BIO ACCENT

BAOJ Neurology

David W. Harrison, et al, BAOJ Neuro 2016 2: 3
2: 021

Research

Test-Retest Reliability of the Ruff Figural Fluency Test

Jared Rowland[1], Michael Knepp[1], Chad Stephens[1], Ryoichi Noguchi[1], Sheri Towe[1], Chris Immel[1], Benjamin B. DeVore[1], Patti Kelly Harrison[1] and David W. Harrison[1*]

[1]Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

Abstract

The Ruff Figural Fluency Test (RFFT), originally designed by Ruff et al. in 1987 [1], is a neuropsychological testing instrument frequently used in clinical, medical, and research settings as an indication of non-verbal fluency reflective of potential right hemispheric deficits. Research on the RFFT has shown that it can be useful in determining potential frontal lobe damage, particularly neural loss in the right frontal region. Given the high utility of the RFFT as a measure of design fluency deficits, various research efforts have been made to establish the test-retest properties of the RFFT. The current study sought to replicate previous efforts provided by Ruff et al. (1987) [1] and Ross et al. (2003) [2] looking at the normative data of the RFFT on an undergraduate population. Multiple factors were taking into consideration, including age, ethnicity, and gender. Findings from the study indicate high test-retest reliability, despite improved scores on the RFFT between test 1 and test 2. Overall, the findings from the current study supported previously demonstrated test-retest properties of the RFFT, further establishing this neuropsychological tool as a strong measure of non-verbal fluency and right hemispheric functioning.

Keywords: Frontal Lobe; Fluency; Psychometrics: Neuropsychological Assessment; Figural Fluency; Verbal Fluency; Test-Retest Reliability, Assessment; Brain; Neuroscience; Neuropsychology

Introduction

Verbal fluency has long been used by neuropsychologists to assess frontal lobe functioning and has been shown to reliably discriminate between normal individuals and those who have experienced a neurological insult, especially to the anterior region of the brain [3,4]. However, deficits in verbal fluency are generally thought to be present after damage occurs either bilaterally to the frontal lobes or primarily to the left frontal lobe. Nonverbal fluency, or design fluency, has also been considered an important indicator of brain injury. It is believed that damage to the right frontal lobe as well as bilateral frontal lobe damage can produce deficits in design fluency [5,4 6].

Jones-Gotman and Milner (1977) developed the first test that was sensitive to right frontal lobe functioning, the Design Fluency Test (DFT) [7]. The strength of this assessment was that it served as a visuospatial analog of Thurstone's written word fluency test [8], but the DFT also lacked published normative data, had poor inter-rater reliability, and there were interpretation difficulties in patients with confounding visuoconstructive or motor deficits [9]. Due to these limitations, the applicability of the DFT has been limited in scope.

The Ruff Figural Fluency Test [1] was developed in response to a lack of tests of non-verbal fluency with adequate psychometric properties and to be analogous to standard tests of verbal fluency. The RFFT was derived from the Five Point Test (Regard, Strauss, & Knapp, 1982) [10] and consists of a series of stimuli that allow an individual to connect an array of five dots within a box to create a design. Individuals are evaluated on the number of unique designs they generate as well as a ratio of repeated designs (perseverative errors) divided by unique designs.

The RFFT was originally tested on a sample of 358 normal individuals predominantly from California and Michigan [1]. It was shown that gender did not affect an individual's score, but scores varied by age and education level. Normative data were collected based on age and education, with three education ranges and four age ranges, resulting in a total of 12 subgroups. Ninety-five individuals representative of the original sample completed the RFFT again after 6 months. The mean for unique designs increased from a mean of 100 (*SD* = 21.8) to a mean of 108 (*SD* = 22.1). A correlation of .76 was found between unique designs created during the initial test and the retest, indicating satisfactory test-retest reliability. The authors noted that normal participants were likely to produce more unique designs upon retest but the differences were negligible. More informative was that perseverative errors increased from 6.53 (*SD* = 5.8) to 7.76 (*SD* = 4.8); a correlation of .36 was found between preservative errors at time one and time two. Ruff and colleagues explained that these between perseverative errors were inherently more variable and thus the correlation between time one and time two was expectedly low. Regarding clinical application of these results it was specified that head-injured patients who significantly reduced perseverative errors upon retest would be demonstrating recovery of ability.

