The Pillbox Test: An Ecological Measure of Executive Functioning and Estimate of Medication Management Abilities

Andrea L. Zartman1,*, Robin C. Hilsabeck2,3, Charles A. Guarnaccia4, Andrew Houtz5

1Department of Psychology, University of North Texas, Denton, TX, USA
2University of Texas Health Science Center, San Antonio, TX, USA
3South Texas Veterans Health Care System, San Antonio, TX, USA
4Department of Psychology, University of North Texas, Denton, TX, USA
5Cognitive Solutions, Fort Worth, TX, USA

*Corresponding author at: VA North Texas Health Care System, Mental Health 4500 S. Lancaster Rd (116A), Dallas, TX 75216, USA. Tel.: +1-214-857-0534; fax: +1-214-857-0531.
E-mail address: andrea.zartman@va.gov (A.L. Zartman).

Accepted 11 February 2013

Abstract

Studies indicate that executive functioning (EF) is a strong predictor of everyday function. However, assessment can be problematic as no single standardized instrument is known to measure all EF domains simultaneously. Thus, the Pillbox Test was developed as a new measure tapping four EF factors through the real-time assessment of medication management, a complex instrumental activity of daily living. The Pillbox Test showed good criterion-related validity and was effective in differentiating graduated levels of executive dysfunction between a mixed neurological group, medical control group, and healthy community control group. This test also showed good convergent validity as it correlated significantly in expected directions with established EF measures in all four of selected EF domains and the Direct Assessment of Functional Status. Finally, a receiver operator characteristic curve found a sensitivity of 75% and specificity of 87.5%, suggesting that the Pillbox Test is a promising new ecological measure of EF.

Keywords: Assessment; Executive functions; Test construction

Introduction

Executive function (EF) is a complex construct that is one of the most useful neuropsychological predictors of everyday functioning (Bell-McGinty, Podell, Franzen, Baird, & Williams, 2002; Boyle, Paul, Moser & Cohen, 2004; Goverover, 2004; O’Bryant et al., 2011; Penades et al., 2010; Sadek, Stricker, Adair, & Haaland, 2011). Although many theoretical models of EF exist (Chan, Shum, Touloumpoulou, & Chen, 2008), one of the most widely accepted is the Lezak’s (1995) four-factor model. This model considers EF to be a multidimensional construct composed of purposive action/self-regulation, planning/attention, volition/inhibition, and effective performance/self-monitoring (Lezak, Howieson, & Loring, 2004).

The purposive action/self-regulation factor, otherwise known as complex purposive behavior, refers to one’s ability to initiate, maintain, and/or adjust behavior when it does not provide the desired outcome. This skill is particularly taxed when one must produce non-routine behaviors or integrate past knowledge with new learning (Lamar, Zonderman, & Resnick, 2002; Lezak et al., 2004). The planning/attention factor refers to one’s ability to sustain attention, consider alternatives, reason sequentially and hierarchically, and demonstrate sound strategy formulation (Duke & Kasznia, 2000; Lamar et al., 2002; Lezak et al., 2004). The volition/inhibition factor refers to an individual’s ability to formulate goals and intentions, with deficits noted when one’s intentions no longer guide behavior (Duke & Kasznia, 2000; Grigsby, Kaye, Baxter, Shetterly, & Hamman, 1998; Lezak et al., 2004). The final factor, effective performance/self-monitoring, also known as meta-cognition, refers to the ability...
to adequately observe one’s own behavior, evaluate task difficulty in relation to personal strengths/weaknesses, choose appropriate strategies for task completion, and recognize/correct errors in an environmental context (Goverover, 2004; Lamar et al., 2002; Lezak et al., 2004; Malloy, Cohen, & Jenkins, 1998).

Persons with executive dysfunction typically manifest deficits in more than one of these four areas (Bryan & Luszcz, 2000; Strauss, Sherman, & Spreen, 2006). Thorough assessment of these deficits can be problematic as no single standardized instrument is known to measure all four EF domains simultaneously. As such, one must carefully choose several measures to achieve this goal (Brauer-Boone, Ponton, Gorsuch, Gonzalez & Miller, 1998; Lamar et al., 2002; Miyake, Friedman, Emerson, Witzki, & Howerton, 2000; Strauss et al., 2006). This is an important clinical consideration as reliance on one EF measure is unlikely to accurately characterize one’s overall EF abilities (Brauer-Boone et al., 1998; Lamar et al., 2002; Miyake et al., 2000; Strauss et al., 2006).

To address the lack of an existing measure assessing all four EF factors, a new process-oriented, performance-based instrument, the Pillbox Test, was designed which evaluates medication management skills. Medication management was chosen as the basis of this new EF measure because it is considered one of the most complex instrumental activities of daily living (IADLs; Burdick et al., 2005; Park, 1999; Stilley, Bender, Dunbar-Jacob, Sereika, & Ryan, 2010) and successful performance requires intact functioning in all four of Lezak’s (1995) EF domains. Additionally, poor medication management not only contributes to the therapeutic failure of disease management but also to the wastage of healthcare resources. For example, non-compliance of antihypertensive medication is associated with increased doctor office visits, increased ER visits, and increased hospitalizations with longer stays (Bennett, Saue, & Shaw, 2005; Salas et al., 2001; Stilley et al., 2010). Furthermore, poor medication management may signal impairment in other cognitive domains, as well as decreased safety for independent living (Burdick et al., 2005).

The cognitive abilities contained within the purposive action/self-regulation domain are most influential on medication adherence as problems with mental flexibility, stimulus-bound behavior, perseveration, suboptimal search strategies, concrete thinking, and the inability to multitask have been associated with incorrect medication administration (Duke & Kaszniak, 2000; Lamar et al., 2002; Lezak et al., 2004; Miyake et al., 2000). As space on medication labels is limited and instructions are generally vague, disjointed, and poorly organized, individuals must rely on making inferences regarding information not explicitly stated (Hayes, 1998; Kendrick & Bayne, 1982; McGraw & Drennan, 2001; Park, Willis, Morrow, Diehl, & Gaines, 1994; Willis et al., 1999). Several studies reveal that both cognitively intact and cognitively impaired individuals have difficulty making inferences from medical information, although those with cognitive impairment make significantly more errors (Hayes, 1998; McGraw & Drennan, 2001; Park et al., 1994; Zahr, 1999). Kendrick and Bayne (1982), for example, found 29% of their cognitively intact elderly sample demonstrated a different understanding of medication administration than the written label. Additionally, only 22% of their sample demonstrated correct administration knowledge of the instruction “take one tablet every 6 h.” Kripalani and colleagues (2006) showed that 38% of individuals from a large, urban general medical clinic were unable to identify all of their medications in spite of being able to look at the bottle, label, and pills themselves.

