The Reliability and Convergent and Divergent Validity of the Ruff Figural Fluency Test in Healthy Young Adults

Thomas P. Ross*
Department of Psychology, College of Charleston, Charleston, SC, USA

*Corresponding author at: Department of Psychology, College of Charleston, Charleston, SC 29424, USA. Tel.: 1-843-953-3339; fax: 1-843-953-7151.
E-mail address: rosttp@cofc.edu (T. P. Ross)
Accepted 17 September 2014

Abstract

The reliability and validity of standard and qualitative scores for the Ruff Figural Fluency Test (RFFT; Ruff, 1988) was examined in 102 healthy undergraduates. Participants (M age = 21.79; SD = 3.7; age = 80% Caucasian) were administered the RFFT and measures assessing executive functions (EF) and other cognitive domains. Inter-scorer reliability was excellent (0.9 range) for most RFFT indices. Test–retest coefficients (M interval = 7 weeks) ranged from 0.64 for the error ratio score to 0.87 for unique designs. RFFT indices correlated with Block Design performance and nonverbal measures of working memory, but were unrelated to measures of verbal fluency, verbal learning, or working memory for verbal material. RFFT novel design output correlated with most measures of EF supporting the convergent validity of this measure. In contrast, correlations between measures of EF and qualitative scores were absent or weak. RFFT score interpretation is discussed in light of relevant models of EF and directions for future research are presented.

Keywords: Ruff Figural Fluency Test; Reliability; Validity; Fluency; Executive functioning

Introduction

The Ruff Figural Fluency Test (RFFT) was developed by Ruff (1988) as a measure of nonverbal fluency analogous to letter fluency tasks such as the Controlled Oral Word Association Test (COWAT; Benton, Hamsher, & Sivan, 1983). Participants are given sheets of paper, each containing several rows of squares and within each square is an arrangement of five dots. Examinees are instructed to connect two or more dots in each square by using straight lines only. The goal is to generate as many designs as possible during a time constraint, while making each figure unique (Ruff, 1988). The RFFT consists of five trials or “parts”, each lasting for 60 s and utilizing a slightly different stimulus presentation. In Part 1, the dots are presented in a concentric arrangement. Parts 2 and 3 employ the same dot arrangement as Part 1, but each include extraneous stimuli that serve as distractors. Diamond-shaped figures are printed in each square as distractors in Part 2, while lines are drawn to connect dots within and between each square as distractors for Part 3. Parts 4 and 5 introduce new, nonconcentric dot configurations without including any distractors (Ruff, 1988).

According to Ruff (1988), the taxonomy called fluency refers to the ability to “utilize one or more strategies that maximize production of responses while simultaneously avoiding response repetitions” (p. 1). Ruff’s (1988) emphasis on strategic responding differs from other views of fluency performance that emphasize working memory ability (e.g., Pennington, Bennetto, McAleer, & Roberts, 1996; Rosen & Engle, 1997) and the regulation of motor programing (Kraybill & Suchy, 2008). These varying viewpoints highlight the complexity of fluency tasks and the need to better understand factors contributing to effective performance on the RFFT and other measures of executive functioning (EF).

Several definitions and models of EF exist which exceed the scope of this manuscript (for a comprehensive review, see Chan, Shum, Toulopoulou, & Chen, 2008). Chan and colleagues (2008) note core elements of EF include reference to planning and reasoning, mental flexibility, working memory, inhibition, strategy generation and implementation, and the effective monitoring, and
regulation of action in the face of novel or non-routine task demands (see also Lezak, Howieson, Bigler, & Tranel, 2012). Reviews of RFFT performance emphasize the ability to initiate and sustain mental productivity, utilize effective strategies for response generation and to self-monitor and regulate responding as critical tasks demands (see Lezak et al., 2012; Ross, Foard, Hiott, & Vincent, 2003; Strauss, Sherman, & Spreen, 2006). Moreover, the RFFT is sensitive to cerebral dysfunction and to anterior brain lesions in particular (see Kraybill & Suchy, 2008 for review). Accordingly, the RFFT is considered a measure of EF and thought to assess abilities similar to other EF measures; however, the RFFT also appears to measure unique elements (e.g., visuospatial skills and motor programming) not represented among EF measures in general or verbal fluency tasks specifically (see Denckla, 1994; Kraybill & Suchy, 2008; Lezak et al., 2012; Ruff, 1988; Strauss et al., 2006).

RFFT performance is assessed by interpreting the total number of unique designs produced across all five trials or parts. The number of unique designs is calculated by subtracting the number of repetitions (i.e., perseverative errors) from the total designs produced. Additionally, an error ratio is calculated by dividing the total number of perseverative designs by the number of unique designs. Although less commonly used, examiners can also calculate qualitative production strategy scores operationally defined as three or more consecutive designs that indicate first, a rotational strategy involving the systematic rotation of a drawing clockwise or counterclockwise; or second, an enumerative strategy whereby subjects systematically add or remove one line (Gardner, Vik, & Dasher, 2013; Ross et al., 2003; Ruff, 1988). Others have applied the term “clustering” to refer to strategy use on figural fluency tasks (e.g., Gardner et al., 2013). Similar to qualitative scores developed for verbal fluency procedures (e.g., Troyer, Moscovitch, & Winocur, 1997), clustering-derived scores (e.g., number of clusters, mean cluster size, and percent designs used in clusters) examine for deliberate strategy use directly while switching (e.g., transition between clusters) is believed to reflect mental flexibility (e.g., Gardner et al., 2013; Hurks et al., 2010; Ross et al., 2003). For the purpose of this manuscript, the term “production strategies” is used synonymously with the terms “clusters” or “strategic clusters.”

The generation and implementation of efficient production strategies is an important cognitive element assessed by the RFFT (Lezak et al., 2012; Ruff, 1988; Vik & Ruff, 1988). Unfortunately, most studies of patient groups do not report data for production strategies or other qualitative scores for the RFFT. Studies of healthy persons report strategy use correlates with novel design output, suggesting these indices assess cognitive operations that underlie effective performance (Gardner et al., 2013; Hurks et al., 2010; Ross et al., 2003). However, persons vary to the extent they employ strategy use on the RFFT, and it is not yet fully understood how strategy utilization contributes to overall performance and whether clustering and switching scores reflect abilities distinct from those assessed by the unique designs produced (Gardner et al., 2013; Ross et al., 2003).

