Repeated Assessment and Practice Effects of the Written Symbol Digit Modalities Test Using a Short Inter-Test Interval

Diana R. Pereira1,2, Patrı´ cio Costa1,2, Joa˜o J. Cerqueira1,2,*

1Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal
2ICVS/3B’s—PT Government Associate Laboratory, Braga/Guimarães, Portugal

*Corresponding author at: Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. Tel.: +351 253 604 800; fax: +351 253 604 809.
E-mail address: jcerqueira@ecsaude.uminho.pt

Accepted 20 April 2015

Abstract

The Symbol Digit Modalities Test (SDMT) is a widely used instrument to assess information processing speed, attention, visual scanning, and tracking. Considering that repeated evaluations are a common need in neuropsychological assessment routines, we explored test–retest reliability and practice effects of two alternate SDMT forms with a short inter-assessment interval. A total of 123 university students completed the written SDMT version in two different time points separated by a 150-min interval. Half of the participants accomplished the same form in both occasions, while the other half filled different forms. Overall, reasonable test–retest reliabilities were found ($r = .70$), and the subjects that completed the same form revealed significant practice effects ($p < .001, d = 1.61$), which were almost non-existent in those filling different forms. These forms were found to be moderately reliable and to elicit a similar performance across participants, suggesting their utility in repeated cognitive assessments when brief inter-assessment intervals are required.

Keywords: SDMT; Practice effects; Repeated assessment; Test–retest reliability; Alternate forms

Introduction

Repeated neuropsychological assessments are necessary procedures in various clinical and research contexts. The evaluation of different cognitive functions over time can generate insightful data regarding, for example, the state and progression of a given clinical condition, possible improvements, and/or declines, and also the impact of adopted interventions, including surgery, pharmacological treatments, and cognitive rehabilitation. Even so a careful reading of the results between assessments is crucial since different factors can impact the subject’s performance. In this realm, practice effects constitute one of the most studied variables that can contribute to bias in repeated cognitive assessment. In simple terms, practice effects refer to an improvement in the task score from the first to the second or following applications of the test attributed solely to the task repetition (McCaflrey & Westervelt, 1995). Various factors such as comfort and familiarity with the test procedures, the development of learning strategies and the memorization of specific test stimuli can contribute to this performance enhancement.

In a general manner, when the same test version is applied in different occasions, improvements in the performance are probable (Benedict & Zgaljardic, 1998; Woods, Delis, Scott, Kramer, & Holdnack, 2006; Zgaljardic & Benedict, 2001) and can even endure 1 year after the baseline assessment (e.g., Basso, Bornstein, & Lang, 1999; Basso, Lowery, Ghormley, & Bornstein, 2001). Furthermore, practice effects tend to be larger between the first and the second evaluation (Baird, Tombaugh, & Francis, 2007; Beglinger et al., 2005; Collie, Maruff, Darby, & McStephen, 2003; Falleti, Maruff, Collie, & Darby, 2006; Monte, Geffen, & Kwapił, 2005; Register-Mihalik et al., 2012), and the performance can continue to improve in the subsequent time points although in a minor magnitude, or even reach a plateau (Bartels, Wegryn, Wiedl, Ackermann, & Ehrenreich, 2010; Beglinger et al., 2005). Additionally, some tests can reveal larger practice effects than others. For instance, research shows that instruments which involve learning a specific rule or strategy, and those related to psychomotor processing speed and with non-verbal items show more gains...
in retesting evaluations in comparison with verbal oriented tests (Baird et al., 2007; Calamia, Markon, & Tranel, 2012; Watson, Pasteur, Healy, & Hughes, 1994). Even when using alternate forms, verbal tests seem more resilient to practice effects comparing to non-verbal tests (Benedict & Žgaljardic, 1998). Learning, memory, and executive functions tasks are also prone to show practice effects (e.g., Basso et al., 1999, 2001; Lemay, Bédard, Rouleau, & Trembley, 2004; Mitrushina & Satz, 1991). The same happens in demanding and complex cognitive tasks, including tests where the development of a strategy is a key element (e.g., Basso et al., 1999, 2001). In contrast, instruments dedicated to explore functions like visual perception/recognition, naming, and attention seem to have less influence from previous testing occasions (e.g., Mitrushina & Satz, 1991; Wilson, Watson, Baddeley, Emslie, & Evans, 2000).