***Corresponding author:** David W. Harrison, Director, Behavioral Neuroscience Laboratory, Psychology Department, College of Science, Virginia Polytechnic Institute, Blacksburg, VA 24061-0436, USA, E-mail dwh.vatech@gmail.com

Sub Date: September 19, 2016, **Acc Date:** September 30, 2016, **Pub Date:** September 30, 2016.

Citation: Jared Rowland, Michael Knepp, Chad Stephens, Ryoichi Noguchi, et al. (2016) Test-Retest Reliability of the Ruff Figural Fluency Test. BAOJ Neuro 2: 021.

The RFFT has been used for a variety of purposes. The measure has been shown to reliably differentiate between normal individuals and those who have experienced a head injury [4], as well as distinguishing individuals with right frontal lobe lesions from those with non-right frontal lobe lesions [11]. Ruff and colleagues (1994) also demonstrated that the RFFT could differentiate among individuals with right frontal lobe lesions from individuals with left frontal lobe lesions, left posterior lesions, and right posterior regions, suggesting that the RFFT can be broadly used to identify frontal lobe deficits, further enhancing its utility as a clinical neuropsychological instrument. Additionally, Foster, Williamson, and Harrison (2005) [12] provided further evidence that the RFFT is sensitive to right frontal lobe functioning using electroencephalography. They showed that individuals who perform poorly on the RFFT exhibited higher delta magnitude over the right frontal lobe than individuals who performed well on the RFFT, indicating that performance on the RFFT is indicative of right frontal lobe functioning.

Examinations of the psychometric properties of the RFFT have replicated the increase in unique designs from test to re-test that was observed in the initial psychometric testing conducted by Ruff et al. (1987) [1]. For instance, Basso, Bornstein, and Lang (1999) [13] administered the RFFT as part of an investigation of practice effects on a number of different executive tests. The RFFT was administered to 50 males (mean age = 32.50, mean education = 14.98 years) at baseline and twelve months later. A significant increase was observed for unique designs, increasing from 52.02 (*SD* = 8.59) at baseline to 58.83 (*SD* = 8.86) at twelve months; no significant increase in the error ratio was observed. This finding suggests the measure is likely to be influenced by practice, and the authors note the magnitude of practice-mediated improvement that one can expect within a 12-month timeframe.

In another study, [14] included the RFFT in an investigation of the ability to malinger on neuropsychological tests. The measure was re-administered after a three-week interval to a group instructed to fake bad performance and a control group given normal instructions about how to complete the measures. Twenty-one individuals were administered the RFFT at each time point, with a mean age of 23.3 years and average education of 13.7. A significant improvement at the second administration was found such that control subjects increased their design scores from a mean of 100.9 (*SD* = 24.5) to a mean of 117.7 (*SD* = 26.9). The malingering group also increased their design scores from a mean of 63.2 (*SD* = 17.8) to a mean of 71.5 (*SD* = 22.1), which is consistent with the assertions of [1] to that effect.

A more recent study was conducted to examine the psychometric properties of the RFFT [2]. They recruited 95 undergraduates to take the RFFT and invited them to take it again after an average interval of seven weeks. Forty-eight individuals were retested and produced a mean of 114.5 (*SD* = 24.6) unique designs compared to a mean of 106.3 (*SD* = 23.1) unique designs produced by the original sample. There was not a significant increase in errors (Time 1 errors = 8.4, Time 2 errors = 8.2). Collectively, the findings from the studies investigating the psychometric properties of the RFFT indicate the existence of practice effects for the number of unique designs created, but studies have not found changes in the error rates among those who complete the measure for the second time.