Psychometric investigations of the RFFT, particularly the use of qualitative scores, are few in number. Inter-rater reliability estimates are reported in the 0.9 range for unique designs, the 0.6–0.7 range for perseverations and error ratio scores, and the 0.8–0.9 range for the production strategy scores (Berning, Weed, Aloia, 1998; Ross et al., 2003; Sands, 1998). Reports of test–retest reliability coefficients range from \( r = .7 \) to .8 for unique designs, \( r = .3 \) to .4 for perseverative designs and error ratios, and from \( r = .7 \) to .8 for production strategy scores (Ross et al., 2003). Although the criterion-related validity of RFFT is well established (e.g., Baldo, Shimamura, Delis, Kramer, & Kaplan, 2001; Ruff, 1988; Suchy, Sands, & Chelune, 2003; Williamson & Harrison, 2003), evidence for its convergent validity with other EF measures is lacking. Studies comparing the RFFT with the Design Fluency Test (Jones-Gotman & Milner, 1977) suggest that these tasks are not equivalent as evident by low correlations in the \( r = .2–.3 \) range (e.g., Demakis & Harrison, 1997). Studies report nonsignificant or very low correlations (\( r = .11–.27 \)) between COWAT and RFFT scores in samples of healthy persons (Abwender, Swan, Bowerman, & Connolly, 2001; Demakis & Harrison, 1997; Gardner et al., 2013; Hurks et al., 2010; Ross, Hanouskova, Giarla, Calhoun, & Tucker, 2007; Ruff, Light & Evans, 1987), while studies of patients have yielded mixed results (e.g., Fama et al., 1999; Soble, Donnell, & Belanger, 2013). Gardner and colleagues (2013) found higher correlations (\( r = .2–.3 \)) between the RFFT and the Sorting and Tower Tests of the Delis–Kaplan Executive Functioning System (Delis, Kaplan, & Kramer, 2001) when compared with nonsignificant correlations between the RFFT and measures of intelligence. Gardner and colleagues (2013) suggested future research on the RFFT should examine its relationship with additional EF measures and other constructs (e.g., working memory and spatial abilities).

The present study sought to add to the existing literature by examining (a) the inter-rater and test–retest reliability of the RFFT; (b) the relationship between measures of strategic responding and novel design output; (c) the associations between RFFT scores, estimates of intelligence, measures of letter fluency, working memory, and verbal learning; and (d) the associations between RFFT scores and other putative measures of EF in a large sample of healthy persons. By addressing these basic research questions, the present study would be of interest to researchers who perceive limited information about the RFFT as a barrier to use, and those who assess nonverbal fluency or EF more generally. The following hypotheses were examined: first, excellent inter-rater reliability and modest stability coefficients would result for most RFFT indices with demonstrable practice effects occurring for select scores (e.g., designs produced); second, production strategy indices would relate positively to unique design output, whereas negative associations should result among production strategy scores and perseverative errors; third, absent or low correlations would result between the RFFT and measures of intelligence, verbal learning, working memory, and verbal fluency as evidence of divergent validity; finally, low-to-moderate correlations would result between the RFFT and measures of EF as evidence for
convergent validity. Due to the exploratory nature of this investigation, no specific hypotheses were offered regarding the relationships between specific RFFT scores and various EF measures, although the resulting pattern of correlations will suggest whether RFFT qualitative scores assess similar or distinct facets of EF relative to the overall score.

**Method**

**Participants**

After review and approval from the institutional review board, undergraduates \( (N = 113) \) were recruited from introductory psychology courses at a mid-sized, liberal arts and sciences college in the Southeast. Individuals were not paid for their participation, but each received credit or extra-credit toward their psychology courses. Persons with a history of neurological disorder, learning disability (LD), attention deficit-hyper activity disorder (ADHD), psychiatric disorder, or psychiatric medication usage were excluded from data analyses using a self-report screening questionnaire described below. Additionally, persons who were not naive to neuropsychological testing by way of self-report were excluded. While not the focus of this investigation, some efforts were made to identify (and eliminate from analyses) individuals who may have put forth poor effort on the neuropsychological measures administered. Participants who obtained <13 recognition hits on the California Verbal Learning Test (CVLT; Delis, Kramer, Kaplan, & Ober, 1987) or produced <90% Discriminability on this measure (see Millis, Putnam, Adams, & Ricker, 1995), or who scored <3 SDs from the sample mean on any other two measures, were eliminated. Using the aforementioned criteria, 11 individuals were eliminated, resulting in the final 102 persons for data analyses.

Of the 102 healthy participants, 84% were female and 88% right handed. Eighty percent of the sample was Caucasian, 10% of the students were African American, 6% were Asian American, and 4% reported other ethnic identities. Participants were between the ages of 18 and 40 years \( (M = 21.7; SD = 3.7) \), and the mean estimated Full Scale IQ using the North American Adult Reading Test (NAART; Blair & Spreen, 1989) was 106.1 \( (SD = 8.2) \). The demographic composition of the present “convenience sample” was commensurate with that of the high-learning institution from which participants were recruited; however, it is not representative of the larger population with respect to census information on race for the USA or the southeastern region (U.S. Census Bureau, 2010). Finally, the present sample included four individuals between the ages of 28 and 40 years, whereas the majority of persons (95%) were aged 18–27. The neuropsychological test performance of the four oldest participants was examined and found to be within \( \pm 1 \) SD of the mean on each measure; therefore, the data from these individuals were retained for analyses. The implications for using this relatively homogenous sample of healthy young persons are discussed later.

**Materials**

The RFFT (Ruff, 1988), as described above, was used to investigate the reliability and validity of traditional and qualitative indices of nonverbal fluency. In addition, several other measures were administered to participants to explore the relationship between RFFT performance and other relevant constructs including general intelligence, working memory, verbal learning, and EF. The inclusion of these diverse measures would allow for an examination of the convergent and divergent validity of the RFFT. Estimates of general intelligence included the number of correct responses on the NAART and the raw scores for the Vocabulary and Block Design subtests of the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997a). Measures used to assess working memory included the raw score of the Spatial Span Backward subtest of the Wechsler Memory Scale-Third Edition (WMS-III; Wechsler, 1997b), the raw score of the Letter-Number Sequencing subtest of the WAIS-III, and the total number of errors on the abstract design form of the Self-Ordered Pointing Task (SOPT; Petrides & Milner, 1982). Verbal learning was assessed using the raw score for the number of correct words recalled across trials 1–5 on the CVLT (Delis et al., 1987).

