Internal consistency, temporal stability, and reproducibility of individual index scores of the Test of Variables of Attention in children with attention-deficit/hyperactivity disorder

Antolin M. Llorente, Alfred J. Amado, Robert G. Voigt, Marcia C. Berretta, J. Kennard Fraley, Craig L. Jensen, William C. Heird

Department of Pediatrics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA
Department Psychiatry, Baylor College of Medicine, Houston, TX, USA
Texas Children’s Hospital, Houston, TX, USA
Department of Educational Psychology, Texas A&M University, College Station, TX, USA

Accepted 31 March 2000

Abstract

Psychometric properties of the Test of Variables of Attention (TOVA) were examined in a cohort of children (n = 63) strictly diagnosed with attention-deficit/hyperactivity disorder (AD/HD). Internal consistency was assessed via correlational analyses to determine the degree of agreement among various test portions. The temporal stability of errors of omission, errors of commission, response time, and response time variability was evaluated using test–retest reliability. Reproducibility of individual scores for the same indices was assessed using the Bland–Altman procedure. Select TOVA index scores exhibited high internal consistency in this cohort. Although the temporal stability of group scores (test–retest reliability) was satisfactory, individual test scores were less reproducible. Temporal stability and individual test–retest score agreement were greater for response time and response time variability than for errors of omission and errors of commission. © 2001 National Academy of Neuropsychology. Published by Elsevier Science Ltd.

Keywords: AD/HD; Hyperactivity; Reliability; Response time; Temporal stability; TOVA

* Corresponding author. Tel.: +1-713-770-3400.
E-mail address: llorente@bcm.tmc.edu (A.M. Llorente).
Attention-deficit/hyperactivity disorder (AD/HD) is characterized by developmentally inappropriate and disabling levels of inattention, impulsivity, and/or excessive motor activity (Laufer & Denhoff, 1957; Chess, 1960; Douglas, 1972; Barkley, 1989, 1990; APA, 1994). In addition, children with AD/HD are at increased risk of developing school difficulties, peer relational problems, and marital and occupational dysfunction later in life (Hechtman et al., 1984; Weiss et al., 1985; Lambert, 1988). With regard to epidemiology, AD/HD is one of the most common neurobehavioral disorders observed in childhood with an estimated prevalence of 3% to 5% of school-age children (APA, 1994; Gadow & Sprafkin, 1997).

Objective laboratory instruments have been developed to assist in differentiating the underlying deficits of AD/HD (Gordon, 1987, 1993; Barkley, 1990; Greenberg & Waldman, 1993) and continuous performance tests (CPTs) (cf., Rosvold et al., 1956) are among the most frequently used measures for such purposes. CPTs are reaction time tasks that rapidly (milliseconds) present stimuli requiring the discrimination of predetermined targets from distracting non-targets. Indices recorded by most CPTs include errors of omission (i.e., failure to respond to the target stimulus) that is interpreted as a measure of inattention (cf., Sostek et al., 1980), errors of commission (i.e., responding to the non-target stimulus) that is considered a measure of impulse control (Sostek et al., 1980), and response time that is interpreted as an index of information processing and reaction time (cf., Conners, 1995).

Although CPTs have potentially reliable and valid components that enhance the multi-method assessment of AD/HD symptomatology (Corkum & Siegel, 1993; Gordon, 1993) and its diagnosis, their ecological validity has not been well established (Barkley, 1991; Kamphaus & Frick, 1996). Other criticisms leveled against CPTs include their lack of specificity or discriminant validity (Barkley, 1991). Although related, yet separate from limitations in discriminant validity (cf., Anastasi & Urbina, 1998), limited information addressing the reliability (internal consistency, temporal stability, etc.) of the inferences derived from CPTs have been presented in the neuropsychological literature and therefore the psychometric properties of CPTs remain poorly understood (see Boivin et al., 1996; Losier et al., 1996; Forbes, 1998 for recent research attempting to address specific properties of CPTs). Nevertheless, both practitioners and investigators continue to use CPTs at an increasing rate to assess AD/HD symptomatology largely because these tests do not rely on personal opinion and subjective judgement as do interviews and self-rating scales (Chronbach et al., 1972; Barkley, 1990; Teicher et al., 1996) and because they have been shown to be sensitive in detecting beneficial medication effects (Coons et al., 1987; cf., Fischer, 1996).