Within the planning/attention EF domain, problems are seen in the preplanning component, often due to a disorganized approach to tasks, which has been related to difficulty performing IADLs, especially medication management (Duke & Kaszniak, 2000; Lamar et al., 2002; Lezak et al., 2004). In taking medication, the person must plan to carry out the appropriate behaviors at the correct time, integrate medication schedules, and observe any special instructions such as food restrictions (Gould, 1999; Park, 1999; Park et al., 1994). Within the general relationships noted above, specific to medication management, the planning/attention factor is also considered a mediator for the volition/inhibition factor by influencing the efficiency and completeness with which one gathers medical information. For example, Isaac and Tamblyn (1993) found the inability to independently complete medication administration, and the number of errors was related to both slowed visual scanning speed and poor initial encoding of information. Additional medication administration problems related to volition/inhibition deficits include difficulty inhibiting the automatic and habitual use of old medication directions. This disinhibition leads to decreased effortful processing and results in the perseveration of previously learned material (Isaac & Tamblyn, 1993; Park et al., 1994).

This is an important clinical consideration when a prescription is changed or a new medical regimen is implemented.

Finally, impairments in the effective performance/self-monitoring domain, such as increased reliance on routine, resistance to change, poor insight into one’s abilities, and environmental dependency/indifference negatively contribute to medication compliance (Gould, 1999; Isaac & Tamblyn, 1993; Park et al., 1994; Willis et al., 1999). For successful medication administration, one must be able to monitor both medication performance and physiological changes that may signal a medical complication. This requires adequate working memory based on feedback received from using (or not using) medication. Additionally, patients must successfully communicate ongoing effectiveness or problems to medical providers and/or family (Gould, 1999; Park, 1999; Park et al., 1994).
Thus, the purpose of the current study was to examine the construct validity of a new measure of EF, the Pillbox Test, which addresses all four of Lezak’s (1995) factors through the real-time assessment of medication management. The following three hypotheses were examined: (1) individuals with neurological impairment will perform worse than those with cardiovascular disease who will perform worse than healthy community controls on traditional and performance-based measures of EF, including the Pillbox Test; (2) the Pillbox Test will correlate moderately with variables from widely established EF measures in each of the four domains purported by Lezak; and (3) the Pillbox Test will demonstrate good sensitivity and specificity in individuals with neurologic impairment and healthy controls.

Method

Participants

Potential participants were recruited into one of three groups: a mixed neurological group (NEURO), a medical control group (MED), and a healthy community control group (CON). The NEURO group was referred by physicians from an outpatient memory clinic at a county hospital or neurorehabilitation day program at a for-profit medical center. The MED group was recruited via flyers posted in a cardiovascular rehabilitation clinic, and the CON group was recruited following presentations on aging at community senior centers. Participants were excluded if English was not their primary language \( n = 2 \), they were unable to read items on the Reitan Indiana Aphasia Screening Test \( \text{Reitan & Wolfson, 1993;} n = 1 \), or vision and hearing were not adequate for study purposes \( n = 2 \). An additional two people declined to complete testing after an interview. Thus, the final sample consisted of 120 of 127 referrals.

The NEURO group was composed of 40 participants (20 men and 20 women) of which approximately 65% were Caucasian, 30% were African American, and 5% were of other ethnicities. Nine participants were previously diagnosed with Alzheimer’s disease (AD), 13 were diagnosed with Vascular Dementia (VaD), and 18 had incurred stroke within 6 months prior to study participation. Diagnoses of AD and VaD were established by a combination of the referring neurologist’s opinion and previous neuropsychological assessment. These diagnoses were not based on tests used in the current study. None of these individuals were on memory-enhancing drugs at the time of testing. The mean age was 68.63 years \( SD = 8.08 \), and the average education was 12.30 years \( SD = 2.86 \). The mean number of medications consumed by this group as verified by the chart review was 6.88 \( SD = 5.44 \).

The MED group also consisted of 40 participants (20 men and 20 women) of which approximately 88% were Caucasian, 10% were African American, and 2% were of other ethnicities. People with cardiovascular disease were chosen as the medical control group, since mild executive dysfunction has been found in patients with cardiovascular disease \( \text{Bennett et al., 2005; Newman et al., 2005} \), which may contribute to medical non-compliance often seen within this population \( \text{Barclay, Weiss, Mattis, Bond, & Blass, 1988} \). All MED participants had either experienced a heart attack within the past year or been diagnosed with a chronic cardiovascular disorder such as congestive heart failure. Additionally, all MED participants denied a prior diagnosis of cognitive impairment or any other conditions that might affect cognition. The mean age was 69.13 years \( SD = 8.12 \), and the average education was 13.91 years \( SD = 3.34 \). This group was consuming an average of 7.08 medications \( SD = 4.12 \) per chart review.

The CON group representing “normal aging” consisted of 40 participants (17 men and 23 women) of which approximately 94% were Caucasian, 3% were African American, and 3% were of other ethnicities. All CON participants denied a prior diagnosis of cognitive impairment or any other conditions that might affect cognition and were living independently. The mean age was 64.10 years \( SD = 7.32 \), and the average education was 15.48 years \( SD = 2.70 \). This group was consuming an average of 3.25 medications \( SD = 2.64 \), which was determined by reviewing their medication bottles during the study session. None of these medications were known to significantly impact cognitive functioning.

Materials

Neuropsychological measures. Based on a literature review, measures were chosen to represent each of Lezak’s (1995) four factors of EF. Traditional neuropsychological measures in addition to a performance-based measure were utilized and are listed under the factor they represent below, along with references supporting their inclusion under each factor. Raw scores were used as dependent variables to facilitate comparisons, since demographically corrected scores for the Pillbox Test have not been derived. Likewise, given the preliminary nature of this work, individual test/subtest scores were examined rather than domain or factor scores in order to gain a deeper understanding of the characteristics of the Pillbox Test.
Volition/Inhibition Factor (Brauer-Boone et al., 1998; Bryan & Luszcz, 2000; Shimamura, 1994; Strauss et al., 2006):

- Stroop Color and Word Interference Test (Stroop; Golden & Freshwater, 2002) Color/Word raw score,
- Geriatric Depression Scale-Short Form (GDS-SF; Yesavage et al., 1983) total raw score.