In keeping with theories of EF as a multidimensional construct (for review, see Chan et al., 2008), several putative measures of EF were included to assess a variety of proposed facets (e.g., cognitive flexibility, monitoring, self-regulation, and planning) including strategy utilization when possible. This study included the following EF measures: the number of correct words, clusters and switches produced across letters CFW on the COWAT (Benton et al., 1983) using the qualitative scoring system developed by Troyer and colleagues (1997); the total time in seconds taken to complete the Color-Naming and Color-Word Interference trials of the Stroop Neuropsychological Screening Test (Trennery, Crosson, DeBoe, & Leber, 1989); the total time in seconds taken to complete the Trail Making Test Parts A and B (TMT; Reitan, 1986); the total score (number of points earned across 12 trials) and planning index score (mean time to first move in seconds) of the Tower of Hanoi (TOH) using the stimuli and scoring methods reported by Humes, Welsh, Retzlaff and Cookson (1997); and the number of perseverative responses on the Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtiss, 1993). Reports exist on the nature, reliability
and validity of these measures and therefore this information is not reported here (for reviews, see Lezak et al., 2012; Mitrushina, Boone, Razani, & D’Elia, 2005; Strauss et al., 2006).

To assess health history and demographic background, a self-report questionnaire was constructed and administered to all participants. The questionnaire screened for previously stated exclusionary criteria including a history of neurological disorders (e.g., ADHD, LD, epilepsy, traumatic brain injury, concussion, stroke, or tumors), psychiatric illness (e.g., depression, anxiety, or psychosis), and other medical disorders known to affect neuropsychological functioning (e.g., hypothyroidism or diabetes). This questionnaire also asked participants to identify whether they have participated in previous testing of their cognitive abilities describe broadly as language, attention, memory, problem solving, or intelligence (a.k.a. IQ testing) related to any health problems (e.g., head injury or concussion) or school/academic issues (e.g., LD, attention deficit disorder, poor academic performance, and behavioral problems) by a mental health professional such as a psychologist.

Procedures and Scoring

After obtaining their informed consent, participants were first administered the health and demographic survey, then the aforementioned battery of neuropsychological measures. All tests were administered and scored in accordance with the published manuals by well-trained examiners under the supervision of a PhD level psychologist. After completing the test administration, all participants were debriefed and invited to participate in the second phase of the present study. The mean interval between Time 1 and Time 2 testing was 45.2 days (range = 31–52 days). Participants retested at Time 2 were administered the RFFT, the health status questionnaire, and any neuropsychological measures not administered at Time 1. The order of test administration was counter-balanced to control for possible order effects. Each of the two testing sessions lasted ~60–75 min. During each session, participants were told they would be asked to complete several cognitive tasks, some of which would be timed and so they would be asked to work quickly in these instances. All participants were also asked to perform their best on each task.

Six raters (under the supervision of the primary investigator who is a PhD level psychologist) scored all RFFT protocols. The raters were advanced undergraduate psychology majors who had received prior training in the administration and scoring of several neuropsychological tests. After completing several of the same training sessions with practice protocols, each rater scored all RFFT protocols independently and these values were used for inter-rater reliability analyses. Each participant’s RFFT protocol was inspected carefully by the primary investigator to determine the correct values for all subsequent analyses (e.g., correlations between the RFFT and other measures).

This study examined both the standard RFFT scoring indices and qualitative indices. As described in Introduction, standard RFFT indices include the number of unique designs generated, the number of perseverations, and an error ratio (Ruff, 1988). As described above, production strategy scores that included rotational and enumerative strategies were also scored. Gardner and colleagues (2013) noted that “blended” strategies can occur in a small percentage of respondents. For example, a participant might generate a rotational strategy using a given figure, and then immediately initiate a second rotational strategy using a new figure that constitutes an enumeration of the figure used in the previous rotational cluster (see Gardner et al., 2013, p. 473). This sequence would be scored as two consecutive rotational strategies using the original criteria specified in Ruff’s (1988) manual. In such instances, this study employed the scoring procedure stated in the RFFT manual. This method is conservative and does not assume two or more consecutive clusters reflect a meta-cluster or subordinate strategy on the part of the test taker. Following the procedure outlined by Ross and colleagues (2003), production strategies for Part 3 of the RFFT were not scored, and therefore qualitative score totals reflect performance across four 60 s trials consisting of Parts 1, 2, 4, and 5. According to Ross and colleagues (2003), the type of distractors used (i.e., lines) on Part 3 confound the interpretation of production strategies as some of the dots are preconnected for the respondent leaving them unsure about whether to incorporate these lines into their own drawings.

The present study followed the procedures outlined by Ross and colleagues (2003) to calculate mean cluster size and percent of designs in clusters. The mean cluster size assesses whether a test-taker, on average, utilized a production strategy to the minimal or maximum extent possible when employed. Mean cluster size is calculated by first determining the sum of the cluster sizes produced and then dividing this sum by the number of clusters generated. Following procedures used by Ross and colleagues (2003), a cluster size of 1 reflects the minimum number of designs (3) necessary to score the occurrence of a production strategy or cluster. A cluster size of 2 indicates the use of minimum numbers of designs required for a cluster (3) plus 1, whereas a cluster size of 3 would be scored when the minimum number of strategic designs (3) plus 2 (for a total of five designs) were drawn consecutively, etc. Therefore, a respondent who produced three clusters having sizes of 1, 3, and 3 would receive a mean cluster size score of 2.5 or \(\frac{3 + 3 + 3}{3}\).

The percentage of designs in clusters index (similar to a strategy ratio score; see Gardner et al., 2013) assesses the degree to which respondents employed strategies throughout the protocol, taking into account the number of designs produced overall. It is calculated by summing the number of designs included in each cluster and then dividing this sum by the total number of designs

\[ \text{percentage of designs in clusters} = \frac{\sum_{i=1}^{n} d_i}{n} \]

where \(d_i\) is the number of designs in cluster \(i\) and \(n\) is the total number of clusters.
produced. For example, if a respondent produced 20 designs total and 10 of these were incorporated into strategic clusters, then the percent of designs included in a cluster would be calculated as 10/20 or 50%. This index, similar in principle to the RFFT error ratio, takes a participant’s overall fluency into account when expressing the use of production strategies (Ross et al., 2003). In addition, switches, defined as transitions between two strategic clusters, a cluster and a single (i.e., unclustered) design, or between two unclustered designs were scored (see Hurks et al., 2010). Following procedures typically employed for qualitative verbal fluency scoring (e.g., Troyer et al., 1997) and previous research examining qualitative scores on the RFFT (e.g., Hurks et al., 2010; Ross et al., 2003) perseverations were included in the calculation of qualitative indices for verbal and nonverbal fluency measures. According to Troyer and colleagues (1997) and Hurks and colleagues (2010), any word (or design) generated provides rich information about the underlying cognitive processes contributing to performance independent of its contribution to the number of correct items produced. For all measures, raw scores were chosen over standardized scores to promote greater score variability and normality given the sample of healthy participants were predominantly from the same age group.