The Test of Variables of Attention (TOVA) is a standardized, fixed-interval (21.6 ± 1.1 min), visual or auditory CPT. It is a language-neutral test in the sense that it uses geometric figures rather than letters or numbers to represent the target and non-target visual stimuli. Successful performance on the TOVA requires sustained attention and up–down discrimination with relatively few other neurocognitive domain contaminations (Greenberg & Waldman, 1993; Greenberg & Kindschi, 1996). An electronic micro-switch is used to distinguish between two identical large outlined rectangles each containing either the target (smaller color square at the top) or the non-target (smaller color square at the bottom) stimulus. The target-to-non-target presentation ratio of the TOVA is constant for each half of the test but the frequency of target presentation is different for each half. The 22.5% and 77.5% of the presentations during the first half (Half 1, infrequent stimulus condition) and second half
(Half 2, frequent stimulus condition) are targets, respectively. Although not all are standardized, indices measured by the TOVA include: errors of omission or failure to respond to the designated target that is interpreted as a measure of inattentiveness; errors of commission or incorrect response to the non-target that is considered a measure of response disinhibition or impulsivity; mean response time or the time required to respond correctly to the target stimulus that is interpreted as a measure of information processing and motor reaction speed; standard deviation of the response time, or response time variability, for correct target stimulus responses that is interpreted as a measure of attentional variability or consistency; multiple responses to the target or reflection of neurological status that is interpreted as a measure of motor hyper-responsivity; post-commission error mean response times for correct responding to the target stimulus; anticipatory responses or guessing that is interpreted as a measure of impulsivity and of the validity of the results; \(d'\) (d prime) score or response sensitivity that is used to interpret the rate of deterioration of performance over time; and AD/HD index score that is indicative of the degree of similarity in performance between the child undergoing assessment relative to the children in the normative sample with AD/HD profiles. Familiarization with the actual test prior to its administration is achieved through a standardized 3-min practice session (Greenberg & Kindschi, 1996; Leark et al., 1996).

The published normative data for the TOVA were primarily obtained from more than 1300 Caucasian children and adolescents ranging in age from 4 to 19 years (Greenberg & Waldman, 1993; Leark et al., 1996). However, participants in the normative sample were excluded from the normalization study if they met one of the following criteria: (1) deviant classroom behavior rating score (greater than two standard deviations) on the Conners’ Parent–Teacher Questionnaire (Abbreviated Form), (2) current use of psychoactive medication, or (3) receiving special education services (Leark et al., 1996). These criteria most likely excluded from the normative sample a large number of children with AD/HD, particularly children with severe inattention and impulsivity commonly seen in clinical and research populations.

Because the magnitude of reliability estimates is influenced by both the qualities of a test (length, test type [speed vs. power], etc.) and the characteristics of the sample (selection criteria, size, etc.) under investigation (Guilford & Fruchter, 1978; Franzen, 1989; Anastasi & Urbina, 1998), the purpose of this study was to assess the internal consistency, temporal stability, and reproducibility of individual test index scores (errors of omission, errors of commission, response time, and response time variability) of the TOVA in a group of children strictly selected and diagnosed with AD/HD. The limited availability of data addressing the TOVA’s psychometric properties including its internal consistency, temporal stability, and individual score reproducibility in children diagnosed with AD/HD, in conjunction with its increased use for assessment and diagnostic purposes, served as impetus for this study.