Another important feature that can be explored throughout repeated assessments is the test–retest reliability (or temporal stability). It provides information about the degree of measurement error and the test score consistency, taking into account the stability of the subject’s ranking positions in the scores distribution across different assessment points (Duff, 2012). Usually, the test–retest reliability is based on the correlation of the scores obtained in the same test, by the same subject, in two distinct occasions (Anastasi & Urbina, 1997). If the correlation is strong and significant, the test is considered to have little change over time and good test–retest reliability. The amount of time between assessments has an impact in this psychometric quality and, as a result, longer test–retest intervals seem to be linked with decreases in the magnitude of the correlation between test and retest evaluations (Calamia, Markon, & Tranel, 2013; McCaffrey & Westervelt, 1995). It is important to note that possible practice effects are not considered in the test–retest reliability measurements (McCaffrey & Westervelt, 1995); therefore, we can have a test with a good correlation coefficient that reveals concomitantly a significant overall score change between assessments.

The present study aims to explore the practice effects and test–retest reliability of two recently developed alternate forms of the Symbol Digit Modalities Test (SDMT; Smith, 1982) in the context of a short inter-assessment interval. In a general manner, the SDMT is a commonly used measure of information processing speed, entailing other components such as attention, working memory, visual scanning, and tracking (Strauss, Sherman, & Spreen, 2006). This task has been applied broadly in different clinical conditions, including Traumatic Brain Injury (TBI) and Multiple Sclerosis (MS), since it is a sensitive measure of information processing speed alterations (e.g., Draper & Ponsford, 2008; Forn, Belenguer, Parcet-Ibars, & Ávila, 2008). Specifically in MS, where deficits in information processing speed appear to be a hallmark (Batista et al., 2012; Forn et al., 2008; Huijbregts, Kalkers, Sonnevile, Groot, & Polman, 2006), the SDMT is incorporated in different neuropsychological batteries of reference (e.g., Rao’s Brief Repeatable Neuropsychological Battery, BRNB; Minimal Assessment of Cognitive Function in MS, MACFIMS; Brief International Assessment of Cognition for MS, BICAMS; Benedict et al., 2002; Langdon et al., 2012). In this context, the SDMT shows high sensitivity to detect cognitive alterations in MS (Dusankova, Kalincik, Havrdova, & Benedict, 2012; Glanz, Healy, Hviid, Chitnis, & Weiner, 2012; Portaccio et al., 2009; Van Schependom et al., 2014).