In the original normative study of the RFFT [1] the college population was not specifically targeted. Based on the normative data provided by the RFFT manual, the appropriate norms for use with a college population would correspond to the age group 16-24 years with 13-15 years of education. In the original normative experiment, there were 29 individuals who fit these criteria [1]. Given that this age range likely included individuals currently enrolled in college, as well as those who have already graduated, it is important to determine if an actual undergraduate sample would differ significantly from the sample reported by Ruff et al. (1987) [1]. Only the study conducted by Ross et al. (2003) [2] could be considered to have accomplished this. The Ross et al. (2003) [2] sample produced a similar number of unique designs to the corresponding age and education subgroup of the original normative sample of 29 individuals, who produced an average of 108.2 (*SD* = 18.6) unique designs and 8.4 (*SD* = 5.3) preservative errors. The original test-retest data reported by Ruff et al. (1987) [1] was not discussed in terms of age and education, but rather as a complete population. Thus, the only test-retest data available for an undergraduate sample is that produced by Ross et al. (2003) [2], which have not been replicated. However, the [2] study found similar results to the [1] study, with individuals producing a higher number of unique designs but a similar number of preservative errors at retest.

The primary goal of this study was to replicate the normative data on undergraduate students provided by the Ruff et al. (1987) [1] and the Ross et al. (2003) [2] studies. We also sought to replicate the test-retest data on undergraduate students provided by the Ross et al. (2003) study. A replication would provide better information on how the RFFT can be used when investigating the undergraduate population in both clinical and research settings. We expect to see a significant increase in unique designs from the initial test to the re-test. We do not expect preservative errors to increase from one test to the next.

Method

Data were collected for this study as part of a larger project investigating the relationship between right frontal and cardiovascular functioning with a number of other domains including: Substance Abuse, Trauma, Anxiety, Temperament, and Emotion Regulation. Participants completed questionnaires online and were then invited to the laboratory session. In the laboratory session individuals were fitted with electrodes in order to gather cardiovascular data. A three-minute cardiovascular reading was taken, followed by administration of the RFFT according to the guidelines set forth in the manual, and a final three-minute cardiovascular reading taken at the end of the session. Data were collected from participants on a longitudinal and cross-sectional basis at two times, with two months between longitudinal data collections. This produced two cohorts, a longitudinal cohort who participated in the study at both

time periods, and a cross-sectional cohort who only participated in the study at one of the two time periods. The test-retest interval was selected to approximate patient care needs in a medical center rehabilitation unit and for the demonstration of progress or recovery of function or for the provision of evidence of decline.

Participants

One hundred and eighty-six undergraduates between the ages of 18 and 24 (M = 18.75, SD = 3.176) were recruited for participation in the laboratory portion of the study at a large university in Southwest Virginia. Two subjects were removed before analyses due to invalid responses on the RFFT. Of the total 186 undergraduates, data were gathered from 38 participants at both time one and time two. These participants ranged from 18 to 24 years (M = 18.11, SD = 4.61). This study was inclusive of all students taking psychology classes who wished to participate in research. The only exclusion criteria was that participants must be 18 years of age or older to participate. The demographics of the study sample were 77% Caucasian, 4% African-American, 1% Hispanic, 15% Asian-American, and 3% other.

The sample for the first trial was 65.8% women; this was most likely due to the high enrollment of women in psychology courses. At the longitudinal follow-up, the sample was 71% women. Most participants were from the eastern seaboard of the United States. This does differ from the Ruff et al. (1987) [1] sample, which was predominately from California and Michigan.

Measures

Ruff Figural Fluency Test [1]. The RFFT consists of five trials, each containing varied presentations of dot matrices. Each trial contains 35 dot matrices on which the subject is asked to connect the dots to create as many unique designs as possible during a one-minute time period. The RFFT is scored by totaling the number of unique designs, as well as design repetitions, also referred to as perseverations, which are scored as errors. Design fluency is then computed by subtracting the total number of preservative errors from the unique design total and using normative data to produce a T score. A T score for a participant's error ratio is also computed using the total number of perseverative errors divided by the total number of unique designs. These T scores are used with normative data provided by the manual to produce qualitative classifications of performance (e.g. Average, Impaired, Superior, etc.) The RFFT has been shown to have good inter-rater reliability [15] as well as good test-retest reliability [1,2].