Planning/Attention Factor (Brauer-Boone et al., 1998; Carlson, Fried, Xue, & Brant, 1999; Golden & Freshwater, 2002; Lamar et al., 2002; Ready, Stierman, & Paulsen, 2001; Senanarong et al, 2005; Strauss et al., 2006):

- Rey–Osterrieth Complex Figure Test (Rey; Osterrieth, 1944) Copy raw score.
- Rey Executive Planning raw score, which was determined using the organizational scoring system of the Rey Complex Figure by Savage and colleagues (1999). This scoring system qualitatively evaluates the large rectangle, diagonal cross, vertical midline, horizontal midline, and vertex of triangle on the right.
- Trail Making Test (TMT; Reitan & Wolfson, 1993) Part A total time in seconds.
- Stroop Word and Color raw scores.
- Clock Draw Test (CDT) total raw score from the 7-min screen (Solomon et al., 1998).
- Controlled Oral Word Association Test (COWAT; Benton, 1968) total raw score.
- Design Fluency Test (DFT; Jones-Gotman & Milner, 1977) novel output raw score.

Purposive Action/Self-Regulation Factor (Boyle et al., 2004; Hanks, Rapport, Millis, & Deshpande, 1999; Kibby, Schmitter-Edgecombe, & Long, 1998; Lamar et al., 2002; Lezak et al., 2004; Ready et al., 2001; Salthouse, Atkinson, & Berish, 2003):

- Rey Delayed Recall raw score,
- TMT Part B total time in seconds,
- Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997) Similarities subtest raw score,
- Wisconsin Card Sorting Test (WCST; Grant & Berg, 1948; Heaton, Chelune, Talley, Kay, & Curtis, 1993) total errors and conceptual responses raw scores,
- Direct Assessment of Functional Status (DAFS; Loewenstein & Bates, 1992) subscale scores. The DAFS is a behaviorally based measure of IADLs developed and validated in a neurologically impaired sample (Loewenstein et al., 1989). This assessment is composed of several subtests in which the participant is asked to complete a particular IADL while the examiner observes. These IADL activities consist of locating a number in a phone book, dialing a telephone, identifying traffic signs, reading a clock, prepare a letter for mailing, counting currency, writing a check, balancing a checkbook, and recording a phone message. Four subscale scores, Communications, Transportation, Financial, and Message, and one Total Scale score are derived.

Effective Performance/Self-monitoring Factor (Bryan & Luszcz, 2000; Ferrer-Caja, Crawford, & Bryan, 2002; Hanks et al., 1999; Salthouse et al., 2003; Strauss et al., 2006):

- COWAT perseverative responses raw score,
- DFT perseverative responses raw score,
- WCST number of categories, perseverative responses, and perseverative errors raw scores.

The Pillbox Test. The Pillbox Test consists of a pillbox and five pill bottles. The pillbox contains four rows labeled as “Breakfast,” “Lunch,” “Dinner,” and “Bedtime” and seven columns labeled for each day off the week, “Sunday” through “Saturday.” The five pill bottles have standardized administration labels and contain colored beads resembling the approximate size of common aspirin or antihypertensive medications as these were the two most commonly prescribed types of medications according to chart reviews of 50 county hospital memory clinic patients (Brose-Zartman & Houtz, 2003). The average age of
these patients was 75 years ($SD = 6.45$). Thirty-eight percent were diagnosed with Dementia of the Alzheimer’s type, 25% were diagnosed with VaD, 8% were diagnosed with dementia of other etiologies, 12% were diagnosed with mild cognitive impairment, and 17% were not diagnosed with any cognitive disorder. Chart reviews also revealed that the average number of medications taken was five, which was congruent with findings of Burdick and colleagues (2005).

The five most common directions were also determined from chart reviews and consultation with hospital pharmacy staff. Medication directions printed on pill bottle labels are as follows: “one tablet daily at bedtime” (orange pills), “one tablet daily in the morning” (blue pills), “one tablet three times a day” (yellow pills), “one tablet twice a day with breakfast and dinner” (green pills), and “one tablet every other day” (red pills). These directions require the participant to make inferences based on vague medication labels. Participants were given credit as long as they did not violate the instructions. Omission errors were counted if the pill was missing from the proper container and not better identified as a misplaced movement error. Misplaced movement errors were pills placed in the correct day and quantity but not according to exact direction. For example, if someone placed an orange pill (one tablet at bedtime) in the breakfast compartment versus the bedtime compartment. In reality, the person would still be taking the correct daily dose but at the improper time which could potentially lead to side effects or other problems. In another example, as long as one yellow pill was placed into compartments three times a day, it did not matter whether the compartments were breakfast, lunch, and dinner or breakfast, dinner, and bedtime and/or if one red pill was placed in the breakfast compartment on Monday but in the dinner compartment on Wednesday as these directions were not time-specific. As such, this would reflect no errors. Commission errors were counted as excess pills (e.g., four yellow pills per day instead of three).

Participants are first asked to read aloud each medication label to verify adequate reading ability. Participants are then instructed to organize the medication in the pillbox as they normally would for 1 week. No further direction or feedback is provided by the examiner for the remainder of the test. This lack of structure was purposeful as EF is most taxed in ambiguous situations (Lamar et al., 2002; Lezak et al., 2004). Participants are allowed 5 min to complete this task. Throughout the test, the examiner records behavioral observations such as the approach to the task. After the participant completes the test, the examiner records the number and type of beads placed into in each pillbox compartment.

This test yields several scores. A straight pass/fail designation is determined by three or more total errors of any type on the task. This criterion was established through consultation with family practice physicians and gerontologists who indicated that this was the number of medication administration errors that would commonly cause medical complications. Additional scores include:

- sum of yellow pill commission and omission errors (one tablet three times a day)
- sum of green pill commission and omission errors (one tablet twice a day with breakfast and dinner)
- sum of blue pill commission and omission errors (one tablet daily in the morning)
- sum of orange pill commission and omission errors (one tablet daily at bedtime)
- sum of red pill commission and omission errors (one tablet every other day)
- number of misplaced movement errors = the sum of number of pills placed in incorrect compartments
- number of total errors = the sum of total omission errors, commission errors, and misplacement errors
- total time to complete the task

Although Pillbox Test scores do not map perfectly onto Lezak’s four EF factors, inferences about these abilities can be made based on successful completion of the task and/or difficulties in specific aspects of performing the task. For the volition/inhibition factor, one must inhibit knowledge of one’s own medication regimen and commence the task. For the planning/attention factor, one must employ an efficient strategy to carry out the task while sustaining attention (i.e., time to complete the task). For the effective performance/self-monitoring factor, one must monitor when to stop placing pills into the pillbox based on administration directions (i.e., misplacement errors), and for the purposive action/self-regulation factor, one must make inferences from instructions, reason alternative solutions, mentally shift between medication instructions, and demonstrate working memory and motor programming (i.e., total errors, omission errors). Behavioral observations, although not quantified, are equally important in identifying executive dysfunction, as well as other cognitive and motor impairments that may affect performance on the Pillbox Test. For example, the complexity of instructions differs and errors on particular pills (e.g., green pill errors) may provide important information about the capacity of the examinee.
Procedure

IRB approval was obtained at the university and both hospitals from which participants were recruited. Written informed consent was obtained from all participants at the beginning of the evaluation session. All testing sessions were conducted in the morning to control for time of day effects on cognitive ability, as this is a common concern for the elderly and neurological populations (Salthouse et al., 2003). Participants completed an intake interview to collect demographic data during which the Reitan Indiana Aphasia Screening Test (Reitan & Wolfson, 1993) was administered to provide a brief check of inclusion/exclusion criteria (e.g., visual acuity, ability to read). Participants meeting inclusion/exclusion criteria were then administered the above measures in the following order according to the standardized instructions: Rey Copy, TMT, WAIS-III Similarities, CDT, COWAT, Rey Delayed Recall, DFT, Stroop, WCST, GDS-SF, DAFS, and Pillbox Test.