As aforementioned, the SDMT assumes a relevant role in diverse research and clinical fields. Nonetheless, the development of alternate forms for repeated testing and the study of its psychometric characteristics remain scarce (Benedict et al., 2012). Additionally, the exploration of these SDMT properties is also limited when considering short inter-assessment intervals. Indeed, shorter inter-test intervals are required to evaluate possible cognitive changes occurring within hours or minutes. Specific examples include fatigue studies (e.g., Johnson, Lange, Deluca, Korn, & Natelson, 1997), clinical investigations with pharmacological agents, in which few hours are needed for the medication to reach the peak effect (e.g., Pietrzak, Snyder, & Maruff, 2010), or studies exploring the impact of surgical interventions in cognition (e.g., cardiac surgery; Bruggemans, Van de Vijver, & Huysmans, 1997; Lewis, Maruff, Silbert, Evered, & Scott, 2006). Considering on one hand the lack of psychometric characterization of SDMT alternate forms and, on other hand, the dearth of SDMT data for short inter-assessment intervals, we planned a simple experimental design in which two recently developed SDMT alternate forms (Benedict et al., 2012) were tested in the context of a brief inter-assessment interval in a group of healthy subjects. It is noteworthy that preliminary data obtained under healthy good-performance conditions can produce relevant psychometric information and clarify the role of specific variables in the performance of neuropsychological tests. Thus, this approach can also contribute to a more attentive design, administration, and interpretation of results in future clinical studies (Kendall & Sheldrick, 2000). In Table 1, we selected and summarized some data from studies using the SDMT with repeated assessment designs and healthy groups. In these cases, it is possible to observe that different versions and inter-assessment intervals have been implemented, ranging from 1 week to 1 year. The reported test–retest reliabilities tend to vary between 0.72 and 0.98, supporting a reasonably high temporal stability. Moreover, previous research already showed that the SDMT is susceptible to practice effects (e.g., Levine, Miller, Becker, Selnes, & Cohen, 2004; Register-Mihalik et al., 2012) and that the use of available alternate forms seems to have a positive impact in controlling for this factor (e.g., Register-Mihalik et al., 2012). In this sense, similar outcomes were anticipated for this study, and we expected to extend these findings considering short inter-assessment intervals and also the use of Benedict’s alternate forms.
<table>
<thead>
<tr>
<th>Study</th>
<th>Participants: N of the sample, N of the most frequent sex, mean age (SD), mean years of formal education (SD)</th>
<th>SDMT version</th>
<th>Number of assessment sessions</th>
<th>Test–retest interval and test–retest reliability</th>
<th>Baseline raw score (mean [SD])</th>
<th>Final assessment raw score (mean [SD])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith (1982)</td>
<td>N = 80.48 women; Age = 34.8 (11.32); Education = 16.2 (2.50)</td>
<td>Original oral and written versions</td>
<td>2</td>
<td>M = 29.40 days written SDMT: r = .80 oral SDMT: r = .76</td>
<td>Written SDMT: 56.79 (9.84) oral SDMT: 64.99 (11.91)</td>
<td>Written SDMT: 60.46 (11.16) oral SDMT: 69.15 (11.97)</td>
</tr>
<tr>
<td>Hinton-Bayre, Geffen, and McFarland (1997) (study 1)</td>
<td>N = 54 professional rugby players, Age = 19.4 (2.1); Education = 12.3 (1.1)</td>
<td>Original version and 3 alternate forms</td>
<td>2</td>
<td>1–2 weeks collapsed for all forms: r = .72</td>
<td>Collapsed for all forms: 55.8 (13.1)</td>
<td>Collapsed for all forms: 57.4 (11.5)</td>
</tr>
<tr>
<td>Levine and colleagues (2004)</td>
<td>N = 1047 healthy male participants from which 465 completed the SDMT, Age = 38.1 (7.8); Education = 16.3 (2.3)</td>
<td>Original version</td>
<td>2</td>
<td>M = 192 days (SD = 53) r = .80</td>
<td>57.3 (8.64)</td>
<td>59.6 (9.47)</td>
</tr>
<tr>
<td>Hinton-Bayre and Geffen (2005)</td>
<td>N = 112 semiprofessional athletes from which 31 did form 1 (Age = 19.7 [SD = 3.2]), 30 form 2 (Age = 21.1 [SD = 4.0]), 26 form 3 (Age = 20.5 [SD = 3.3]), and 25 form 4 (Age = 20.7 [SD = 3.7])</td>
<td>Original written version and three alternate forms (from Hinton-Bayre et al., 1997)</td>
<td>2</td>
<td>1–2 weeks Forms 1 and 2: ICC = .97 Forms 1 and 3: ICC = .87</td>
<td>Form 1: 53.5 (9.6) Form 2: 55.8 (10.5)</td>
<td>Form 3: 58.4 (9.9) Form 4: 58.0 (13.1)</td>