Procedure

Upon arrival to the lab, participants were seated in a normal classroom chair and desk, and given an informed consent form to read and sign. For this study, testing sessions included between one and eight participants. Participants began each section with a three-minute cardiovascular baseline recording. Following this recording, experimenters passed out RFFT booklets, ink pens, and read aloud standardized instructions. The instructions were given before the start of the first practice trial page and participants were monitored as they completed these items to ensure that the instructions were being followed. Any participant not following instructions were offered clarifications about how to complete the task, and questions regarding the instructions were also answered at this time. Participants were informed that they were not allowed to go back to a previous page once that section had been completed.

The RFFT administration continued through all five parts in this manner: task instruction, completion of practice items under monitoring, and completion of the regular items. Afterwards, the experimenters collected the RFFT booklets and another cardiovascular recording was taken before the laboratory session was finished. Following the laboratory session, both experimenters scored each RFFT booklet. Any discrepancies were discussed and evaluated by the experimenters.

Data Quantification

The total number of unique designs and perseverative errors were calculated for each participant. Error ratio was calculated as the number of perseverative errors divided by the total number of unique designs. Classification of the data was done on total unique design creation and error ratio based on previous norms for 16-24 year old individuals with 13 – 15 years of education provided by the RFFT administration manual.

Results

Homogeneity of Sample

T-tests were conducted to ensure there were no differences between individuals participating for the first time at time 1 and those participating for the first time at time 2. The two groups did not differ significantly with regards to age [$t(180) = .661$, $p > .10$] or GPA [$t(175) = -.233$, $p > .10$] The two groups also did not differ in their performance on the RFFT. There were no significant differences with regard to number of unique designs [$t(182) = .167$, $p > .10$], perseverative errors [$t(182) = 1.773$, $p < .078$], or error ratio [$t(182) = 1.665$, $p < .098$]. Due to the lack of differences, the two waves were combined and considered a single group referred to as Time 1 in subsequent data analyses.

Gender Differences

There were no gender differences on Time 1 RFFT performance with regard to number of unique designs [$t(182) = -.521$, $p > .10$], perseverative errors [$t(182) = 1.508$, $p > .10$], or error ratio [$t(182) = 1.412$, $p > .10$]. There was also no difference in GPA [$t(175) = -.578$, $p > .10$]; however females were significantly older than males in this study [$t(180) = 2.395$, $p < .018$], with the average female age being 19.23 years and the average male age being 18.75 years.

Age

The age range for this sample was 18 through 24. There were no differences at Time 1 on total number of unique designs [$F(6,175) = .077$, $p > .10$], total number of perseverative errors [$F(6,175) = .464$, $p > .10$], or error ratio [$F(6,175) = .643$, $p > .10$], that related to age. Due to this lack of difference, individuals of different ages were combined in subsequent analyses.

Ethnicity

Due to small numbers of certain minority groups, comparisons

were only conducted between Caucasians, African-Americans, and Asian-Americans. No significant differences were found among any of the groups on the number of unique designs produced [$F(2,173) = .667$, $p > .10$], the number of errors produced [$F(2,173) = 1.616$, $p > .10$], or the error ratio [$F(2,173) = .421$, $p > .10$]. There were also no significant differences in age [$F(2, 171) = .277$, $p > .10$] or GPA [$F(2, 166) = .596$, $p > .10$]. Due to the lack of differences, ethnic groups were combined in subsequent analyses.

Follow up Sample

Comparisons were conducted to examine if the group of participants who completed the RFFT in the second administration differed in any way from the participants who only completed the RFFT at the first administration. Comparisons between the two groups on RFFT performance at Time 1 reveal no significant differences on total unique designs [$t(182) = -1.472$, $p > .10$], perseverative errors [$t(182) = -1.268$, $p > .10$], or error ratio [$t(182) = -.714$, $p > .10$]. There were also no significant differences between the groups with regards to gender [$t(182) = .769$, $p > .10$], age [$t(180) = 1.163$, $p > .10$] or GPA [$t(175) = .654$, $p > .10$].