Results

Prior to analyses, data were inspected for univariate normality. Data regarding skewness and kurtosis were within acceptable ranges using guidelines recommended by Tabachnick and Fidell (1991); therefore, no data transformations were necessary. The means and SD for all RFFT scores for the current sample (N = 102) are shown in Table 1. No gender differences were observed on any indices (p > .05); therefore, data for males and females are not presented separately. Participants’ performance across RFFT Parts 1, 2, 4, and 5 at Time 1 was fairly comparable, with a few exceptions (see Table 2). Using the Bonferoni method of correction for the number of comparisons (0.05/18 = p < .002), the mean number of novel designs observed for Part 1 was significantly less than the values observed for Part 2 (t(101) = 3.76, p < .002, Cohen’s d = .36), Part 4 (t(101) = 5.19, p < .002, Cohen’s d = .42) and Part 5 (t(101) = 7.14, p < .002, Cohen’s d = .61). A similar pattern was observed for switch scores, as participants generated significantly fewer switches on Part 1 when compared with Part 2 (t(101) = 3.10, p < .002, Cohen’s d = .25) and Part 5 (t(101) = 7.89, p < .002, Cohen’s d = .74). In addition, the mean number of switches on Part 2 was significantly less than the value observed for Part 5 (t(101) = 5.5, p < .002, Cohen’s d = .50). Most notably, there were no differences observed for the mean number of strategic clusters generated on each of the RFFT parts. The mean difference for unique designs is consistent with patterns observed by

Table 1. Mean and SD for RFFT indices at Time 1 and Time 2 (N = 102)
a
<table>
<thead>
<tr>
<th>RFFT score</th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Unique designs</td>
<td>93.72</td>
<td>21.62</td>
</tr>
<tr>
<td>Perseverations</td>
<td>5.67</td>
<td>6.02</td>
</tr>
<tr>
<td>Error ratio</td>
<td>.0605</td>
<td>(.0677)</td>
</tr>
<tr>
<td>Clusters</td>
<td>4.62</td>
<td>3.66</td>
</tr>
<tr>
<td>Mean cluster size</td>
<td>1.72</td>
<td>0.90</td>
</tr>
<tr>
<td>% designs in strategies</td>
<td>22.25</td>
<td>18.61</td>
</tr>
<tr>
<td>Switches</td>
<td>61.35</td>
<td>18.44</td>
</tr>
</tbody>
</table>

Notes: *Mean testing interval was 45.2 days.
*p-value set at .007 for the number of comparisons using Bonferoni method of correction.

Table 2. Performance across RFFT parts for novel designs, clusters and switching scores

<table>
<thead>
<tr>
<th>RFFT part</th>
<th>Unique designs b</th>
<th>Strategic clusters c</th>
<th>Switches d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>1</td>
<td>16.9</td>
<td>5.1</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>18.8</td>
<td>5.3</td>
<td>1.3</td>
</tr>
<tr>
<td>4</td>
<td>18.9</td>
<td>4.2</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>19.9</td>
<td>4.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Notes: *p-value set at .002 for the number of comparisons using Bonferoni method of correction.

bPart 1 significantly differed from Parts 2, 4, and 5 for Mean # of unique designs.

cNo significance differences were observed between parts for Mean # of clusters.

dPart 1 significantly differed from Parts 2, 4, and 5; Part 2 differed from Part 5 for Mean # of switches.
The ICCs for each RFFT summary score are displayed in Table 3. Inter-rater reliability coefficients were excellent using Cicchetti and Sparrow’s (1981) criteria for evaluating practical significance. Of the standard RFFT scores, the highest reliability was observed for unique designs, although the reliability coefficients for perseveration and error ratio scores were also acceptable. Excellent reliability ($r_{icc} \geq .87$) was observed for all qualitative scores, with the percentage of designs included in strategies having the highest reliability using Cicchetti and Sparrow’s (1981) criteria (see Table 3).

Estimates of stability ranged from $r_{icc} = .64$ for error ratio scores to $r_{icc} = .87$ for unique designs (see Table 3). For RFFT qualitative indices, the coefficients of stability ranged from $r_{icc} = .65$ for mean cluster size to $r_{icc} = .83$ for the total number of switches. Using criteria suggested by Nunnally (1978, p. 245), all RFFT indices were above the acceptable range of 0.70 with the exception of the error ratio and mean cluster size scores.

Paired sample $t$-tests were used to examine for differences between participants’ Time 1 versus Time 2 scores using the Bonferroni method of correction for the number of comparisons ($0.05/7 = p < .007$). Upon retesting, participants’ number of unique designs improved significantly [$t_{(101)} = 9.41$, $p < .007$, Cohen’s $d = .63$], as subjects produced 15 more unique designs, on average, at Time 2 ($M = 108.5$, $SD = 24.8$) when compared with Time 1 ($M = 93.7$, $SD = 21.6$). No significant differences between Time 1 versus Time 2 performance were observed for perseverative responses [$t_{(101)} = 0.24$, $p > .05$, Cohen’s $d = .02$] or the error ratio score [$t_{(101)} = 1.27$, $p > .05$, Cohen’s $d = .15$]. Regarding qualitative indices, significant differences were observed upon retesting for the number of switches [$t_{(101)} = 30.32$, $p < .007$, Cohen’s $d = .52$] with higher scores observed upon repeat testing. No differences were observed upon retesting for total number of strategic clusters produced [$t_{(101)} = 2.32$, $p > .007$, Cohen’s $d = .21$] or mean cluster size [$t_{(101)} = 0.82$, $p > .007$, Cohen’s $d = .09$]. Although nonsignificant using a conservative $p$-value, the mean difference observed across testing sessions for the number of clusters constitutes a “small” effect size using criteria by Cohen (1988). In contrast, participants produced 10 more switches and 15 more designs on average, constituting “medium” sized practice effects upon retesting. Interpretation of effect size (e.g., Cohen’s $d$) can be preferable to statistical significance which is dependent on sample size (Cohen, 1988).