</tr>
<tr>
<td>Duff and colleagues (2010)</td>
<td>N = 127 community-dwelling older adults, 103 women; Age = 78.7 (7.8); Education = 15.5 (2.5)</td>
<td>n/a (the same version was used throughout sessions)</td>
<td>3</td>
<td>M = 103 days (SD = 16) ICC = .94 (this value was obtained from a randomly selected subsample of 17 MS participants)</td>
<td>Only for the oral administration MS: 45.1 (11.6) Controls: 57.6 (12.5)</td>
<td>n/a</td>
</tr>
<tr>
<td>Akbar, Honarmand, Kou, and Feinstein (2011)</td>
<td>N = 119 participants with MS; 38 healthy subjects MS: 90 women; Age = 44.7 (8.5); Education = 15.0 (2.2) Controls: 29 women; Age = 41.8 (11.0); Education = 15.9 (1.7)</td>
<td>Computerized version; oral paper version (from the Rao’s Brief Repeatable Neuropsychological Battery, BRNB)</td>
<td>1 (2 for the temporal consistency calculations)</td>
<td>M = 103 days (SD = 16) ICC = .94 (this value was obtained from a randomly selected subsample of 17 MS participants)</td>
<td>Only for the oral administration MS: 45.1 (11.6) Controls: 57.6 (12.5)</td>
<td>n/a</td>
</tr>
<tr>
<td>Duff and colleagues (2011)</td>
<td>N = 26 participants with amnestic MCI with minimal PE; 25 MCI with large PE; 57 cognitively intact MCI minimal PE: 19 women; Age = 83.2 (6.7); Education = 15.1 (2.1) MCI large PE: 22 women; Age = 81.6 (6.4); Education = 15.8 (3.0) Cognitively intact: 46 women; Age = 77.1 (7.9); Education = 15.4 (2.7)</td>
<td>n/a</td>
<td>2</td>
<td>1 week MCI minimal PE: 32.5 (9.3) MCI large PE: 41.1 (8.8) Cognitively intact: 40.8 (7.8)</td>
<td>MCI minimal PE: 33.6 (10.4) MCI large PE: 42.2 (8.9) Cognitively intact: 44.2 (8.9)</td>
<td>n/a</td>
</tr>
<tr>
<td>Study</td>
<td>Sample Description</td>
<td>Procedure Details</td>
<td>Correlations</td>
<td>Mean Scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Benedict and colleagues (2012) | N = 25, 19 women; Age = 42.0 (15.6); Education = 14.8 (1.9) | Original version; two alternate forms from the BRNB; two alternate forms created in the context of the study | Collapsed for all forms between times 1 and 2: \( r = .84 \)  
Times 2 and 3: \( r = .86 \)  
Times 3 and 4: \( r = .89 \)  
Times 4 and 5: \( r = .90 \)  
Between new form 1 and new form 2 (used in our study): \( r = .86 \) | Collapsed for all forms: 59.3 (11.7)  
Collapsed for all forms: 64.9 (13.5) |
| Duff, Callister, Dennett, and Tometich (2012) | N = 268 community-dwelling older adults, 211 women; Age = 73.3 (7.6); Education = 15.3 (2.6) | n/a (the same version was used throughout sessions) | \( r = .86 \)  
Between sessions 1 and 2: \( M = 1.8 \) days (SD = 0.61)  
Between sessions 2 and 3: \( M = 1.6 \) days (SD = 0.59)  
Sessions 1 and 2: \( r = .795 \)  
Sessions 2 and 3: \( r = .743 \)  
Sessions 1 and 3: \( r = .621 \) | Collapsed for all forms: 56.2 (11.6)  
College group: 50.04 (14.53)  
High school group: 41.00 (5.85)  
College group: 33.74 (0.22)  
High school group: 41.95 (5.94) |
| Register-Mihalik and colleagues (2012) | N = 40, 20 women and 20 men | Three distinct alternate forms | Between sessions 1 and 2: \( M = 1.8 \) days (SD = 0.61)  
Between sessions 2 and 3: \( M = 1.6 \) days (SD = 0.59)  
Sessions 1 and 2: \( r = .795 \)  
Sessions 2 and 3: \( r = .743 \)  
Sessions 1 and 3: \( r = .621 \) | College group: 50.04 (14.53)  
High school group: 41.00 (5.85)  
College group: 33.74 (0.22)  
High school group: 41.95 (5.94) |
| Duff (2014)                   | N = 167 community-dwelling older adults, 136 women; Age = 78.6 (7.8); Education = 15.4 (2.5) | n/a | \( r = .86 \)  
Between sessions 1 and 3: \( M = 1.8 \) days (SD = 0.61)  
Between sessions 2 and 3: \( M = 1.6 \) days (SD = 0.59)  
Sessions 1 and 2: \( r = .795 \)  
Sessions 2 and 3: \( r = .743 \)  
Sessions 1 and 3: \( r = .621 \) | Collapsed for all forms: 56.2 (11.6)  
College group: 50.04 (14.53)  
High school group: 41.00 (5.85)  
College group: 33.74 (0.22)  
High school group: 41.95 (5.94) |
| Goretti and colleagues (2014)  | N = 273 from which 243 completed the baseline and the retest assessment, 180 women; Age = 38.9 (13.0); Education = 14.9 (3.0) | Oral version | \( r = .815 \)  
Between sessions 1 and 3: \( M = 1.8 \) days (SD = 0.61)  
Between sessions 2 and 3: \( M = 1.6 \) days (SD = 0.59)  
Sessions 1 and 2: \( r = .795 \)  
Sessions 2 and 3: \( r = .743 \)  
Sessions 1 and 3: \( r = .621 \) | Collapsed for all forms: 56.2 (11.6)  
College group: 50.04 (14.53)  
High school group: 41.00 (5.85)  
College group: 33.74 (0.22)  
High school group: 41.95 (5.94) |

