurorehab@onenet.com.ar

Accepted for publication: June 17, 2008. First published online: August 1, 2008.

The FPT has proved to be a sensitive measure in the assessment of frontal lobe functioning. For example, Lee et al. (1997) found that the FPT was able to discriminate patients with frontal lobe dysfunction from patients with psychiatric disorders and non-frontal neurological disease.

To date, only one study has been published on the reliability of the FPT. Using a 10-minute extended-time version of the FPT, Santa-María, Martin, Morrow, and Gouvier (2001) examined the minute-by-minute production of unique designs (UD) and percentage of perseverative errors (PPE). They found that UD decreased significantly each minute until the ninth minute. The PPE increased significantly each minute until the eighth minute. As a consequence, these researchers affirmed that extending the time of the test would increase its sensitivity.

No study has been published on the reliability of the test addressing the issues of internal consistency and stability. Internal consistency refers to the degree of within-scale item correlation (i.e., the degree to which the items that constitute the test measure the same dimension). Stability refers to the test–retest reliability of the instrument—i.e., it is "... an index of the extent to which scores are likely to fluctuate as a result of time sampling error" (Urbina, 2004, p. 125). The current study was designed to assess these two critical dimensions of reliability on the FPT: internal consistency and stability.

METHOD

Study 1. Internal consistency

Participants. A total of 212 participants were included in this study. The sample was comprised of people recruited from several sources: psychology students, people attending a teaching program for older adults, and relatives of individuals from the former groups. Data were gathered from July 2003 to January 2005. A total of 65% of the participants were female. The average age of the sample was 47.79 ± 21.76, and the average level of education was 13.10 ± 4.55 years. Most participants were right handed (86%), with the remaining participants being classified as either ambidextrous (11%) or left-handed (3%). Individuals were excluded from participation if they had any history of neurological disease, psychiatric diagnosis, diabetes, thyroid disease, Chagas disease (a frequent infectious disease in some areas in Argentina producing heart dysfunction), head trauma, stroke, heart attack, non-controlled high blood pressure, coma, drug intake, alcoholism, sleep disorders, learning disabilities and chronic headaches.

Procedure. Administration of the FPT was performed by 10 psychology students or graduate psychologists. To insure standardized administration and scoring, all raters underwent training during a 1-month period. After reviewing the test manual, raters practiced administering the instrument, first to other raters and then to individuals who matched the sample used in the final studies. Data gathered during the training sessions were not included in the final analyses. Test administration was performed in different settings according to the availability (classrooms, laboratories, etc.) but under the same circumstances: a quiet, well-lit room containing only the rater and study participant.

At the beginning of each session, participants were given a pencil and a copy of the 40 five-dot matrices presented on an 8½ in. × 11 in. sheet of paper. Each matrix was identical, with four dots arranged in a square pattern around a central dot (as in a dice cube). The matrices were arranged in eight rows and five columns. The first two matrices were used as samples drawn by the examiner. Participants were asked to produce as many different figures as possible by connecting the dots within each matrix. They were also informed that they did not need to connect all dots in order to make a valid design (connecting fewer than five dots was acceptable as a design), and that only straight lines were appropriate. The rater then used the first two matrices to demonstrate the task. After the task had been demonstrated, the test started. If the participant repeated a design, the rater stated: ''Remember you must not repeat any design.'' This reminder was given only after the first repetition.

If the participant completed the first test sheet, an additional sheet was provided. This additional sheet was placed next to the original sheet on the opposite side of the dominant hand (i.e., to the left side if the participant was right-handed), allowing for observation of designs made on the first sheet. The participants were instructed to continue producing designs on the new sheet but to avoid repeating designs produced on the first sheet. The time limit was 3 minutes.

The test was scored by counting the total number of UD and the number of repeated designs (perseverative errors) produced by the participant. Because the number of UD and perseverative errors are not independent (a greater number of designs results in greater likelihood of repetitions), the influence of individual differences in overall design production must be controlled. An index score, the PPE, was calculated so that participants could be rated on perseverative behavior independent of total designs produced. The following formula was used: (perseverative errors/total unique designs) × 100.

The internal consistency was evaluated with the split-half procedure. However, as some authors caution (Anastasi & Urbina, 1998; Martinez-Arias, 1995), the standard split-half procedure is not appropriate for timed tests. In a typical split-half reliability design, the test items are divided into two halves (first half vs second half, or even vs odd items). When applied to timed tests the reliability coefficient is adversely influenced by practice, fatigue, and other factors not related to the construct being assessed.