The use of strategic responding was frequent in the present sample, as 85.3% of participants employed either a rotational or enumerative strategy at Time 1. However, most participants (76% of the sample) generated six or less strategic clusters at Time 1. Rotational strategies (which occurred in 74.5% of the present sample) were more commonly observed relative to enumerative strategies (which occurred in 52% of the present sample). A similar pattern was observed at Time 2, whereby 86% of respondents employed one form of strategy with the rotational variety being more commonly observed (80.4%) relative to the enumerative type (59.8%).

The correlations among RFFT quantitative and qualitative scores are displayed in Table 4. The number of unique designs produced correlated positively with the number of perseverative responses; however, the error ratio score was unrelated to the unique designs produced. Not surprising, the perseveration and error ratio scores were highly correlated given the interdependent manner in which these values are calculated. Clustering indices and switching correlated positively with the number of unique designs produced. Perseverative responses and error ratio scores correlated negatively with the percentage of designs included in clusters while, correlating positively and more robustly with switching. The error ratio score (which takes the participant’s overall productivity into account) was more consistently correlated with measures of clustering relative to the number of perseverations. Finally, switching correlated negatively with total strategic clusters produced and the percentage of designs used in clusters, but was not related to mean cluster size. Based on the observed correlations among RFFT variables (shown in Table 4), perseverations and the mean cluster size scores were eliminated from subsequent analyses due to lower or less consistent correlations with other variables and redundancy with conceptually similar scores that were retained (i.e., error ratio and percentage of designs used

---

Table 3. Reliability coefficients ($N = 102$) for RFFT scores

<table>
<thead>
<tr>
<th>Index</th>
<th>Inter-rater reliability$^a$</th>
<th>Coefficient of stability$^{a,b}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique designs</td>
<td>.98</td>
<td>.87</td>
</tr>
<tr>
<td>Perseverations</td>
<td>.94</td>
<td>.79</td>
</tr>
<tr>
<td>Error ratio</td>
<td>.84</td>
<td>.64</td>
</tr>
<tr>
<td>Total clusters</td>
<td>.92</td>
<td>.78</td>
</tr>
<tr>
<td>Mean cluster size</td>
<td>.87</td>
<td>.65</td>
</tr>
<tr>
<td>% designs in strategy</td>
<td>.95</td>
<td>.78</td>
</tr>
<tr>
<td>Total switches</td>
<td>.91</td>
<td>.83</td>
</tr>
</tbody>
</table>

Notes: $^a$Intra-class correlation coefficients ($r_{icc}$).
$^b$Mean testing interval was 45.2 days.

---

Ruff (1988) and the modest differences seen between trials (i.e., different letters) on measures of verbal fluency (see Lezak et al., 2012; Ross, 2003; Strauss et al., 2006). Therefore, participants’ performance across each part was combined into a composite or summary score for all subsequent analyses, a procedure consistent with that specified in the RFFT manual (Ruff, 1988).

The ICCs for each RFFT summary score are displayed in Table 3. Inter-rater reliability coefficients were excellent using Cicchetti and Sparrow’s (1981) criteria for evaluating practical significance. Of the standard RFFT scores, the highest reliability was observed for unique designs, although the reliability coefficients for perseveration and error ratio scores were also acceptable. Excellent reliability ($r_{icc} \geq .87$) was observed for all qualitative scores, with the percentage of designs included in strategies having the highest reliability using Cicchetti and Sparrow’s (1981) criteria (see Table 3).

Estimates of stability ranged from $r_{icc} = .64$ for error ratio scores to $r_{icc} = .87$ for unique designs (see Table 3). For RFFT qualitative indices, the coefficients of stability ranged from $r_{icc} = .65$ for mean cluster size to $r_{icc} = .83$ for the total number of switches. Using criteria suggested by Nunnally (1978, p. 245), all RFFT indices were above the acceptable range of 0.70 with the exception of the error ratio and mean cluster size scores.

Paired sample $t$-tests were used to examine for differences between participants’ Time 1 versus Time 2 scores using the Bonferroni method of correction for the number of comparisons ($0.05/7 = p < .007$). Upon retesting, participants’ number of unique designs improved significantly [$t_{(101)} = 9.41$, $p < .007$, Cohen’s $d = .63$], as subjects produced 15 more unique designs, on average, at Time 2 ($M = 108.5$, $SD = 24.8$) when compared with Time 1 ($M = 93.7$, $SD = 21.6$). No significant differences between Time 1 versus Time 2 performance were observed for perseverative responses [$t_{(101)} = 0.24$, $p > .05$, Cohen’s $d = .02$] or the error ratio score [$t_{(101)} = 1.27$, $p > .05$, Cohen’s $d = .15$]. Regarding qualitative indices, significant differences were observed upon retesting for the number of switches [$t_{(101)} = 30.32$, $p < .007$, Cohen’s $d = .52$] with higher scores observed upon repeat testing. No differences were observed upon retesting for total number of strategic clusters produced [$t_{(101)} = 2.32$, $p > .007$, Cohen’s $d = .21$] or mean cluster size [$t_{(101)} = 0.82$, $p > .007$, Cohen’s $d = .09$]. Although nonsignificant using a conservative $p$-value, the mean difference observed across testing sessions for the number of clusters constitutes a “small” effect size using criteria by Cohen (1988). In contrast, participants produced 10 more switches and 15 more designs on average, constituting “medium” sized practice effects upon retesting. Interpretation of effect size (e.g., Cohen’s $d$) can be preferable to statistical significance which is dependent on sample size (Cohen, 1988).

The use of strategic responding was frequent in the present sample, as 85.3% of participants employed either a rotational or enumerative strategy at Time 1. However, most participants (76% of the sample) generated six or less strategic clusters at Time 1. Rotational strategies (which occurred in 74.5% of the present sample) were more commonly observed relative to enumerative strategies (which occurred in 52% of the present sample). A similar pattern was observed at Time 2, whereby 86% of respondents employed one form of strategy with the rotational variety being more commonly observed (80.4%) relative to the enumerative type (59.8%).