Notes: ICC = intraclass correlation coefficient; M = Mean; MCI = Mild Cognitive Impairment; MS = Multiple Sclerosis; n/a = not applicable/not available; PE = Practice Effects; r = Pearson’s correlation coefficient; SD = standard deviation; SDMT = Symbol Digit Modalities Test.
Methods

Participants

The study conduction was approved by the local ethical committee for research in the health sciences (Ethics Subcommittee for Life and Health Sciences of University of Minho). A total of 123 healthy university students collaborated in the present study, 77 (62.6%) women and 46 (37.4%) men, aged between 19 and 37 years old ($M = 22.4$, $SD = 3.54$, 16% above 25 years old), 117 right-handed and 6 left-handed, and with 14.9 average years of formal education ($SD = 1.96$). The subjects were recruited during a class at the School of Health Sciences, University of Minho, Portugal. There were no reports of neurologic and/or psychiatric conditions, abusive consumption of substances, such as alcohol and drugs with known impact in the cognitive functioning, and no presence of sensorial/motor variations with significant interference in the test performance.

Materials

Symbol Digit Modalities Test. The SDMT (Smith, 1982) was created as a measure of cognitive screening for children and adults. It is a substitution task that covers diverse neurocognitive functions, including information processing speed, psychomotor functioning, attention, working memory, and visual scanning (Strauss et al., 2006). This test has two possible ways of administration, one written that can be used for individual and group settings and one oral for individual administrations and for subjects with motor complications. The test requires the substitutions of random geometric figures for a specific number, according to a key that contains nine different geometric designs paired with a single 1–9 Arabic number. In the written version, the one used in this study, a sheet of paper with the key on the top, 120 blank boxes paired with one specific design, and 10 blank boxes for initial practice is presented to the subjects, followed by the instructions. Individuals have 90 s to complete the blank boxes with the expected number of the key, and they are instructed to work as fast and accurate as possible. The SDMT takes, approximately, 5 min to administer, and the score corresponds to the total correct substitutions accomplished within the 90 s. The score ranges from 0 to 120, with higher scores pointing to a better performance. This is a simple test, not time-consuming, easily scored, and it is also a well-accepted measure for different subject groups, including clinical groups (Berrigan et al., 2014; Possa, 2010; Rogers & Panegyres, 2007; Walker et al., 2012). As abovementioned, two particular alternate forms developed by Benedict and colleagues (2012) and shown to be equivalent to the Smith’s original version (1982) were used (WPS Publishing, Torrance, CA).

Procedure

After giving written informed consent, participants completed a brief questionnaire with relevant personal and medical information, such as years of formal education, occupation, handedness, relevant diseases, and chronic medication use. This initial self-report facilitated the screening of identifiable neurological and psychiatric conditions, sensory-motor alterations, and pharmacological treatments with possible interference in the subjects’ performance. The experimental design consisted of two evaluations separated by, approximately, 150 min. In each occasion, the participants were asked to perform one of the written alternate forms: form 1 or form 2. The forms presentation was counterbalanced across participants and a $2 \times 2$ design was used implying two testing conditions: same condition, in which half of the participants filled the same version—1 or 2—in both occasions, and different condition where the remaining half undertook distinct forms, form 1 in the first assessment and form 2 in the second or vice versa. In both time points, the SDMT instructions were presented according to the SDMT manual (Smith, 1982) simultaneously to all the participants, which were also asked to complete the first 10 training items. Following this initial stage, participants completed one of the two forms during 90 s. Between the two testing phases, participants were engaged in their regular classes.