1. Method

1.1. Participants

An original pool of 250 volunteers screened for potential participation in this study resulted in a research sample of 63 participants who had been previously diagnosed
with AD/HD and placed on an effective dose of stimulant medication (e.g., methylphenidate [mean dosage = 25.7 ± 12.9 mg/day] administered one to three times per day; duration of treatment averaged 35.4 ± 22.2 months) by their primary care physician or pediatrician. All participants were enrolled in a study (Voigt et al., 1998) evaluating the effectiveness of dietary docosahexaenoic acid (DHA) supplementation (cf., Mitchell et al., 1987; Arnold et al., 1989) to ameliorate or reduce inattention, poor impulse control, and excessive motor activity. Children who met any preset exclusion criteria including history of head injury or seizures, diagnosis of a lipid metabolic disorder, any chronic medical condition (e.g., heart disease), a significant life event within the previous 6 months (e.g., divorce in the family), use of long-chain polyunsaturated fatty acid (LCPUFA) supplementation within 6 months, requirements for special education services for mental retardation or borderline intellectual functioning, autism, Asperger or Tourette Syndrome, pervasive developmental disorder-NOS, anxiety or mood (e.g., dysthymia, depression, or mania) disorder, history of prematurity, or maternal use of alcohol, illicit drugs or tobacco during pregnancy were excluded. Children with learning disabilities or oppositional-defiant disorder were not excluded consequent to the high comorbidity between these disorders and AD/HD (see Barkley, 1990; Semrud-Clikeman et al., 1992).

In addition to meeting DSM-IV diagnostic criteria for AD/HD, and because any combination of six out of nine items from the DSM-IV AD/HD criteria, Inattention, meets DSM-IV diagnosis, all participants in this study had to meet five of the six following items dealing with inattention within the Inattention subclassification (1a [“fails to give close attention,” 98.5% of participants met 1a], 1b [“difficulty sustaining attention,” 92.4% of participants met 1b], 1c [“does not seem to listen,” 98.5% of participants met 1c], 1d [“often does not follow instructions,” 100% of participants met 1d], and 1 h [“often easily distracted,” 100% of participants met 1 h]), because inattention has been shown to be the core symptom of this disorder (cf., Douglas, 1972). Implementation of exclusion criteria resulted in a sample of sixty-three 6- to 12-year-old (M = 9.5 years; SD = 1.5; 52 males, 11 females) children with AD/HD available for this study. Five children met DSM-IV criteria for AD/HD, Predominantly Inattentive Type while 58 children met DSM-IV criteria for AD/HD, Combined Type. A total of 50 children were Caucasian, 10 were African American, and 3 were Hispanic. Data from forty-nine children were available for analyses.

1.2. Materials

Initial diagnostic interview was conducted using the American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, DSM-IV (APA, 1994) to establish a diagnosis of AD/HD. Each participant was evaluated employing two objective neuropsychological measures of attention. The Children’s Color Trails Test 1 and 2, an orthographic (paper and pencil) measure capable of assessing attention in children (Williams et al., 1995) and similar to the Trailmaking Test A and B (Reitan, 1979), was employed as a measure of alternating attention. The visual TOVA (Leark et al., 1996) was used to assess sustained attention and concentration.
1.3. Procedure

The original pool of volunteers was initially screened by telephone to exclude those with any preset exclusion criteria. A total of 160 volunteers met one or more of these exclusion criteria. The remaining 90 children were further evaluated by a developmental pediatrician to confirm the diagnosis of AD/HD according to established DSM-IV criteria (APA, 1994), to insure subjects met the more stringent criteria from the general DSM-IV nosology established for this study (five out of six items from the Inattention subclassification previously noted), and to exclude those not meeting diagnostic criteria or those unable to participate in the study (they did not consent to participation or were unavailable for three visits). DSM-IV protocols for the remaining participants available for this study were independently reviewed by a child psychologist and determined to meet DSM-IV, AD/HD criteria establishing an interdiagnostician (inter-rater) reliability of 0.82 between clinicians. Finally, a reliability coefficient of 0.84 was obtained between the TOVA Interpretation Index (Leark et al., 1996) (first TOVA assessment) and the DSM-IV, AD/HD criteria. The remaining 70 children were randomized to receive either DHA or a placebo in a double-blind fashion along with their usual effective dose of stimulant medication (see Voigt et al., 1998). Data from only 63 children were available for investigation.