RFFT Performance

Descriptive data for performance on the RFFT at baseline and at the longitudinal follow-up is presented in Table 1. Paired *t*-tests were run to compare Time 1 and Time 2 scores on the RFFT for participants who completed both the initial session and the longitudinal follow-up session. There was a significant difference between unique design scores at Time 1 and Time 2 [$t(37) = -7.50$, $p < .001$], with a greater number of unique designs created at Time 2 ($M = 107.53$, $SD = 29.35$) compared to Time 1 ($M = 89.24$, $SD = 23.85$). However, there were no differences between Time 1 and Time 2 for the total number of perseverative errors committed [$t(37) = .576$, $p > .10$] or for the error ratio [$t(37) = 1.37$, $p > .10$]. This result indicates that participants increased their number of unique designs but not their errors at Time 2, which is similar to previous research [2,12].

Individuals were classified into qualitative groups using the norms provided by the RFFT manual; these classifications can be seen in Table 2. Classifications based on total unique designs produced at Time 2 were significantly different from the classifications based on the Time 1 administration ($\chi^2 = 47.32$, $df = 20$, $p < .005$). Of the 38 participants, 9 participants had the same classification for unique design between Time 1 and Time 2, 3 participants had a lower (worse) classification at Time 2, and 26 participants increased in classification level at Time 2. Also, classification based on error ratio was found to improve from Time 1 to Time 2 ($\chi^2 = 30.75$, $df = 12$, $p < .005$). Of the 38 participants, 13 individuals had the same classification for error ratio from Time 1 to Time 2, 7 individuals had a lower classification at Time 2, and 18 individuals increased in classification level at Time 2. This finding in error ratio appears to be driven by the increases in total unique designs, not by any changes in the amount of errors that a particular subject is making from Time 1 to Time 2.

Table 1: Descriptive statistics for RFFT performance in the current study.

Time 1	N	Minimum	Maximum	Mean	SD
Unique Designs	184	29	154	83.7	26.12
Errors	184	0	30	3.51	4.26
Error Ratio	184	0	.33	.04	.05
Time 2	**N**	**Minimum**	**Maximum**	**Mean**	**SD**
Unique Designs	38	41	168	107.53	29.35
Errors	38	0	17	3.84	4.31
Error Ratio	38	0	.20	.04	.04

Test-Retest Reliability of RFFT in an Undergraduate College Sample

Correlation analyses were run between the Time 1 and 2 administrations of the RFFT, including only individuals who completed both administrations. The total number of unique designs was significantly correlated between Time 1 and 2 ($r = .86$, $p < .005$). Total number of perseverative errors and error ratio were also significantly correlated between Time 1 and Time 2 (perseverative errors, $r = .42$, $p < .01$; error ratio, $r = .33$, $p < .05$), though not as strongly as unique designs. These correlations were also examined for gender differences. For men ($n = 11$), the total number of unique designs ($r = .89$, $p < .001$), total number of perseverative errors ($r = .61$, $p < .05$), and error ratio ($r = .66$, $p < .05$) were significantly correlated between baseline and follow up. These correlations are much higher than in the combined sample. For women ($n = 27$), the total number of unique designs ($r = .84$, $p < .001$) remained correlated, however, the total number of perseverative errors was only marginally correlated ($r = .37$, $p < .06$) and error ratio was not significantly correlated ($r = .24$, $p > .10$).