To address the limitations of performing split-half analyses on timed tests, Anastasi and Urbina (1998) proposed a different approach to reliability assessment. The total time of the test was divided into quarters. In the case of the FPT, four time periods of 45 seconds each were constructed. The first quarter was then added to the fourth quarter, comprising the first half (Part A). The second quarter was then added to the third quarter, making the second half (Part B). A Spearman-Brown correlation for Parts A and B was then calculated to determine the internal consistency coefficient. The above procedure was employed to assess the internal consistence of both scores, UD and PPE.

In order to divide the participants' performance into quarters without interfering with them, the raters had an identical test sheet and placed marks on the rectangle the participant had just completed at the end of every time quarter. These marks were transferred to the same rectangles on the sheets of the participants once

they had finished the test to allow the counting of designs and repetitions, and calculation of the scores.

Study 2. Stability

Participants. A total of 142 participants were included in this study. The inclusion criteria and recruitment procedure were the same as Study 1. A total of 53% of the participants were female. In comparison to the sample in the first study, participants were younger and better educated. The average age of the sample was 30.5 ± 14.5, and the average level of education was 14.7 ± 2.9 years. Most participants were right-handed (82%), with the remaining participants being classified as either ambidextrous (11%) or left-handed (7%). Five test-takers, who also participated in Study 1, collected the data. These data were collected in a 5-month period.

Procedure. The administration procedure was identical to Study 1 with the exception of the quarter marking system, which was excluded here. Test–retest interval had a range of 20 to 79 days, with a mean of 37.7 ± 14.3 days.

Pearson correlation coefficients and repeated measures *t*-tests were employed in the statistical analysis. The analyses of the influence of demographic variables on the scores of the test were not performed in this study due to the very narrow range of age and education of the sample.

RESULTS

Study 1. Internal consistency

The reliability coefficients obtained were $r = .80$ ($p < .01$) for UD, and $r = .48$ ($p < .01$) for PPE. Table 1 shows the mean and standard deviation of UD and PPE for the total time of the test. Table 2 illustrates the mean and standard deviation for each time quarter. Pearson correlation with age and education were modest for both scores. These are shown in Table 3. Although an ANOVA showed no significant differences between males and females for both scores there was a trend towards significance, with men producing more UD and less PPE than women: UD $= F(1, 210) = 3.31$; $p = .07$; and PPE $= (1, 210) = 2.99$; $p = .09$. Table 4 exhibits the mean scores of UD and PPE for males and females.

Table 1 Mean and standard deviation of unique designs and percentage of perseverative errors: Study 1

	Mean	*SD*
Unique designs	26.63	9.71
% Perseverative errors	9	9.35

Study 2: Stability

The stability coefficient was $r = .78$ ($p < .01$) for UD, while it was $r = .51$ ($p < .01$) for PPE. Means and standard deviations for both scores are presented in Table 5.

In order to evaluate the influence of learning in both administrations a repeated measures t-test was performed. Differences between test and retest performance were significant for UD, $t(142) = -12.55$, $p < .01$, but not for PPE $t(142) = 1.53$, $p < .13$.

The significant differences in the mean scores of UD between test and retest demonstrate that the FPT is susceptible to learning effects. Therefore, a statistical index that reflects the magnitude of expected change across evaluations is necessary.

Table 2 Mean and standard deviation of unique designs and percentage of perseverative errors by time quarter

Quarter #	Unique designs	% Perseverative errors
Q1	9.71 (±4.36)	3.33 (±7.3)
Q2	6.6 (±2.93)	10.46 (±17.5)
Q3	5.52 (±2.65)	4.09 (±6.08)
Q4	4.72 (±2.58)	19.55 (±34.65)

Table 3 Correlation of test scores with demographic variables

Variable	Unique designs	% Perseverative errors
Age	−.53 ($p < .000$)	.30 ($p < .000$)
Education	.55 ($p < .000$)	−.08 ($p < .15$)

Table 4 Mean and standard deviation of males and females for both scores

Score	Males	Females
Unique designs	28.28 (±10.76)	25.75 (±9.02)
% Perseverative errors	7.48 (±7.3)	9.8 (±10.22)

Table 5 Mean and standard deviation for both scores at test and retest

Score	Test	Retest
Unique designs	31.8 (±9)	38.5 (±10)
% Perseverative errors	5.9 (±7)	5.1 (±5.9)

Table 6 Standard error of prediction (SEP) and 90% confidence interval for both scores of the Five-Point Test

	R_{Y1Y2}	SEP	90% CI
Unique designs	.78	6.25	±10
Percentage of perseverative errors	.51	5.08	±8

The confidence interval values are rounded to the nearest whole digit.