Statistical Analysis

For the analysis, subjects were aggregated according to the two previously described conditions, same and different, since the alternate forms had been shown to be equivalent (Benedict et al., 2012). All the SDMT values reported are based in the raw scores obtained from the total number of correct substitutions.

A mixed-design ANOVA was performed with condition as the between-subjects factor and the testing session as the within-subjects factor, in order to explore potential main and interaction effects between the two described factors on practice effects. Statistical significant interactions were further explored by using the appropriate $t$-tests. Forms equivalence was tested by comparing mean scores at baseline with an independent samples $t$-test. Test–retest reliabilities of both forms and between forms were obtained by computing Pearson’s product–moment correlations coefficients across the two time points.
Discussion

As convention, the results were considered statistically significant for $p < .05$, and when this significant condition was reached, partial eta squared ($\eta^2_p$) for ANOVA, and Cohen’s $d$ for $t$-tests were reported as measures of effect size. The statistical procedures were performed with the IBM Statistical Package for the Social Sciences Statistics software for windows, version 22 (IBM SPSS Statistics for Windows, Armonk, NY).

Results

Subjects assessed with the same and different forms in the two occasions were similar regarding age, gender, and years of formal education. Of note, the groups that completed a same or a different form had a similar distribution in terms of age ($t(117) = .166, p = .86, d = .17, years of education t(121) = 1.66, p = .099, d = 0.30, and sex $\chi^2_{(1, N=123)} = 1, p = .415, \phi = 0.07$. A summary of the main sociodemographic characteristics of the participants in the two conditions can be consulted in Table 2.

In the context of practice effects examination, the mixed-design ANOVA revealed a significant main effect of test session, $F(1,121) = 10.76, p = .001, \eta^2_p = 0.08$. More importantly, there was a significant condition $\times$ test session interaction, $F(1,121) = 82.81, p < .001, \eta^2_p = 0.41$, revealing that performance across evaluations varied according to the condition.

Indeed, although there were statistically significant differences between the two occasions in the same condition, $t(60) = -2.54, p < .001, d_c = 1.61$, performance in the different condition group was not significantly different, $t(61) = -0.69, p = .49$. This implies that practice effects are important when similar forms are used, but mitigated by the use of alternate forms between assessments (Table 3 and Fig. 1). It is noteworthy that, in the first assessment session, no significant differences were found between the participants in the same group ($M = 60.67, SD = 9.57$) and those in the different group ($M = 61.92, SD = 10.83$), $t(121) = -0.676, p = .50$, whereas in the second evaluation session, the same group revealed a better performance ($M = 75.71, SD = 13.16$) than the different group ($M = 62.63, SD = 9.71$), $t(121) = 6.28, p < .001, d = 1.13$. Additionally and supporting previous data that these forms are equivalent, performance in the first assessment for both forms was similar (form 1: $M = 62.25, SD = 9.80$; form 2: $M = 60.37, SD = 10.58$), $t(121) = 1.02, p = .31$.

Test–retest reliability was analyzed separately according to experimental condition, different or same (Table 3). Importantly, reliability was at a similar level for the group of participants which completed different forms ($r = .70, p < .001$) and the group exposed to the same form in both occasions ($r = .70, p < .001$; Fig. 1). Overall, these data indicate that both forms are moderately reliable when using the same or alternate forms throughout distinct assessment time points.

Discussion

Brief test–retest intervals of hours up to a few days have been implemented in diverse contexts to explore possible cognitive changes occurring within a short span of time or to attenuate the impact of some day-by-day variable factors with known influence.
in the test performance (Falleti et al., 2006). In spite of this, test properties under such short repeated administrations remain poorly characterized. In the present work, we assessed the practice effects and test–retest reliabilities of the two SDMT alternate forms in a group of university students using 2.5 h inter-test interval.

Our results showed significant practice effects when the same form was administered, but not when participants undertook different, although equivalent forms. These results are in line with previous investigations that report practice effects in the SDMT when similar forms are applied, even with longer test–retest intervals (e.g., Erlanger et al., 2014; Hinton-Bayre et al., 1997; Levine et al., 2004; Register-Mihalik et al., 2012). More importantly, they also extend to a much shorter test–retest interval previous observations of attenuated practice effects with the application of SDMT alternate forms (e.g., Register-Mihalik et al., 2012).