Data Analysis

Chi-squared tests and one-way analyses of variance (ANOVs) were performed to determine if there were significant group differences in demographic variables. If significant group differences in age, education, gender, and/or ethnicity were identified, then the variables were used as covariates in subsequent analyses. One-way univariate and multivariate ANOVAs were used to test hypothesis 1 that significant group differences would be found on all traditional and performance-based measures. If significant group differences were found, Tukey’s LSD post hoc tests were conducted to identify which groups differed significantly from one another. For variables that violated assumptions of normality, results were confirmed using non-parametric analyses, although effects of covariates could not be examined in these cases. Spearman’s rho correlation coefficients were utilized to investigate relationships between the Pillbox Test and demographic variables and to test hypothesis 2 that Pillbox Test variables would correlate significantly with existing EF measures which tap each of Lezak’s (1995) four factors. Given the large number of variables (i.e., 28), only traditional EF measures that were significantly different among groups and Pillbox variables with p-values of < .001 in differentiating among the groups were analyzed. To test hypothesis 3, a receiver operator characteristic (ROC) curve analysis was conducted to determine the sensitivity and the specificity of the pass/fail criterion by examining the number of total errors made on the Pillbox Test. A p-value of ≤ .05 was utilized for all analyses; corrections for multiple comparisons were not employed give the preliminary nature of this work.

Results

Significant group differences were found in age, $F(2,117) = 4.98, p = .008$, education, $F(2,117) = 11.37, p < .001$, ethnicity, $\chi^2(6) = 15.76, p = .02$, and the number of medications consumed, $F(2,117) = 9.91, p < .001$, but not gender $\chi^2(2) = 0.60, p = .74$. As shown in Table 1, the CON group was significantly younger and taking significantly fewer medications than the NEURO and MED groups, which did not differ significantly from each other. The CON and MED groups were significantly more educated and comprised of a significantly greater proportion of Caucasian participants than the NEURO group but did not differ significantly from each other.

Controlling for age, education, and ethnicity in the omnibus one-way and multivariate ANOVAs, significant differences were found among the three groups on traditional EF measures including the Rey, $F(6,224) = 2.48, p = .02$, TMT, $F(4,226) = 6.12, p < .001$, CDT, $F(2) = 5.11, p = .007$, Stroop, $F(6,224) = 4.18, p = .001$, WCST, $F(10,220) = 2.52, p = .007$, and GDS-SF, $F(2) = 8.41; p < .001$, but not on the COWAT, $F(4,226) = 1.22, p = .30$, or DFT, $F(4,226) = 0.54, p = .71$. As seen in Table 2, post hoc analyses showed that the NEURO group performed significantly worse than the CON group.

Table 1. Demographic characteristics by group

<table>
<thead>
<tr>
<th></th>
<th>NEURO (n = 40)</th>
<th>MED (n = 40)</th>
<th>CON (n = 40)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
</tr>
<tr>
<td>Age (years)</td>
<td>68.63a</td>
<td>8.08</td>
<td>69.13b</td>
<td>8.12</td>
</tr>
<tr>
<td>Education (years)</td>
<td>12.30a,b</td>
<td>2.86</td>
<td>13.91b</td>
<td>3.34</td>
</tr>
<tr>
<td># medications</td>
<td>6.68a</td>
<td>5.44</td>
<td>7.08b</td>
<td>4.12</td>
</tr>
<tr>
<td>% Men</td>
<td>50</td>
<td></td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>% Caucasian</td>
<td>65</td>
<td></td>
<td>87.5</td>
<td></td>
</tr>
</tbody>
</table>

Notes: NEURO = mixed neurological group; MED = medical control group; CON = healthy community control group.

Groups with the same subscript are significantly different from one another.