The correlations among RFFT quantitative and qualitative scores are displayed in Table 4. The number of unique designs produced correlated positively with the number of perseverative responses; however, the error ratio score was unrelated to the unique designs produced. Not surprising, the perseveration and error ratio scores were highly correlated given the interdependent manner in which these values are calculated. Clustering indices and switching correlated positively with the number of unique designs produced. Perseverative responses and error ratio scores correlated negatively with the percentage of designs included in clusters while, correlating positively and more robustly with switching. The error ratio score (which takes the participant’s overall productivity into account) was more consistently correlated with measures of clustering relative to the number of perseverations. Finally, switching correlated negatively with total strategic clusters produced and the percentage of designs used in clusters, but was not related to mean cluster size. Based on the observed correlations among RFFT variables (shown in Table 4), perseverations and the mean cluster size scores were eliminated from subsequent analyses due to lower or less consistent correlations with other variables and redundancy with conceptually similar scores that were retained (i.e., error ratio and percentage of designs used
Design performance also correlated with RFFT clustering and switching, but not with the error ratio index (see Table 5). The strongest correlation was observed between RFFT unique designs and WAIS-III Block Design Performance. Block type-I error in clusters. The aforementioned reduction of RFFT indices included in subsequent analyses also served to reduce the possibility of type-I error.

Correlations between RFFT scores and estimates of intelligence, working memory, and verbal learning are presented in Table 5. Correlations among measures other than the RFFT are reported when relevant and used in the computation of partial correlation coefficients. Performance on the NAART correlated positively with the number of strategic clusters produced on the RFFT and the percentage of designs included in clusters. WAIS-III Vocabulary subtest performance correlated negatively with the RFFT error ratio, while positive associations (ranging from \( r = .23 \) to \( r = .26 \)) were observed for qualitative indices assessing clustering. The strongest correlation was observed between RFFT unique designs and WAIS-III Block Design Performance. Block Design performance also correlated with RFFT clustering and switching, but not with the error ratio index (see Table 5).

WAIS-III Letter-Number Sequencing scores were not associated with performance on any RFFT measure. Spatial Span Backward and SOPT scores were associated with clustering indices, but not with novel design output, error ratio, or switching scores (see Table 5). The resulting correlation between WAIS-III Block Design and SOPT performance (not shown in Table 5) was significant at \( r = -.20, p = .037 \). A partial correlation was computed to examine whether a significant association between SOPT and RFFT total clusters would result after controlling for variance shared with WAIS-III Block Design performance. The resulting partial correlation remained significant \( (r = .23, p = .022) \).

Participants’ scores on trials 1–5 of the CVLT were unrelated to any RFFT index. A modest, positive association was observed between the CVLT semantic clustering ratio and the total number of strategies produced on the RFFT. This association could not be explained by a shared association with estimates of intelligence, as vocabulary performance, for example, did not correlate with CVLT semantic clustering in the present sample. The remaining zero-order correlations among the group of measures presented in Table 5 did not warrant any further partial correlation analyses.

The observed correlations between RFFT scores and putative measures of EF are shown in Table 6. Correlations among executive measures other than the RFFT are reported when relevant and used in the computation of partial correlation coefficients. No associations were observed between any of the RFFT indices and novel words produced or switching scores on the COWAT. Clustering on the COWAT was associated with RFFT novel design output and switching, but not with clustering on the RFFT. Unique design production was correlated modestly with most executive measures (e.g., \( r = .2 \) range), with the largest correlation observed between unique designs and TMT Part B performance.
The correlations among RFFT qualitative indices and EF measures were less consistent relative to the pattern of results seen for the unique designs produced. TMT-A performance correlated with clustering, the percentage of designs used in clusters and switching, while TMT-B related to clustering only. Stroop Color-Naming performance correlated with clustering, while no significant associations were observed between Stroop Color-Word Naming interference and RFFT qualitative scores. As can be seen in Table 6, perseverative responding on the WCST did not correlate with any RFFT qualitative score. A measure of planning (i.e., TOH time to first move) correlated with novel design output and switching, but not with clustering on the RFFT.

Given that some RFFT indices were correlated with TMT-A and TMT-B, and TMT parts A and B performances were highly correlated in the present sample \( r = .83, p < .001 \), some partial correlations were examined. When controlling for variance shared with TMT-A performance, the resulting partial correlation between TMT-B and RFFT unique designs remained significant \( r = -.22, p = .034 \). In contrast, partial correlation between TMT-B and RFFT clustering was no longer significant after controlling for variance shared with TMT-A. Given the pattern of correlations observed, a partial correlation was also computed to better examine the association between Stroop Color-Word interference trial performance and the number RFFT unique designs produced. After controlling for variance shared with the Stroop Color-Naming trial performance, the resulting partial correlation between the Stroop Color-Word interference trial and RFFT unique designs was nonsignificant. The remaining zero-order correlations among the group of EF measures did not warrant any further partial correlation analyses.

**Discussion**

The objectives of the present study were met and the majority of the predictions were supported. This study examined the reliability and validity of the standard and qualitative scores for the RFFT in a healthy sample of college students. The means and SD for RFFT scores were comparable with patient samples (e.g., Basso, Bornstein, & Lang, 1999; Ruff, 1988) and with studies that utilized samples of healthy college students (e.g., Berning et al., 1998; Demakis, 1999; Ross et al., 2003).

As predicted, the inter-rater reliability estimates obtained for RFFT scoring indices in the present study are similar to those reported in previous investigations, suggesting the standard and qualitative measures can be scored reliably with sufficient practice in individuals with less than graduate-level training (Berning et al., 1998; Ross et al., 2003; Sands, 1998). The coefficients of stability obtained in the present study are also consistent with previous reports of greater reliability estimates for RFFT unique designs (e.g., \( r \approx 0.7 \)) when compared with perseverative responses and error ratio scores (Basso et al., 1999; Ross et al., 2003; Ruff, 1988). Although the pattern is similar, the coefficients of stability observed in the present study are slightly higher than those reported by Ruff (1988) and Basso and colleagues (1999). The present study differed from prior investigations of patient populations that included much longer time intervals between retesting (e.g., 6–12 months), but is highly consistent with previous studies of healthy college students employing short test intervals (Demakis, 1999; Ross et al., 2003).