In the specific case of brief test–retest intervals, it would be expected that the use of alternate forms would be less effective in controlling practice effects, especially because the participants can recall similar test features, including instructions and test materials. Concerning this point, it is important to note that test itself comprise an initial training period, giving a first opportunity to familiarize with the test procedures. Therefore, the possible contribution of this factor to the practice effects, even when alternate forms are used, is probably stabilized from the beginning. Notably, the mitigation of the expected performance improvement when alternate forms are used supports the notion that item-specific practice has an important role in the SDMT-associated practice effects. As a result, when the items are slightly modified in alternate forms, it is possible to attenuate a significant performance enhancement due solely to item-specific learning. Thus, our data support the notion that alternate forms are relevant to diminish practice effects (Benedict & Zgaljardic, 1998), especially the ones associated with item-specific training (Calamia et al., 2012; Woods et al., 2006; Zgaljardic & Benedict, 2001). Even so, there are other strategies to control for practice effects worthy to mention, including: (a) the edification of dual base lines, in which the subject do enough practice trials to establish a stabilized, pre-baseline performance (e.g., Duff, Westervelt, McCaffrey, & Haase, 2001; Watson et al., 1994); (b) inclusion of a paired control group (Watson et al., 1994); (c) implementation of statistical procedures designed to have in consideration changes related to practice. In this last case, the Reliable Change Index (RCI) has been widely used (Lewis et al., 2006), since it provides information about how big a difference between two evaluations must be in order to consider a change as clinically relevant (RCI values were also calculated for this study and can be consulted in Supplementary material). The adaptation and combination of...
these proposals according to the nature of each situation seems the most careful approach for dealing with possible practice effects when interpreting the results obtained from repeated assessments.

Regarding the test–retest reliability, it would be expected to be >0.70 as recommended (Burlingame, Lambert, Reisinger, Neff, & Mosier, 1995), especially because we used a short inter-assessment interval that is theoretically associated with higher reliability coefficients (Slick, 2006). Although considering a .70 correlation as reasonable, the values reported by other studies with healthy participants tend to be above .70 (Table 1), and in the specific case of the Benedict’s and colleagues work (2012) that used the same alternate forms also in healthy subjects, they report a coefficient of .86. One of the possible explanations for this finding resides in the demographic specificities of our sample, since some groups can show a more variable pattern across time than others (Slick, 2006) and this has a reflection in terms of reliability. In Table 1, if we look for the studies with samples of highly educated adults with mean age below 40 (e.g., Goretti et al., 2014; Hinton-Bayre et al., 1997; Levine et al., 2004; Register-Mihalik et al., 2012; Smith, 1982), even using distinct inter-assessment intervals and different SDMT forms and versions, it is possible to verify Pearson’s correlation coefficients between .62 and .82, so there is some variability for this task and the value we found here can be viewed as satisfactory.

In this line of thought, it is important to recognize that variables such as cultural, ethnical, educational, and age-related factors can play an essential role in the test performance. Nevertheless, investigations aimed at a better clarification of the conjurred or independent contributions of these variables in the field of repeated testing performance tend to show different results according to the neuropsychological tasks and the population cohorts. Concerning the SDMT, the results tend to be controversial. Although some studies emphasize no major impact of variables like age and education in its results (e.g., Sheridan et al., 2006), others point to strong correlations between SDMT performance and the mentioned variables (e.g., Harris, Wagner, & Cullum, 2007; Vogel, Stokholm, & Jørgensen, 2013). In the case of practice effects, a study of Duff and colleagues (2012) revealed no significant correlation between the performance in SDMT and different demographic, clinical variables (e.g., age; formal education; depression; global cognition). These issues were not explored in the present study, which is an imperative limitation. In this sense, the results obtained here, derived from a group of healthy young adults highly educated, may not be generalized for other populations with different characteristics and also for subjects with a given clinical condition. Additionally, we used the written version and a group administration context, so it is not possible to perceive in what extent our results are applicable to the SDMT oral version, nor to individual testing settings. Even so, it is important to note that the raw scores obtained at baseline are close to the ones reported by other studies in different cultural settings especially with younger highly educated subjects (e.g., Bate, Mathias, & Crawford, 2001; Goretti et al., 2014; Jorm, Anstey, Christensen, & Rodgers, 2004; Nissley & Schmitter-Edgecombe, 2002; Tables 1 and 3).