313

Table 2. Means and standard deviations (SD) of EF measures by group

<table>
<thead>
<tr>
<th></th>
<th>NEURO (n = 40)</th>
<th>MED (n = 40)</th>
<th>CON (n = 40)</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Rey copya</td>
<td>26.16a,b</td>
<td>7.99</td>
<td>31.33a</td>
<td>5.29</td>
<td>33.11b</td>
</tr>
<tr>
<td>Rey recall</td>
<td>10.84a</td>
<td>6.18</td>
<td>14.68a</td>
<td>6.32</td>
<td>17.79a</td>
</tr>
<tr>
<td>Rey execa</td>
<td>11.25a,b</td>
<td>2.57</td>
<td>12.81a</td>
<td>1.60</td>
<td>13.38b</td>
</tr>
<tr>
<td>Trails Aa</td>
<td>77.35a,b</td>
<td>50.04</td>
<td>47.23a</td>
<td>19.91</td>
<td>35.98b</td>
</tr>
<tr>
<td>Trails Ba</td>
<td>175.15a</td>
<td>80.94</td>
<td>119.03a</td>
<td>68.80</td>
<td>82.95a</td>
</tr>
<tr>
<td>CDT total</td>
<td>4.93a,b</td>
<td>1.31</td>
<td>5.83a</td>
<td>1.24</td>
<td>5.75a</td>
</tr>
<tr>
<td>Similaritiesa</td>
<td>15.40a</td>
<td>7.03</td>
<td>19.85a</td>
<td>6.18</td>
<td>24.45a</td>
</tr>
<tr>
<td>COWAT total</td>
<td>23.38</td>
<td>13.73</td>
<td>29.08</td>
<td>9.76</td>
<td>33.47</td>
</tr>
<tr>
<td>COWAT psv</td>
<td>0.58</td>
<td>1.04</td>
<td>0.48</td>
<td>1.22</td>
<td>0.58</td>
</tr>
<tr>
<td>DFT novel</td>
<td>4.25</td>
<td>3.14</td>
<td>5.20</td>
<td>3.92</td>
<td>6.20</td>
</tr>
<tr>
<td>DFT psv</td>
<td>3.12</td>
<td>2.91</td>
<td>3.10</td>
<td>3.32</td>
<td>3.25</td>
</tr>
<tr>
<td>Stroop word</td>
<td>63.88a,b</td>
<td>22.43</td>
<td>75.90a</td>
<td>17.38</td>
<td>92.40a</td>
</tr>
<tr>
<td>Stroop color</td>
<td>43.43a</td>
<td>17.00</td>
<td>56.15a</td>
<td>12.74</td>
<td>63.40a</td>
</tr>
<tr>
<td>Stroop CW</td>
<td>19.50a</td>
<td>8.84</td>
<td>28.58a</td>
<td>9.28</td>
<td>32.85a</td>
</tr>
<tr>
<td>WCST # cata</td>
<td>2.85</td>
<td>2.24</td>
<td>3.73a</td>
<td>2.15</td>
<td>5.18a</td>
</tr>
<tr>
<td>WCST errora</td>
<td>53.63a</td>
<td>26.94</td>
<td>44.53a</td>
<td>23.97</td>
<td>24.00a</td>
</tr>
<tr>
<td>WCST psv respa</td>
<td>43.63a</td>
<td>35.91</td>
<td>26.60a</td>
<td>19.65</td>
<td>13.40a</td>
</tr>
<tr>
<td>WCST psv errorsa</td>
<td>37.20a</td>
<td>27.64</td>
<td>23.60a</td>
<td>16.00</td>
<td>12.08a</td>
</tr>
<tr>
<td>WCST concepta</td>
<td>46.85a</td>
<td>20.03</td>
<td>57.70a</td>
<td>19.44</td>
<td>67.10a</td>
</tr>
<tr>
<td>GDS-SF total</td>
<td>9.30a,b</td>
<td>5.71</td>
<td>4.78a</td>
<td>4.42</td>
<td>3.43b</td>
</tr>
</tbody>
</table>

Notes: NEURO = mixed neurological group; MED = medical control group; CON = healthy community control group. Rey = Rey Complex Figure Test; Rey exec = executive plan score; Trails = Trail Making Test; CDT = Clock Draw Test; COWAT = Controlled Oral Word Association Test; COWAT psv = perseverative responses; DFT = Design Fluency Test; DFT novel = novel output score; DFT psv = perseverative responses; Stroop = Stroop Word and Color Test; Stroop CW = Color Word subscale; WCST = Wisconsin Card Sort Test; WCST # cat = number of categories; WCST error = total errors; WCST psv resp = number of perseverative responses; WCST psv errors = number of perseverative errors; WCST concept = conceptual responses; GDS-SF = Geriatric Depression Scale-Short Form. Groups with the same subscript are significantly different from one another.

*Findings did not change when non-parametric analyses were utilized to account for unequal variances.

group on the Rey Copy, Delayed Recall, and Executive Plan Scores; TMT Parts A and B; CDT total score; Similarities; Stroop Word, Color, and Color/Word; WCST number of categories, total errors, perseverative responses and errors, and conceptual responses; and GDS-SF. The NEURO group also performed significantly worse than the MED group on all of these measures except WCST total errors. The MED group performed significantly worse than the CON group on Rey Delayed Recall, TMT-Part B, Similarities, Stroop Color and Color/Word, and WCST number of categories, total errors, perseverative responses and errors, and conceptual responses.

On performance-based measures, significant group differences were found on both the DAFS, $F(10,220) = 8.92, p < .001$, and the Pillbox Test, $F(16,214) = 3.36, p < .001$, after correcting for age, education, and ethnicity. As shown in Table 3, the NEURO group performed significantly worse than the MED and CON groups on the DAFS Total score and the Communication, Financial, and Message subscale scores, whereas the MED group performed significantly worse than the CON group on the DAFS Message subscale only. On the Pillbox Test, the NEURO group performed significantly worse than the other two groups on total errors, misplacement errors, total time, yellow pill errors, green pill errors, blue pill errors, and orange pill errors. The MED group performed significantly worse than the CON group on green pill errors only. When considering the pass/fail criterion, there were significant differences among the groups, $\chi^2(2) = 29.27, p < .001$, with a significantly greater portion of the NEURO group failing the Pillbox Test (75%) compared with both the MED (43%) and CON (15%) groups, $\chi^2(1) = 8.72$ and $p = .003$ and $\chi^2(1) = 29.09$ and $p < .001$, respectively, and a significantly greater portion of the MED group failing compared with the CON group, $\chi^2(1) = 7.34, p = .007$.

Significant correlations were found between the primary Pillbox Test scores and age, education, and ethnicity but not gender. Correlations with age ranged from $.20$ (number of green pill errors) to $.34$ (total time to complete task), with education from $-.17$ (number of yellow pill errors) to $-.31$ (total time to complete task), and with ethnicity from $.21$ (total time to complete task) to $.38$ (total number of errors). The Pillbox variables correlated significantly in the expected direction with most neuropsychological measures, including the DAFS (Table 4). Correlations between the Pillbox variables and established EF measures ranged from $-.004$ (i.e., number of green pill errors and GDS-SF) to $-.60$ (i.e., total errors and DAFS Communication). For the volition/inhibition factor, the strongest correlation was seen between total time and Stroop Color/Word ($r = -.46$). For the planning/attention factor, the strongest correlations were found between total errors and Rey...
Table 3. Means and standard deviations (SD) of performance based measures by group