Reliable Change Intervals

Reliable change intervals were calculated to demonstrate the magnitude of change that would be necessary in order for that change to be attributed to more than simply practice effects. We chose to use the Standard Error of Prediction as suggested by Charter (1996) [16] and used previously by Basso et al. (1999) [13] to examine practice effects on the RFFT since measurement error is likely correlated between the test administrations. This method computes a 90% confidence interval for the follow up score based on the test-retest reliability and standard deviation. This confidence interval is then bracketed around an estimated true score, which is developed from the mean, test-retest reliability, and the individual's observed score. Scores falling within the confidence interval bracket are likely due to measurement error, practice effects, or chance and do not represent change based on cognitive or neurological improvement or deterioration. Scores falling outside of the confidence interval bracket therefore likely represent change beyond what would be expected from practice effects or measurement error. Thus, scores falling outside the 90% confidence interval likely represent true change.

Table 3 presents descriptive statistics describing the confidence interval as well as the number of participants who fell outside of that interval. This table shows that the 90% confidence interval for

Table 2: Classifications based on Ruff et al. (1986) normative data at Time 1 and Time 2 for Unique Designs and Error Ratio.

Designs					
Time 1	**N**	**%**	**Time 2**	**N**	**%**
Impaired	56	30.4	Impaired	4	10.5
Borderline	23	12.5	Borderline	1	2.6
Low Average	30	16.3	Low Average	6	15.8
Average	58	31.5	Average	12	31.6
High Average	10	5.4	High Average	11	28.9
Superior	4	2.2	Superior	0	0
Very Superior	3	1.6	Very Superior	4	10.5
Errors					
Time 1	**N**	**%**	**Time 2**	**N**	**%**
Borderline	10	5.4	Borderline	1	2.6
Low Average	2	1.1	Low Average	2	5.2
Average	65	35.3	Average	9	23.7
High Average	22	12	High Average	4	10.5
Superior	85	46.2	Superior	22	57.9

Table 3: Average estimated true score, test retest correlation, SEP, and the 90% Confidence interval for Unique Designs and Error Ratio.

	M	R_{Y1Y2}	SEP	90% CI	Increase	Decrease
Unique Designs	92.8	0.86	14.98	+/- 24.56	13	0
Error Ratio	0.015	0.33	0.038	+/- 0.062	1	0

unique designs was +/- 24.56 and for error ratio it was +/- 0.062. This means that an individual's score could increase by up to 24 designs above their estimated true score and it would still be due to chance.

Discussion

The purpose of this study was to investigate the test-retest properties of the Ruff Figural Fluency Task [1] in a college population and to replicate previous findings within the literature. Analyses indicated that age, gender, and ethnicity were not associated with performance on the RFFT. This finding is consistent with previous studies [2] given that the age range of the current sample falls within a single age range of the original normative sample [1]. Also consistent with previous studies is the finding that the number of unique designs produced increases significantly upon re-administration, while the number of perseverative errors does not [1,2]. Due to the increase in unique designs, the classification level based on the RFFT manual improved significantly for the majority of individuals for both unique designs and error ratio in the current study. The test re-test correlation of unique designs was much higher than that of perseverative errors or error ratio scores. Once again, this is similar to other investigations [1,2].

Table 4 offers a comparison of the baseline and longitudinal follow-up performance on the RFFT in the current study with those conducted by Ruff et al. (1987) and Ross et al. (2003) [1,2]. As

Table 4: A comparison between studies examining RFFT performance in a population of undergraduate age.

Time 1	**Current Study**	**Ross et al. (2003)**	**Ruff et al. (1987)***	**Ruff et al. (1987)****
Unique Designs	83.7(*SD*=26.12)	106.3 (*SD* = 23.1)	108.2 (*SD* = 18.6)*	100 (*SD*=21.8)
Errors	3.51(*SD*=4.26)	8.4 (*SD*=8.8)	8.4 (*SD* = 5.3)*	6.53 (*SD* = 5.8)
Time 2	**Current Study**	**Ross et al. (2003)**	**Ruff et al. (1987)***	**Ruff et al. (1987)****
Unique Designs	107.53(*SD*=29.35)	114.5 (*SD* = 24.6)	108 (SD=22.1)	108 (SD=22.1)
Errors	3.84(*SD*=4.31)	8.2 (*SD*=8.1)		8.2 (*SD*=8.1)

*These numbers represent the subset falling in the 16-24 year age range and 13-15 years of education. There are no time 2 numbers for this sample because the normative retest sample was a representative sample not segregated by age or education.