As anticipated, significant practice effects were observed for the unique designs produced and switching scores, but not for perseverations, error ratio scores, the number of clusters or mean cluster size. Although nonsignificant using a conservative \( p \)-value, a small effect size was observed for the number of clustering strategies generated across testing sessions; however, the
mean increase upon retesting was < 1. In contrast, the medium-sized practice effect observed for the number of unique designs (mean increase of 15 drawings) is meaningful and could result in significant T-score differences that impact test interpretation. The observed increase is highly consistent with previous research using healthy samples (see Demakis, 1999; Ross et al., 2003; Strauss et al., 2006). Therefore, these findings add to a growing body of evidence demonstrating that meaningful gains on some RFFT scores are obtained upon retesting when short time intervals are employed between assessments. These findings may have bearing on RFFT interpretation, as some authors emphasize task novelty as an essential element of EF and associated theories of prefrontal-mediated cognition (see Chan et al., 2008; Denckla, 1994). Therefore, significant practice effects would presumably compromise the assessment of EF. However, clinical studies employing longer testing intervals have not yielded such striking differences between assessments. Therefore, any conclusions that the RFFT is not suitable for repeated assessments or for examination of EF theories are unwarranted, but the length of the testing interval must be considered when evaluating research on this measure.

Qualitative indices of strategy utilization were all positively correlated with unique design output, but only to a modest degree. Inefficiency marked by preservative responding (as measured by the RFFT error ratio score) was negatively associated with clustering indices and positively associated with switching. Switching scores correlated strongly with unique design production; however, data do not support the interpretation that switching is an index of cognitive flexibility in the present sample. Participants, on average, produced only 4-5 clusters total across all trials while generating an average of 94 designs. Therefore, the overwhelming majority of switch scores represented transitions between un-clustered designs which can be interpreted as a lack of clustering rather than indicating flexible shifting between different strategies (see Abwender et al., 2001).

The resulting pattern of correlations among quantitative and qualitative scores is consistent with those observed by Ross and colleagues (2003) and with Gardner and colleagues (2013) who reported correlations between strategy use indices and unique designs in the \( r = -0.3-0.5 \) range. The present findings provide mixed support for Lezak and colleagues (2012) interpretation that the RFFT imposes two important requirements that must be adhered to simultaneously: to be productive and to avoid repetitions. Contrary to this position, repetitions were positively associated with unique design output. Moreover, the use of production strategies is not necessary to avoid repetitions, or a significant negative correlation should have resulted between these variables as was observed for the error ratio score. Similarly, the use of design production strategies or clustering was not needed by all respondents in the present sample in order to achieve effective performance (as measured by unique designs). Twenty-three percent of the sample employed no or 1 cluster only across the four RFFT trails. Of these individuals, 65% still managed to perform within normal limits using the norms provided by Ruff’s (1988) test manual. Gardner and colleagues (2013) noted that 24% of their sample performed within normal limits on the RFFT without using any strategies. Moreover, the use of strategies did not result in design production within the normal range for all participants in the Gardner and colleagues study. Taken together, these observations suggest that the absence of strategic responding on the RFFT may not be indicative of a “deficit” per se, nor does the ability to use strategies appear to be a special or “unusual cognitive asset” (see Gardner et al., 2013, p. 480). In contrast, the use of strategic responding may be more indicative of a preferred cognitive style or approach to this task favored by some but not all participants (Ross et al., 2003) or some indication of other cognitive abilities (Gardner et al., 2013). While not required in an absolute sense, a growing number of studies indicate that the use of strategic clusters on the RFFT is clearly associated with greater design production and efficiency for a significant portion of participants, so it would seem premature to abandon Lezak and colleagues (2012) interpretation of RFFT performance at present.

These findings highlight the importance of interpreting unique design output in light of other indicators of efficiency (e.g., percentage of designs included in clusters, error ratio scores) to determine whether a respondent employed a thoughtful versus careless test-taking approach. Some participants clearly generated a high number of unique designs at the expense of more repetitions, while others produced designs in a more efficient manner by means of strategy use (which did not correlate with repetitions). Ross and colleagues (2003) reported that the percentage of the total designs incorporated into strategic clusters may be a more useful index than mean cluster size in assessing the extent to which test-takers optimized strategy use. Similar to the present study, Ross and colleagues (2003) found higher reliability coefficients for the percentage of designs used in clusters relative to mean cluster size. Because the number of stimuli (i.e., dots) imposes a limitation on the possible cluster sizes, the percentage of designs in clusters may better assess the variability of strategy use among a sample of participants. For this reason, the mean cluster size index may be less useful for assessing RFFT performance when compared with verbal fluency measures where more cluster size variability can be expected.

Evidence for the divergent validity of the RFFT (as a measure of EF psychometrically distinct from other relevant constructs) was mixed. RFFT indices correlated with select WAIS-III subtest scores and measures of working memory to a modest degree and some interesting patterns emerged. Block Design performance correlated most strongly with design output and significant associations were also observed for qualitative measures of clustering and switching. In contrast, WAIS-III Vocabulary and NAART performance was associated with qualitative indices of strategic responding, but not with unique design production. Similarly, measures of working memory for nonverbal information correlated with strategy use, but not unique design output. Finally,
working memory for verbal information and verbal learning did not correlate with any RFFT index. The one exception being that semantic clustering on the CVLT was related to clustering on the RFFT. Taken together, the aforementioned correlations suggest a pattern of shared variance likely attributed to the nonverbal information processing requirements imposed by these measures. This interpretation is consistent with reports that novel design production on the RFFT is highly sensitive to right-hemisphere lesions and/or imposes substantial right-hemisphere processing requirements relative to the left-hemisphere (Foster, Williams, & Harrison, 2005; Ruff, 1988; Ruff, Allen, Farrow, Niemann & Wylie, 1994; Williams & Harrison, 2003). The finding that estimates of verbal intelligence correlate with RFFT strategy utilization (but not overall design output) is interesting, yet difficult to explain. While it is tempting to speculate higher verbal ability may contribute to a strategic test-taking approach in some manner distinct from contributions to overall performance, the correlations were very modest. Moreover, if verbal ability was somehow facilitating the deployment of RFFT production strategies, it is difficult to explain why similar correlations were not observed for trials 1–5 on the CVLT. Others have suggested the use of strategies may be conceptualized as a more “sophisticated” cognitive approach that may relate to cognitive abilities other than EFs (Gardner et al., 2013). Although the present study does not allow for such conclusions, the data are consistent with the idea that strategy use (conceptualized as more sophisticated responding) may relate to general intelligence, while overall performance is influenced by both “g” and nonverbal (e.g., spatial memory and reasoning) skills in particular. This interpretation should be explored in future research.