<table>
<thead>
<tr>
<th></th>
<th>NEURO (n = 40)</th>
<th>MED (n = 40)</th>
<th>CON (n = 40)</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>DAFS total</td>
<td>55.90, b</td>
<td>7.94</td>
<td>62.85, b</td>
<td>2.66</td>
<td>64.60, b</td>
</tr>
<tr>
<td>DAFS comm</td>
<td>11.18, b</td>
<td>2.23</td>
<td>13.65, b</td>
<td>0.62</td>
<td>13.68, b</td>
</tr>
<tr>
<td>DAFS tran</td>
<td>11.68</td>
<td>1.58</td>
<td>12.20</td>
<td>0.97</td>
<td>12.60</td>
</tr>
<tr>
<td>DAFS fin</td>
<td>15.98, b</td>
<td>3.03</td>
<td>18.12</td>
<td>1.14</td>
<td>18.52, b</td>
</tr>
<tr>
<td>DAFS mes</td>
<td>1.97, a</td>
<td>1.19</td>
<td>3.33</td>
<td>0.89</td>
<td>3.85</td>
</tr>
<tr>
<td>Pill total</td>
<td>29.70, a,b</td>
<td>30.25</td>
<td>8.57</td>
<td>13.48</td>
<td>130.5</td>
</tr>
<tr>
<td>Pill me</td>
<td>13.25, a,b</td>
<td>25.32</td>
<td>3.38</td>
<td>5.68</td>
<td>190.55</td>
</tr>
<tr>
<td>Pill time</td>
<td>259.75, a,b</td>
<td>62.15</td>
<td>204.35</td>
<td>77.40</td>
<td>190.55</td>
</tr>
<tr>
<td>Pill yellow</td>
<td>10.85, a,b</td>
<td>11.50</td>
<td>2.60</td>
<td>6.86</td>
<td>0.93</td>
</tr>
<tr>
<td>Pill green</td>
<td>9.98</td>
<td>9.75</td>
<td>4.15</td>
<td>7.08</td>
<td>0.88</td>
</tr>
<tr>
<td>Pill blue</td>
<td>2.48, b</td>
<td>4.99</td>
<td>0.33</td>
<td>1.33</td>
<td>0.00</td>
</tr>
<tr>
<td>Pill orange</td>
<td>3.43, a,b</td>
<td>6.10</td>
<td>0.43</td>
<td>1.50</td>
<td>0.40, a</td>
</tr>
<tr>
<td>Pill red</td>
<td>2.98</td>
<td>5.67</td>
<td>0.93</td>
<td>2.73</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Notes: NEURO = mixed neurological group; MED = medical control group; CON = healthy community control group; DAFS = Direct Assessment of Functional Status; DAFS comm = DAFS communication subscale; DAFS tran = DAFS transportation subscale; DAFS fin = DAFS financial; DAFS mes = DAFS message; Pill total = Pillbox total errors; Pill me = Pillbox misplacement errors; Pill time = Pillbox total time; Pill yellow = Pillbox yellow pill errors; Pill green = Pillbox green pill errors; Pill orange = Pillbox orange pill errors; Pill red = Pillbox red pill errors. Groups with the same subscript are significantly different from one another.

Table 4. Correlations of primary Pillbox Test scores with existing measures of EF

<table>
<thead>
<tr>
<th></th>
<th>Total errors</th>
<th>Total time</th>
<th>Yellow pill errors</th>
<th>Green pill errors</th>
<th>Pass/Fail ≥ 3</th>
<th>Pass/Fail ≥ 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volition/Inhibition EF factor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stroop C/W</td>
<td>−0.38***</td>
<td>−0.46***</td>
<td>−0.35***</td>
<td>−0.33***</td>
<td>0.37***</td>
<td>0.40***</td>
</tr>
<tr>
<td>GDS-SF</td>
<td>0.13</td>
<td>0.21*</td>
<td>0.19*</td>
<td>−0.004</td>
<td>−0.28***</td>
<td>−0.25***</td>
</tr>
<tr>
<td><strong>Planning and Attention EF factor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rey copy</td>
<td>−0.57***</td>
<td>−0.36***</td>
<td>−0.56***</td>
<td>−0.42***</td>
<td>0.47***</td>
<td>0.48***</td>
</tr>
<tr>
<td>Rey exec</td>
<td>−0.48***</td>
<td>−0.32***</td>
<td>−0.51***</td>
<td>−0.33***</td>
<td>0.42***</td>
<td>0.43***</td>
</tr>
<tr>
<td>Trails A</td>
<td>0.57***</td>
<td>0.43***</td>
<td>0.48***</td>
<td>0.43***</td>
<td>−0.46***</td>
<td>−0.43***</td>
</tr>
<tr>
<td>Stroop word</td>
<td>−0.24***</td>
<td>−0.26**</td>
<td>−0.18*</td>
<td>−0.19*</td>
<td>0.11</td>
<td>0.16</td>
</tr>
<tr>
<td>Stroop color</td>
<td>−0.37***</td>
<td>−0.38***</td>
<td>−0.30***</td>
<td>−0.36***</td>
<td>0.35***</td>
<td>0.36***</td>
</tr>
<tr>
<td>CDT total</td>
<td>−0.38***</td>
<td>−0.20*</td>
<td>−0.43***</td>
<td>−0.29***</td>
<td>0.35***</td>
<td>0.33***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Purposive Action and Self-Regulation EF factor</strong></th>
<th>Total errors</th>
<th>Total time</th>
<th>Yellow pill errors</th>
<th>Green pill errors</th>
<th>Pass/Fail ≥ 3</th>
<th>Pass/Fail ≥ 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rey recall</td>
<td>−0.38***</td>
<td>−0.40***</td>
<td>−0.40***</td>
<td>−0.29***</td>
<td>0.47***</td>
<td>0.42***</td>
</tr>
<tr>
<td>Trails B</td>
<td>0.56***</td>
<td>0.54***</td>
<td>0.49***</td>
<td>0.49***</td>
<td>−0.56***</td>
<td>−0.55***</td>
</tr>
<tr>
<td>Similarities</td>
<td>−0.52***</td>
<td>−0.33***</td>
<td>−0.44***</td>
<td>−0.46***</td>
<td>0.44***</td>
<td>0.48***</td>
</tr>
<tr>
<td>WCST total error</td>
<td>0.42***</td>
<td>0.31***</td>
<td>0.39***</td>
<td>0.38***</td>
<td>−0.43***</td>
<td>−0.39***</td>
</tr>
<tr>
<td>WCST concept</td>
<td>−0.38***</td>
<td>−0.30***</td>
<td>−0.33***</td>
<td>−0.34***</td>
<td>0.26**</td>
<td>0.26**</td>
</tr>
<tr>
<td>DAFS total</td>
<td>−0.57***</td>
<td>−0.25***</td>
<td>−0.55***</td>
<td>−0.47***</td>
<td>0.54***</td>
<td>0.55***</td>
</tr>
<tr>
<td>DAFS comm</td>
<td>−0.60***</td>
<td>−0.32***</td>
<td>−0.56***</td>
<td>−0.45***</td>
<td>0.40***</td>
<td>0.39***</td>
</tr>
<tr>
<td>DAFS financial</td>
<td>−0.52***</td>
<td>−0.17</td>
<td>−0.47***</td>
<td>−0.41***</td>
<td>0.48***</td>
<td>0.48***</td>
</tr>
<tr>
<td>DAFS message</td>
<td>−0.55***</td>
<td>−0.31***</td>
<td>−0.55***</td>
<td>−0.45***</td>
<td>0.50***</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Effective Performance/Self-Monitoring EF factor</strong></th>
<th>Total errors</th>
<th>Total time</th>
<th>Yellow pill errors</th>
<th>Green pill errors</th>
<th>Pass/Fail ≥ 3</th>
<th>Pass/Fail ≥ 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCST # cat</td>
<td>−0.42***</td>
<td>−0.27**</td>
<td>−0.39***</td>
<td>−0.38***</td>
<td>0.37***</td>
<td>0.35***</td>
</tr>
<tr>
<td>WCST psv resp</td>
<td>0.40***</td>
<td>0.28**</td>
<td>0.39***</td>
<td>0.35***</td>
<td>−0.43***</td>
<td>−0.41***</td>
</tr>
<tr>
<td>WCST psv err</td>
<td>0.43***</td>
<td>0.26**</td>
<td>0.41***</td>
<td>0.38***</td>
<td>−0.44***</td>
<td>−0.42***</td>
</tr>
</tbody>
</table>