Outcome variables, obtained at baseline and after 2 and 4 months of study, included plasma and erythrocyte phospholipid fatty acids concentrations and objective assessment of inattention and impulsivity using the Children’s Color Trails Test 1 and 2 and the TOVA. The objective measures of attention were performed while each participant was off stimulant medication for approximately the previous 24 hours so that scores on the TOVA did not reflect the effect of the stimulant medication (Winsberg et al., 1982). The objective psychological screening was performed in the morning in 90% of the participants. Written informed consent was obtained from each participant’s parent or legal guardian and verbal assent was obtained from each child. The study was approved by the Institutional Review Board of the Baylor College of Medicine and affiliated hospitals.

1.4. Data analyses

The internal consistency of the TOVA was evaluated by calculating the extent of relationship between various portions of the test (e.g., Half 1 and Half 2 at Visit 3, subsequent to completion of DHA supplementation). A Pearson product–moment correlation (rxy) was computed between the total number of correct target responses (errors of omission and commission) on Half 1 and the total number of correct target responses (errors of omission and commission) on Half 2 (Half 1 vs. Half 2). Identical statistical analyses were used to individually assess the degree of response consistency between each half of the test and the entire test, independently correlating the number of total correct target responses (errors of omission and commission) on Half 1 and Half 2 with the number of total correct target responses (errors of omission and commission) for the entire test (Half 1 vs. Total and Half 2 vs. Total).

Due to the nature of the TOVA (partly a speed and power test; cf., Anastasi & Urbina, 1998), the temporal stability of the test was computed using test–retest reliability
(r_{xy}). Pearson product–moment correlations (stability) were calculated across the three administrations for errors of omission, errors of commission, response time, and response time variability.

Although the reliability of neuropsychological measures has been estimated historically by correlation coefficients, and these procedures are appropriate for determining the repeatability of group scores, such procedures provide little information about the reproducibility of individual test scores (Altman & Bland, 1983). In fact, it is possible for scores from groups of participants to exhibit high reliability but limited individual score reproducibility. Therefore, the Bland–Altman (Altman & Bland, 1983; Bland & Altman, 1986) procedure was used to calculate the limits of agreement between individual errors of omission, errors of commission, response time, and response time variability between visits (V_1-V_2, V_1-V_3, and V_2-V_3). Limits of agreement were calculated using the difference between the scores from two test administrations (Bland & Altman, 1986). In addition, individual differences were plotted against the average of the test–retest values to determine if agreement between test scores were related to response scores.

2. Results

Because one of the chief aims of the parent investigation from which these data were obtained was to determine whether DHA reduced the symptoms of AD/HD, and because such an effect could threaten the internal validity of this study (test of the original null hypothesis), a repeated measures ANOVA was conducted to determine whether long-term DHA administration reduced AD/HD symptomatology. However, dietary DHA supplementation failed to selectively reduce AD/HD symptoms in children receiving this fatty acid in the parent investigation (Voigt et al., 1998). Since the children were tested when off stimulant medication, the lack of DHA effect permitted the unaltered study of the psychometric properties of the TOVA in this cohort.

2.1. Internal consistency

Internal consistency correlations are listed in Table 1. The correlation assessing the relationship among the two independent test halves (total correct target responses) (Half 1

<table>
<thead>
<tr>
<th>Test segment</th>
<th>r_{xy}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half 1 vs. Half 2</td>
<td>0.93***</td>
</tr>
<tr>
<td>Half 1 vs. Total</td>
<td>0.96***</td>
</tr>
<tr>
<td>Half 2 vs. Total</td>
<td>0.99***</td>
</tr>
</tbody>
</table>

**Correlation significant at the 0.001 level (two-tailed test); V_1 and V_2 correlations are similar.**
vs. Half 2) revealed an internal consistency in the ‘High’ range \( (r_{xy} = 0.93) \). Internal consistencies between Half 1 (infrequent stimulus condition) total correct target responses vs. total correct target responses for the entire test and Half 2 (frequent stimulus condition) total correct target responses vs. total correct target responses for the entire test (Half 1 vs. Total and Half 2 vs. Total) also were in the ‘High’ range \( (r_{xy} = 0.96 \) and \( r_{xy} = 0.99 \), respectively).