Although modest in size, the correlations between novel design production on the RFFT and several well-known measures of EF support the convergent validity of this scoring index as a measure of the EF construct. Moreover, partial correlations provided further evidence that the relationships observed were not solely attributable to shared method variance independent of the construct assessed. For example, the finding that RFFT unique design production correlated with TMT-B (even after controlling for shared variance with TMT-A) suggests this association cannot be explained solely as function of the visual scanning and motor speed requirements imposed or more generally because these procedures were both timed. If this were the case, then similar correlations should have resulted with the other speed-dependent measures (e.g., COWAT indices). In contrast, is more likely these modest correlations are attributable to other task demands shared by EF measures (e.g., cognitive flexibility, set-shifting or novel or non-habitual responding, divergent thinking) or a common relationship with other constructs with which the overall production score appears to be more sensitive when compared with qualitative indices (e.g., clustering and switching). Additionally, results suggest the RFFT may more directly assess EF subsystems required to process nonverbal or visuospatial information. This interpretation is consistent with theories of executive control that stress centralized functions that are fractionated according to domain- or content-specific operations (e.g., Fuster, 1985; Goldman-Rakic, 1987; Shallice & Burgess, 1991). Moreover, the absence of a correlation between overall output on verbal (letter) and design (figural) fluency measures observed in this study and others provide support for this position.

Evidence for the validity of the cluster and switch scores (as measures of EF) was weaker relative to findings observed for novel design production. Correlations between RFFT clustering and switching indices and EF measures were not observed consistently. When present, correlations were modest and comparable in magnitude to those observed for estimates of verbal intelligence and nonverbal working memory performance (i.e., $r = .2$ range) and much lower than the resulting correlation between novel design output and Block Design performance. Moreover, EF measures viewed as highly sensitive to planning ability (e.g., TOH planning index) and strategic utilization in particular (e.g., COWAT Clusters) did not correlate with clustering on the RFFT despite correlating with the overall number of unique designs produced. Although production strategies did correlate with TMT Part B and Stroop Color-Word Interference performance, these relationships were no longer significant after controlling for variance shared with TMT-A and Stroop Color-Naming performance, respectively. Therefore, there was little evidence that strategy use on the RFFT provided a superior or more “precise” assessment of EF (or select facets of this construct) relative to the overall score. Variance shared between the RFFT and other EF tasks was accounted for most effectively by the novel designs produced in the present sample. However, as the present study is one among very few investigations of qualitative scores, it is premature to conclude that strategy use (as assessed by clustering and switching on the RFFT) is not an indication of EF, as results may vary across samples (e.g., healthy persons vs. patient populations) and EF measures examined.

Although practitioners must keep in mind the present findings were generated from young, healthy participants, and some findings are relevant to clinicians and consistent with several widely accepted assumptions guiding the practices of neuropsychologists. Stated generally, first, measures of nonverbal fluency such as the RFFT assess different abilities relative to verbal fluency measures; second, EF is a multifaceted construct that should be assessed comprehensively with several measures and EF measures may relate differentially to other cognitive skills and domains; third, similar to other EF measures, abilities such as intelligence and working memory can influence performance on the RFFT; finally, although evidence for reliability exists, qualitative scoring indices (e.g., cluster and switch scores) for the RFFT must be interpreted with caution as evidence for validity (as a measure of EF) is modest at present.

The present study is among few examining the reliability and validity of RFFT scores, and qualitative scores in particular. This study has several strengths, including a large sample size and the use of a relatively homogenous sample of persons free from
neurological disease. In addition, the present study included several raters to assess reliability and several measures of EF, as well as estimates of intelligence, working memory, and verbal learning. However, the present study was not without limitations.

The present investigation utilized a convenience sample of healthy college undergraduates who were predominantly young, Caucasian females and not representative of the population at large. Therefore, these results may not generalize to samples that differ on demographic variables known to affect neuropsychological test performance (e.g., age) or samples drawn from patient populations. More research is needed on the psychometric properties of the RFFT and additional normative data using patients and samples from underrepresented groups (e.g., ethnic minorities and older adults).

The present investigation employed a short testing interval (i.e., 45.2 days on average) which admittedly is not a stringent examination of score stability, but this provided useful information about the striking practice effects that occur between short testing intervals. Practice effects on measures administered over short intervals are highly relevant in cases where frequent repeated assessments are employed (e.g., post-acute rehabilitative settings — see Lynch, 1990; and forensic examinations — see Putnam, Adams, & Schneider, 1992). These results may have bearing on such situations, as an average gain of 15 designs was observed upon retesting.

Although the present study sought to assess several putative domains of EF, several measures (e.g., Cognitive Estimation Test and Multiple Errands Test) were not included because of time constraints. Therefore, it is possible RFFT scores may relate differentially to other tests sensitive to EF components not assessed in the present investigation. Also, the present investigation did not include a comprehensive measure of intelligence but instead relied on select WAIS-III subtests known to correlate highly with IQ composite scores. Given that RFFT correlated substantially with Block Design performance relative to EF measures, further examination of the contribution of intelligence seems warranted. For example, it may be that RFFT is better conceptualized as another measure of performance IQ or perhaps fluid intelligence given its high association with Block Design performance relative to EF tasks.

The focus of the study was to examine the association between RFFT, EF, and other relevant constructs, namely estimates of intelligence, working memory, and verbal learning. Future studies should include additional measures to better assess the contribution of other cognitive abilities (e.g., visual spatial skills) to effective performance on this task. This suggestion is in keeping with imaging studies (e.g., PET) that demonstrate blood flow correlates of RFFT performance in frontal and parietal areas (Woo et al., 2010).

Future studies should examine the RFFT using other research paradigms (e.g., dual-task methodology) to better isolate and elucidate the task demands imposed by this measure. Additionally, the literature on the RFFT would be enhanced by additional factor analytic research given the dearth of such studies involving this measure. The modest correlations observed among EF tasks support a multidimensional view of the EF construct which is in keeping with several theoretical positions (for reviews, see Chan et al., 2008; Lezak et al., 2012). Factor analytic studies may better elucidate the patterns of shared variance among this complex set of measures. Finally, more information is needed concerning the external validity of the RFFT (and other EF tasks) to better determine how such test scores might inform intervention planning and other recommendations by practitioners.

References