Notes: Yellow pill direction = one tablet three times a day; Green pill direction = one tablet twice a day with breakfast and dinner; Pass/Fail ≥ 3 = three or more errors; EF = executive function; Stroop C/W = Stroop Color/Word score; GDS-SF = Geriatric Depression Scale-Short Form total score; Rey = Rey Complex Figure copy score; Rey exec = Rey executive plan score; Trails = Trail Making Test; CDT = Clock Draw Test; WCST = Wisconsin Card Sort Test; WCST concept = number of conceptual responses; WCST # cat = number of categories completed; WCST psv resp = number of perseverative responses; WCST psv err = number of perseverative errors; DAFS = Direct Assessment of Functional Status; DAFS me = DAFS message subscale score.

*p < .05.
**p < .01.
***p < .001.
copy \( r = -.57 \), total errors and TMT Part A \( r = .57 \), and the number of yellow pill errors and Rey copy \( r = -.56 \). For the purposive action/self-regulation factor, the strongest correlations occurred between total errors and DAFS Total, Communication, and Message \( (r = -.57, -.60, \text{ and } -.55, \text{ respectively}) \) and TMT Part B \( r = .56 \), between total time and TMT Part B \( r = .54 \), and between the number of yellow pill errors and DAFS Total, Communication, and Message \( (r = -.55, -.56, \text{ and } -.55, \text{ respectively}) \). For the final factor, effective performance/self-monitoring, the strongest correlations were found between pass/fail and WCST perseverative errors and responses \( (r = -.44 \text{ and } -.43, \text{ respectively}) \), between total errors and WCST perseverative errors \( (r = .43) \), number of categories \( (r = -.42) \), and perseverative responses \( (r = .40) \), and between the number of yellow pill errors and perseverative errors \( (r = .41) \). When examining correlations within each of the three groups separately, significant correlations were found between Pillbox Test variables and measures representing each of Lezak’s four-factor EF except in the CON group where no Pillbox Test scores correlated significantly with any measures of effective performance/self-monitoring (Supplementary Table S1). These findings were largely congruent with expectation in relation to the level of executive dysfunction within each of these patient groups and the psychometric properties of each of the standardized EF measures highlighted in this table.

An ROC analysis indicated that the area under the curve was significant (i.e., 0.82, 95% CI: 0.73–0.92; \( p < .001 \)) when using the total number of Pillbox Test errors to predict membership in the NEURO or CON groups (Figure 1). Two or more total errors on the Pillbox Test provided a sensitivity level of 77.5% and a specificity level of 77.5%, ≥3 total errors provided a 75.0% sensitivity level and an 80% specificity level, ≥4 total errors kept the sensitivity level at 75% and increased the specificity level to 85%, and ≥5 total errors also had a 75% sensitivity level and increased the specificity level slightly to 87.5%. Correlations between existing EF measures and cutscores of ≥3 and ≥5 are shown in Table 4 and reveal moderate correlations in the expected directions on most measures.

Discussion

The purpose of the current study was to examine the construct validity of a new performance-based measure of EF, the Pillbox Test, which was hypothesized to address all four of Lezak’s (1995) EF factors (i.e., purposive action/self-regulation, planning/attention, volition/inhibition, and effective performance/self-monitoring) through the real-time assessment of medication management. The following three hypotheses were examined and generally supported: (1) individuals with neurological
impairment will perform worse than individuals with cardiovascular disease who will perform worse than healthy community controls on traditional and performance-based measures of EF, including the Pillbox Test; (2) the Pillbox Test will correlate moderately with variables from widely established EF measures in each of the four domains purported by Lezak; and (3) the Pillbox Test will demonstrate good sensitivity and specificity in individuals with neurologic impairment and healthy controls.

With regard to hypothesis 1, the NEURO group performed significantly worse than the MED and CON groups on seven of eight Pillbox Test measures, most established EF measures, and four of five DAFS subscales, an existing performance-based measure of IADLs. The MED group performed significantly worse than the CON group on one Pillbox Test variable (i.e., green pill errors), many of the traditional EF measures, and one of the DAFS subscales (i.e., Message). These findings indicate that the Pillbox Test has good criterion-related validity and is effective in differentiating graduated levels of executive dysfunction as the worst performance was seen in the NEURO group, with the MED group demonstrating less difficulty but more than controls, consistent with mild executive dysfunction often noted in cardiovascular samples. An exploratory analysis within the NEURO group comparing patients with and without dementia revealed that patients with dementia performed worse on all Pillbox variables, albeit not always significantly so (data not shown), providing further preliminary support of its ability to differentiate among levels of cognitive impairment.

The Pillbox Test also showed good convergent validity as it correlated significantly in expected directions with established EF measures in all four of Lezak’s EF domains, as well as with DAFS subscales (hypothesis 2). Interestingly, the strongest correlations noted in each factor were related to some of the most commonly used measures in each respective domain: Stroop Word/Color score for the volition/inhibition factor (Strauss et al., 2006), Rey Copy and TMT Part A for the planning/attention factor (Brauer-Boone et al., 1998; Savage et al., 1999), TMT Part B and DAFS Total score for the purposive action/self-regulation factor (Hanks et al., 1999), and WCST perseverative errors, perseverative responses, and number of categories for the effective performance/self-monitoring factor (Kibby et al., 1998). Especially noteworthy are the moderate correlations between Pillbox Test variables, TMT Part B and WCST variables as the latter two tests are frequently considered the “gold standard” in measuring EF. Pillbox Test scores were significantly correlated with the four most sensitive WCST scores for determining executive dysfunction: raw error score, number of categories completed, perseverative raw response score, and perseverative error score (Crawford, Bryan, Luszcz, Obonsawin, & Stewart, 2000; Ready et al., 2001). While the WCST has been criticized for its length and difficulty of administration (Lezak, 1995; Lezak et al., 2004; Strauss et al., 2006), the Pillbox Test could provide similar information in a shorter amount of time and may be less frustrating.