### 2.2. Temporal stability (test–retest reliability)

Table 2 lists test–retest correlations (reliability) for errors of omission, errors of commission, response time, and response time variability. The results revealed statistically significant temporal stability in the ‘Moderate’ range from baseline through four months. The correlations representing temporal stability ranged in magnitude from \( r_{tt} = 0.51 \) to \( r_{tt} = 0.82 \). They reached such a magnitude despite the extensive length of time elapsed between assessments (see Discussion). Response time and response time variability exhibited greater temporal stability than errors of commission and errors of omission.

### 2.3. Reproducibility of individual TOVA index scores

Table 3 summarizes individual test–retest mean difference scores with limits of agreement using the method of Bland–Altman (Altman & Bland, 1983). The Bland–
Errors of omission and commission scores exhibited bias across all visit comparisons. The observed bias in these two indices was the result of increases in the difference of scores from one examination to the next as a function of increasing average scores. Response time and response time variability exhibited less bias associated with increasing average scores of these variables than that exhibited by errors of omission and commission, exhibiting greater reproducibility of individual scores across all visit comparisons.

3. Discussion

The present results suggest that the TOVA possesses robust internal consistency when used with children strictly selected and diagnosed with AD/HD in a fashion similar to that employed to categorize this cohort. The TOVA exhibited internal consistency in the ‘High’ range when the two halves of the test were correlated with each other as well as when each

---

1 Bland–Altman plots for TOVA indices comparing all visits are available from AML.
half of the test was independently correlated with the entire test. The correlation coefficient between total correct target responses on Half 2 and total correct target responses on the entire test was slightly higher relative to the correlation between total correct target responses on Half 1 and total correct target responses on the entire test. The increase in magnitude is most likely the result of practice effects obtained from doing Half 1 first or consequent to regression effects. Although it could be concluded on the basis of a superficial analysis that the high internal consistency observed in the TOVA between the total number of correct target responses of each independent test half and the total number of correct target responses on the entire test was simply the outcome of administering a lengthy test (324 total targets, etc.), a more exhaustive examination of the present findings, coupled with a thorough understanding of the intricacies of this instrument, would refute such a conclusion. Such an inference fails to appreciate that these two internal consistency coefficients, obtained from independently correlating all errors of omission and commission for targets only from each half of the test with the entire test, fell in the High range despite the fact that each test half has significantly different target presentation frequencies (Half 1, infrequent stimulus condition; Half 2, frequent stimulus condition).

The present results also revealed satisfactory temporal stability when measured using correlational procedures, the traditional fashion in which psychological instruments have been gauged in the past. Although the correlations representing the temporal stability ($r_{tt}$) of the TOVA in this study fell below the level recommended for clinical decision-making (approximately $r_{tt} = 0.80$, cf., Sattler, 1988; Anastasi & Urbina, 1998), with the exception of the response time correlation between administrations at V2–V3 ($r_{tt} = 0.82$), it should be noted that the aforementioned recommendations usually apply to test administrations with significantly shorter test–retest intervals (2 to 4 weeks, cf., Sattler, 1988) than those used in this study (8 and 16 weeks). It is highly likely that shorter intervals between test administrations would have had a positive (incremental) effect on the magnitude of the stability coefficients since increases in time interval between test administrations tend to result in decreases in the magnitude of these correlations (see Anastasi & Urbina, 1998). In the same vein, the homogeneous composition of this research sample most likely had a negative effect on the test–retest correlations. As a group under investigation becomes increasingly similar (e.g., a research population) secondary to exclusionary or other homogenizing selection criteria, the stability coefficient from such populations tends to decrease (cf., Anastasi & Urbina, 1998). Therefore, $r_{tt} = 0.51–0.81$ scores are most likely satisfactory for such long intervals between test administrations as used in this investigation. It is also important to note that increasing the time between test administration appeared to pose less threat to the temporal stability of the response time relative to errors of omission and commission. Nevertheless, from an applied standpoint, the application of these coefficients should be limited to the test intervals presented here as well as similar populations of children with AD/HD.