**These numbers represent the entire normative sample.

this table shows, the current sample produced the lowest overall means both at Time 1 and Time 2; however, the current sample also produced the largest increase in unique designs from Time 1 to Time 2. The difference in overall production of designs could be due to the inclusion criteria for participation in each of the studies. The only inclusion criterion for the current study was that individuals be over the age of 18 and a currently enrolled student. The other two studies had more stringent criteria, including an absence of head injury, etc. However, the only possible exclusion criteria (neurological impairment, learning disorder, seizures, ADHD, etc.) endorsed by more than two participants was the experience of a concussion, which 40 of the original 184 participants reported and 11 of the 38 follow up participants reported. Due to the possible biasing nature of experiencing a concussion, the RFFT performance of these individuals was compared to those individuals not reporting a concussion. There were no significant differences for total unique designs created ($t[181] = .118$, $p > .10$), perseverative errors ($t[181] = .282$, $p > .10$), or error ratio ($t[181] = .106$, $p > .10$). This indicates that there was no difference in performance between those individuals reporting having experienced a concussion and those not reporting having experienced a concussion. It is important to consider that this is self-report data without corroborating medical evidence.

Test-retest reliability data indicate a gender effect, such that the RFFT is more reliable for males than females. Previously, only initial test performance has been examined for gender effects, and they have not been found. However, this study was not designed to examine gender differences in test-retest administrations and the analysis was conducted with only 11 males and needs to be replicated.

Reliable change indices were also created for this sample in order to examine the role of practice effects and measurement error in the observed increases in unique designs on the RFFT. This indicated that individuals needed to improve their scores drastically (unique designs > 24 and error ratio >.062) above an estimated true increase (average estimated change for unique designs = 9.1, for error ratio = -0.025). 13 individuals met criteria for unique designs, but only one for error ratio. There was no gender difference in the individuals who improved their scores ($\chi^2 = 0.032$, $df = 1$, $p > .10$), indicating that the test-retest gender difference is not present in individuals demonstrating significant improvements.

Implications

The replication of an increase in unique designs but not perseverative errors with the re-administration of the RFFT is important for both clinical and research purposes. If the RFFT is being used as a way to track clinical improvement, then the increase in the production of unique designs at later administrations will be important to consider because a clinician requires that the assessment indicate true improvement on the design task that is not attributable to practice effects. The same is true of evaluating an intervention in a clinical trial or research setting. It would be tempting to suggest an increase in performance or classification implies improvement in frontal lobe functioning [1]. Originally suggested that this increase represented an increase in cognitive flexibility; however, it is difficult to attribute such a broad increase in cognitive flexibility to a single administration of the RFFT. It seems just as likely that the increase in performance is simply the result of practice effects. Ruff and colleagues (1987) also suggested that a decrease in perseverative errors would be a more important indicator of improvement than an increase in unique designs. Based on research into the effects of multiple administrations of the RFFT [2, 13] this seems to be a reasonable suggestion given the apparent stability of perseverative errors in individuals who are not receiving an intervention.

The RCI analysis conducted in this study, as well as the one conducted by [13], indicates that an individual must improve their performance by a large margin in order for the change to be outside of the realm of practice effects. However, each of these analyses is representative of the sample used in the study, not necessarily the population in general. It would be possible to develop norms based on an RCI across age and education level, similar to that in the RFFT manual for initial performance. These norms could then be used to determine levels of change necessary to be considered significant. This would be an important step in improving the reassessment of executive functioning.

An alternative suggested by McCaffrey and Westervelt (1995) and endorsed by Basso, Bornstein, and Lang (1999) [13] is the use of multiple baselines. The use of multiple baselines is suggested to reduce the impact of practice effects by having them included in the initial testing; therefore the presence of practice effects at re-testing will not impact the interpretation of improvements. Whichever course of action is taken, it is clear that tests of executive functioning naturally improve over time and that steps need to be taken to account for this natural improvement in the areas of both treatment and research.