The standard error of measurement is the usual index employed to calculate this change. However, Charter (1996) has pointed out that this method is inappropriate in the case of a measure susceptible to the effects of practice, because the standard error of measurement assumes that error between the assessments is uncorrelated. Thus, the standard error of measurement is more appropriate for those situations where practice effects are irrelevant, for instance, attitude or interests scales. Charter suggested a different index for tests where the measurement error is likely correlated (Charter, 1996): the standard error of prediction (SEP). The SEP is calculated as follows:

$$\mathrm{SEP} = \mathrm{SD}_{Y2}\sqrt{\left(1 - {r_{Y1Y2}}^2\right)}$$

In this equation, SD_{Y2} is the standard deviation of scores in the retest, and r_{Y1Y2} is the correlation between test scores across the assessment intervals. A 90% confidence interval was obtained by multiplying the SEP by +1.64. Table 6 exhibits the SEP as well as the 90% confidence intervals for each test score.

DISCUSSION

The current study was designed to assess two critical dimensions of reliability on the FPT: internal consistency and stability. Several levels of reliability within the scores of FPT were revealed, ranging from unacceptable to good reliability. Cicchetti (1994) elaborated a scale to classify the levels of clinical significance of reliability coefficients. According to Cicchetti, a coefficient below.70 is unacceptable; when it is between .70 and .79 the level is fair; when it is between .80 and .89 it is good; when it is 90 or above, the level of clinical significance is excellent.

Using the above classification system the internal consistency of FPT is good for UD ($r = .80$), but unacceptable for PPE ($r = .48$). Regarding stability, the reliability coefficient is fair for UD ($r = .78$), and unacceptable for PPE ($r = .51$). Thus, only the production of UD seems to have acceptable levels of reliability. Because PPE varied so much within and across administrations, this score should be interpreted very cautiously; i.e., changes in this score observed in the successive performance of a given patient should be interpreted with the assistance of additional data to decide if they truly mean an improvement or a worsening of his/her figural fluency ability. However, it is important to note that low reliability is not rare among executive tests (Basso, Bornstein, & Lang, 1999; Burguess, 1997). Several possibilities for low reliability have been proposed. Among them, the loss of

a novelty effect (Lezak, 1995), as well as the participation of procedural memory in the test-taking strategies (Basso et al., 1999), have been mentioned.

Analyzing the low internal consistency coefficient obtained for PPE it might be hypothesized that the nature of repetitions changes during the course of the task. Along the test the UD production decreases while the PPE percentage increases. Thus, at the beginning, the test seems to measure the ability of the participants to avoid repetitions while they have enough ideas to create new designs (figural fluency). Towards the end of the FPT, however, the instrument seems to measure the ability of the participants to avoid the production of repeated designs once their figural fluency decreases. Following Santa-Maria et al. (2001) this tendency seems to extend until the eighth minute. Nonetheless, performance of participants seems to be quite different as regards this ability. A more careful inspection of the data within each quarter allows one to see a considerable increase of PPE in quarters 2 and 4. However, in both quarters the standard deviation is also significantly increased. This effect might be interpreted as the particular behavior of a sample subgroup. It seems that for some individuals it is more difficult than for others to avoid the production of repeated designs.

Taking into account the low internal consistency coefficient found in this study, the suggestion of Santa-Maria et al. (2001) of increasing the extension of the task in order to increase its sensitivity gains further support. The issue of time may be particularly important in the case of patients with a mild executive dysfunction who may not exhibit a significant PPE in the 3-minute version. On longer versions, however, difficulties may be revealed. This does not seem to be the case of patients with more severe brain injury. In the Lee et al. (1997) study, the 3-minute version was sensitive enough to discriminate patients with brain damage from psychiatric patients without brain lesions. The patients in the brain damage group of this study included neurological conditions such as cerebrovascular accidents, intractable epilepsy, tumors, closed-head injuries, and neurodegenerative disorders.