The reproducibility of individual scores using the Bland–Altman procedure revealed less agreement than that observed for the entire sample. There was more bias in individual errors of omission and commission scores than in response time and response time variability. The observed bias was the result of increases in the difference of scores from one examination to the next as a function of increasing average scores. The fact that omissions and commissions
exhibited greater bias (less individual test–retest score agreement) relative to response time and response time variability may be associated with the nature and measurement of response time (less fluctuation in scores), a relatively basic measure of reaction time measured in milliseconds representing the amount of time required to process and respond correctly to a target, compared to errors of omission and commission (greater fluctuation in scores). However, this finding is not surprising as it is consistent with the work of other investigators (Boivin et al., 1996) addressing other psychometric properties of the TOVA.

The variability observed in individual TOVA index scores, although expected, has significant implications for clinical applications, including the assessment and diagnosis of AD/HD. Consistent with the recommendations for the TOVA as described in its manual (cf., Greenberg & Kindschi, 1996) and the Standards for Educational and Psychological Testing (American Psychological Association, 1985), TOVA scores should be interpreted within the context of a more comprehensive evaluation including the examinee’s history, neuropsychological test profile, and neurobehavioral characteristics since no individual assessment component provides sufficient information to interpret brain–behavior relationships. In addition, as the present investigation suggests, TOVA test scores from single examinations should be interpreted with greater caution, particularly errors of commission and omission, as they may exhibit substantial inter-test (test-to-test session) variability (see Leark et al., 1996). The extensive variability is most likely associated with the fluctuations in sustained attention sometimes observed in the symptom complex of AD/HD, factors associated with the variables being measured (e.g., stability of reaction time relative to other dependent variables), as well as developmental variables in the case of children.

While the children participating in this study were selected using a stringent protocol, including DSM-IV AD/HD criteria, selecting participants that met five out of six symptoms from the DSM-IV AD/HD Inattention subclassification, high concordance between child psychologist and developmental pediatrician, and a high inter-method correlation between the TOVA Interpretation Index and DSM-IV AD/HD diagnostic, it is possible that a group of children selected on the basis of a similar but different narrow criteria employing behavioral rating scales may exhibit different psychometric properties on the TOVA. For this reason, the coefficients presented here are applicable to children selected on the basis of similar criteria as reliability estimates vary as a function of the selection parameters and characteristics of the sample under investigation (see Anastasi & Urbina, 1998).

One criticism of the TOVA, and similar CPTs, stems from the lack of specificity (see Kamphaus & Frick, 1996) and ecological validity (Barkley, 1991) of these procedures (see Boivin et al., 1996; Forbes, 1998 for recent research addressing the validity of the TOVA). As noted above, although CPTs may show appropriate specificity when differentiating ‘normal’ children from children with attentional problems (see Corkum & Siegel, 1993), it has been inconsistent in distinguishing children with AD/HD from other clinical diagnostic groups (Barkley, 1990), or children with different AD/HD diagnostic subtypes. Therefore, studies that attempt to delineate the discriminant validity (specificity), reliability, and other psychometric properties of the TOVA on different populations merit attention.
Acknowledgments

Portions of this research were supported in part by grants from Martek Biosciences Corp., Columbia, MD, USA and the TOVA Foundation, St. Paul, MN, USA to Antolin M. Llorente and Robert G. Voigt. Select results from this study were presented at the 19th Annual National Academy of Neuropsychology Conference, San Antonio, TX, November, 1999. The authors wish to express their sincere thanks to Robert A. Leark, PhD, for his comments and suggestions on an earlier draft of this manuscript.

References