Future studies could provide a better understanding of these findings by investigating if the administration of the RFFT increases performance on other measures of cognitive flexibility, such as the Wisconsin Card Sorting Task or Trail making part B. Administering the RFFT at one point and other measures at a later point could provide information about true change versus practice effects.

It would also be useful to know if the observed practice effects diminish with time. Administering the RFFT, or other executive measures, more than just twice could shed light on how individuals improve on these measures over time. It is quite possible that there is a ceiling effect and that after a certain point improvement on the measure would stop or plateau. It is also possible that some tasks or parts of tasks (errors on RFFT) do not show practice effects until after several administrations. So far, two administrations do not appear to have an effect on the number of perseverative errors produced on the RFFT. It may be that four or five administrations would reveal a significant decrease in perseverative errors as individuals become more familiar with the task and have more time to develop and refine strategies for keeping track of designs already produced.

References

1. Ruff RM, Light RH, Evans RW (1987) The Ruff Figural Fluency Test: A Normative Study with Adults. Developmental Neuropsycholog 3(1): 37-51.
2. Ross, TP, Foard EL, Hiott FB, Vincent A (2003) The reliability of production strategy scores for the Ruff Figural Fluency Test. Archives of Clinical Neuropsychology 18: 879-891.
3. Baldo JV, Shimamura AP (1998) Letter and Category Fluency in Patient with Frontal Lobe Lesions. Neuropsychology 12(2): 259-267.
4. Ruff RM, Evans R, Marshall LF (1986) Impaired Verbal and Figural Fluency after Head Injury. Archives of Clinical Neuropsychology 1: 87-101.
5. Baldo JV, Shimamura AP, Delis DC, Kramer J, Kaplan E, et al. (2001) Verbal and design fluency in patients with frontal lobe lesions. Journal of the International Neuropsychological Society 7: 586-596.
6. Harrison DW (2015) Brain asymmetry and neural systems: Foundations in clinical neuroscience and neuropsychology Springer Publishing Company (Neuroscience).
7. Jones-Gotman Milner (1977) Design fluency: The invention of nonsense drawings after focal cortical lesions. Neuropsychologia 15: 653-674.
8. Thurstone & Thurstone (1943) The Chicago tests of primary mental abilities. Chicago IL: Science Research Associates.
9. Lee GP, Strauss E, Loring DW, McCloskey L, Haworth JM, et al. (1987) Sensitivity of figural fluency on the five points test to focal neurological dysfunction. The Clinical Neuropsychologist 11(1): 59-68.
10. Regard M, Strauss E, Knapp P (1982) Children's production on verbal and nonverbal fluency tasks. Perceptual and Motor Skills 55(3): 839-844.
11. Ruff, Allen, Farrow, Niemann, Wylie, et al. (1994) Figural Fluency: Differential Impairment in Patients with Left Versus Right Frontal Lobe Lesions. Archives of Clinical Neuropsychology 9: 41-55.
12. Foster PS, Williamson JB, Harrison DW (2005) The Ruff Figural Fluency Test: heightened right frontal lobe delta activity as a function of performance. Archives of Clinical Neuropsychology 20: 427-434.
13. Basso MR, Bornstein RA, Lang JM (1999) Practice Effects on Commonly Used Measures of Executive Function Across Twelve Months. The Clinical Neuropsychologist 13(3): 283-292.
14. Demakis, G.J. (1999) Serial Malingering on Verbal and Nonverbal Fluency and Memory Measures: An Analog Investigation. Archives of Clinical Neuropsychology 14(4): 401-410.
15. Berning LC, Weed NC, Aloia MS (1998) Interrater reliability of the Ruff Figural Fluency Test. Assessment 5(2): 181-186.
16. Charter RA (1996) Revisiting the standard errors of measurement, estimate, and prediction and their application to test scores. Perceptual and Motor Skills 82: 1139-1144.