The initial pass/fail designation of three or more total errors on the Pillbox Test was based on consultation with family practice physicians and gerontologists who indicated this as the threshold of errors commonly associated with medical complications. An ROC curve found this criterion resulted in the sensitivity of 75% and the specificity of 80% (hypothesis 3). However, increasing the threshold to five or more total errors resulted in the same sensitivity (75%) but increased the specificity to 87.5%, thus lowering the false positive rate to only 12.5%. When five or more errors were applied as the pass/fail criterion to the current study, significant differences among the NEURO, MED, and CON groups remained.

Theoretically, the Pillbox Test appears to compensate for documented weaknesses of traditional EF measures in several ways. First, while no single standardized instrument is known to measure all EF domains (Brauer-Boone et al., 1998; Lamar et al., 2002; Miyake et al., 2000; Strauss et al., 2006), the Pillbox Test appears to tap the four EF domains purported by Lezak. Thus, by considering Pillbox Test performance both quantitatively and qualitatively, not only can one identify the presence of impairment, but one may also determine which particular EF domain(s) is affected, resulting in the ability to apply specific corrective actions and inform treatment planning.

Other advantages of the Pillbox Test over existing EF measures are its lack of reliance on novelty and the minimal interactions required by the examiner. For these reasons, the Pillbox Test is expected to have higher test–retest reliability than other standardized measures of EF which are commonly low due to benefits of prior exposure (Miyake et al., 2000; Strauss et al., 2006) and greater ecological validity as the examiner does not function as the patient’s “frontal lobes” by employing corrective actions such as those applied during Trails-B (Manchester, Priestley, & Jackson, 2004; Richardson, Nadler, & Malloy, 1995). Additionally, this test requires the patient to make inferences from ambiguous instructions, an ability suggested as most responsive to poor medication management but not readily assessed by other neuropsychological tests (Hayes, 1998; McGraw & Drennan, 2001; Park et al., 1994).

Finally, the utility of traditional EF measures in predicting one’s ability to carry out IADLs has been criticized due to the lack of strong correlations between performance on standardized testing and real-life functions within a specific domain (Diehl, Willis, & Schaie, 1995; Duke & Kaszniak, 2000; Grigsby, Kaye, & Robbins, 1995). As the Pillbox Test replicates an IADL task, this criticism does not apply. In fact, in addition to identifying possible executive dysfunction, Pillbox Test performance may provide important information about the level of comprehension patients have regarding their medical regimen and opportunities to implement strategies to improve medication management. In the long term, this may lead to earlier diagnosis and more cost-effective healthcare with increased medication adherence and lower hospitalization rates due to mismanagement of
medications. Additionally, while patients’ families may have difficulty comprehending how their loved ones’ performances on traditional neuropsychological tasks relate to real-world functioning, showing them a concrete example in the form of the patient’s difficulties organizing a pillbox can make it easier to understand the provider’s concerns. This is especially important as families tend to overestimate patients’ abilities to perform supposedly over learned IADL tasks such as medication management (Park, 1999).

Because the concept of assessing medication management is not new (Farris & Phillips, 2008), it is important to point out how the Pillbox Test differs from existing performance-based medication management measures, such as the Hopkins Medication Schedule (HMS; Carlson et al., 2005), the Medication Management Ability Assessment (MMAA; Patterson et al., 2002), and the Medication Management Test (Gurland et al., 1994). First and foremost, as has been stated above, the Pillbox Test was designed a priori to assess the four factors of EF proposed by Lezak (1995). No other medication management measure was developed to assess EF specifically, although varying degrees of EF are likely required. Toward this end, instructions for the Pillbox Test are very brief and lack structure (i.e., organize this pillbox with these medications for 1 week) so that the examinee’s executive abilities are assessed as fully as possible. This is in contrast to other medication management measures, most of which were developed to assess capacity to manage medications and ask specific questions of the examinee (e.g., if you took three pills every day, how many days would the pills last?) or provide specific instructions about what to do (e.g., read and explain five prescription labels, recognize tablet colors). Another difference in administration is that the Pillbox Test asks examinees to fill in compartments of a “weekly” pillbox with “five” medications, while other measures ask examinees to fill the compartments of a “daily” pillbox (e.g., HMS, MMAA) or to fill a weekly pillbox with fewer than five medications (e.g., Standardized Medication Task; Isaac et al., 1993). Because the Pillbox Test provides less structure and requires the organization of more medications over a longer time period than most other measures, it may be more difficult and thus more sensitive to detecting subtle cognitive dysfunction and particularly executive dysfunction.

Findings of this study should be interpreted in the context of study limitations. First, the results of this study were based on a limited sample size (n = 120). Second, due to source location characteristics, collateral information regarding each participant’s ability to perform IADLs was not collected. Third, neuroimaging was unavailable for the cardiac group, which may have provided more information concerning the integrity of the anatomy and function of the brain. Fourth, divergent validity of the Pillbox Test was not addressed and convergent validity of Lezak’s (1995) four-factor model could not be fully established due to the limited number of existing neuropsychological measures that formally assess each of the four EF factors. For example, few tests have been developed to quantify specifically self-monitoring or self-corrective behavior (i.e., factor four). Thus, perseverative responses on several tests were used as convergent validity measures but likely provide too narrow a representation of this EF factor. Finally, while these initial findings of the Pillbox Test suggest that it may be clinically useful as a measure of medication management, it was designed as an ecological measure of EF and not for assessing all aspects involved in medication adherence such as ability to open pill bottles, remembering to take medications on time, and obtaining refills. Healthcare providers wishing to thoroughly assess capacity to manage medications would be better served administering a measure designed specifically for this purpose (see Farris & Phillips, 2008, for a review of measures).

In summary, the Pillbox Test appears to be a promising measure of EF. Initial findings indicate that it demonstrates good construct validity, tapping all four domains of EF purported by Lezak, it does not rely on novelty or the examiner to act as the patient’s frontal lobes, it involves a concept familiar to most people and thus is likely less threatening, and it is brief and easy to administer and score, making it a potential measure for cognitive screening, especially in primary care settings. However, additional studies of test–retest reliability and validity in large and diverse samples with a variety of clinical diagnoses, including well-defined groups with specified cognitive impairment, are needed to firmly establish the utility of the Pillbox Test. The proposed cutscores, in particular, will require further validation before being applied more widely to other patient populations with questionable abilities to comply with medication regimens (e.g., moderate-to-severe traumatic brain injury, human immunodeficiency virus). Moreover, research on whether a time limit should be imposed may be helpful, since there is no time limit when organizing a pillbox at home.

Supplementary Material

Supplementary material is available at Archives of Clinical Neuropsychology online.

Funding

There were no sources of financial support for this study.
Conflict of Interest

None declared.

References