Regarding the stability of the PPE, the low coefficient obtained also indicates a great variability in the performance of the participants across administrations. Most of the participants produced fewer repetitions on the retest, but an inspection of these data allowed us to see that several participants obtained a higher PPE score. A qualitative analysis of data might be necessary to help clarify the nature of this variability in performance.

Similar results with reference to stability have been found with tests akin to the FPT. The Ruff Figural Fluency Test (RFFT, Ruff, 1988) consists of five sheets of paper with one of the sheets identical to the FPT and the other four including interfering elements and asymmetrical and irregular positioning of the points. Using the RFFT, Basso et al. (1999) obtained a coefficient of $r = .71$ for UD, and .39 for the PPE, across a 12-month testing interval. Still, the UD SEP scores are similar for both studies (10 points for UD in both studies). This might be interpreted to suggest that even when RFFT is a modified version of the FPT both seem to be very similar in essence.

The SEP scores allow us to state that an important increase or decrease in the UD and PPE production is necessary to affirm that there has been a significant change in an individual in a second or subsequent assessment. For example, if we were evaluating the progress of a treatment administered to an individual,

an increase of more than 10 UD, or a reduction of more than 8% in the PPE, should be necessary to affirm that the participant has improved. Conversely, a decrease of less than 10 UD or an increase of less than 8% in the PPE would not allow us to state that the condition of an individual is worsening.

The above considerations regarding the significant change in the test scores are directly applicable to normal samples. Nevertheless, these coefficients could be different if obtained from clinical samples. Research around this topic is not consistent. While some studies suggest that these coefficients are different in clinical than in normal samples (Heaton et al., 2001), others have not found evidence of this effect even when 10 different clinical groups where compared to a normal sample (Zhu, Tulsky, Price, & Chen, 2001).

Finally, the results of this study on the influence of age and education on the scores of the FPT differ significantly from the results of other studies (Lee et al., 1997; Santa-Maria et al., 2001). Specifically, others have found nonsignificant or much lower correlations than the ones found in this study. This is likely due to the difference in the samples' constitution. The samples of the cited studies had a much more limited range of age and education. Basically, their samples were comprised of young and educated people. Samples in this study included a wide range of age and education. The age and education ranges in the sample from Study 1 were 15–94 and 1–28 years, respectively.

In summary, the FPT has acceptable reliability coefficients for UD but not for PPE. Therefore interpretation of PPE scores, as well as their changes, should be carefully interpreted considering additional clinical and psychometric data.

REFERENCES

Anastasi, A., & Urbina, S. (1998). *Tests Psicológicos* [Psychological testing]. México: Prentice-Hall.

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist*, *13*(3), 283–292.

Burguess, P. W. (1997). Theory and methodology in executive functions and research. In P. Rabbitt (Ed.), *Methodology of frontal and executive functions*. Hove, UK: Psychology Press.

Charter, R. A (1996). Revisiting the standard errors of measurement, estimate, and prediction and their application to test scores. *Perceptual and Motor Skills*, *82*, 1139–1144.

Cicchetti, D. V. (1994). Guidelines criteria and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychological Assessment*, *6*, 284–290.

Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., Marcotte, T. D., et al. (2001). Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology*, *16*, 75–91.

Lee, G. P., Strauss, E., Loring, D. W., Mc Closkey, L., Haworth, J. M., & Lehman, R. (1997). Sensitivity of figural fluency on the Five-Point Test to focal neurological dysfunction. *The Clinical Neuropsychologist*, *11*(1), 59–68.

Lezak, M. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Martínez Arias, R. (1995). *Psicometría. Teoría de los tests psicológicos y educativos* [Psychometrics. Theory of psychological and educational tests]. Madrid: Síntesis.

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

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

Santa-Maria, M. P., Martin, J. A., Morrow, C. M., & Gouvier, W. D. (2001). On the duration of spatial fluency measures. *International Journal of Neuroscience*, *106*(3–4), 125–130.

Urbina, S. (2004). *Essentials of psychological testing*. New York: John Wiley & Sons.

Zhu, J., Tulsky, D. S., Price, L., & Chen, H.-Y. (2001). WAIS-III reliability data for clinical groups. *Journal of the International Neuropsychological Society*, *7*, 862–866.