Accordingly, the results found here can be a possible addition to the normative data specifically for the population cohort of European Portuguese university students.

Another important limitation regards the number of conducted assessments and the number of alternate forms used. On one hand, the results found here support that the two alternate forms created by Benedict and colleagues (2012) are moderately reliable considering a short inter-assessment interval but, on the other hand, other forms and related practice effects could be tested, including the original version by Smith (1982), the Hinton-Bayre and colleagues (1997) alternate forms, and the BRNB versions. Moreover, in the research and in clinical practice, it is common to have various repeated evaluations across time. Thus, it would be important to test these SDMT alternate forms in distinct brief and long inter-assessment intervals, with more evaluation time points, in order to get some approximations with the diversity and emergent needs present in different clinical/research context. Another point that needs additional exploration is the possible association between test–retest interval length and the magnitude of the practice effects. In this context, some studies already show no differences between the practice effects obtained in various inter-assessment times (e.g., Baird et al., 2007; Hinton-Bayre & Geffen, 2005); for instance, in the study of Baird and colleagues (2007), the practice effects were similarly noticeable for test–retest intervals of 3 months, 1 week, and 20 min.

As final remarks, the findings presented here support that the SDMT alternate forms used (Benedict et al., 2012) are moderately reliable and equivalent, suggesting its usefulness for serial neuropsychological evaluations. Even considering that the results were extracted from a specific healthy cohort of participants, this study gathers some data regarding the scores stability and practice effects of two SDMT alternate forms in the context of a short inter-assessment interval. This information can be useful for specific normative comparisons (Kendall & Sheldrick, 2000) and for the design of future investigations with other population cohorts, including with clinical conditions. The development and psychometric study of SDMT alternate forms is crucial, since this test can be applied successfully in diverse ethnic and cultural populations (Harris et al., 2007; O’Bryant, Humphreys, Bauer, McCaffrey, & Hilsabeck, 2007). Similarly, it is a promising cognitive screening tool in different clinical conditions, including MS (Morrow, Jurgensen, Forrestal, Munchauer, & Benedict, 2011; Strober, Rao, Lee, Fischer, & Rudick, 2014) and TBI (Draper & Ponsford, 2008), in which repeated cognitive evaluations over time are essential. Our results also support the SDMT as a reliable instrument to administrate across brief test–retest periods, and this fact can be advantageous for surgical and pharmaceutical interventions that require such short intervals (e.g., Bruggemans et al., 1997; Lewis et al., 2006; Pietrzak et al., 2010). Nevertheless, more investigations are warranted to clarify how different properties of cognitive tests may change.
in the context of repeated assessment, including possible variations associated with practice effects. So it is pertinent to test brief or long inter-assessment time points, and different cultural, sociodemographic, and clinical characteristics (McCaffrey & Westervelt, 1995; Putnam, Adams, & Schneider, 1992; Slick, 2006). Moreover, the practice effects can be studied, on one hand, to elucidate their impact in serial testing so that their influence is accounted for when significant cognitive changes are expected and, on other hand, as a measure of cognitive performance. More specifically, the absence or the diminished development of expected practice effects has been suggested as an important marker of neuropsychological dysfunction (Duff et al., 2010). Overall, new and old neuropsychological instruments require consistent investigations regarding several psychometric characteristics, practice effects associated and even variations linked to different population cohorts and clinical groups.

Supplementary Material

Supplementary material online material is available at Archives of Clinical Neuropsychology online.

Funding

This work was supported by European Union FEDER funds through Programa Operacional Factores de Competitividade (COMPETE) for ADI Project “DoIT—Desenvolvimento e Operacionalização da Investigação de Translação” (“MyHealth-PPS4”; project n° 13853).

Conflict of Interest

None declared.

Acknowledgements

Diana R. Pereira is a recipient of a scholarship under the MyHealth project. The authors would like to thank particularly to all the participants involved in this study and to the anonymous reviewers for the significant comments and suggestions that helped to improve the quality of the manuscript.

References


