Published in final edited form as:
Clin Neuropsychol. 2009 September ; 23(7): 1213–1231. doi:10.1080/13854040902855358.

RELIABILITY CONCERNS IN THE REPEATED COMPUTERIZED ASSESSMENT OF ATTENTION IN CHILDREN

T. Andrew Zabel[1,2], **Christian von Thomsen**[1,2], **Carolyn Cole**[1,2], **Rebecca Martin**[1], and **E. Mark Mahone**[1,2]

[1]Kennedy Krieger Institute, Baltimore, MD, USA

[2]Johns Hopkins University School of Medicine, Baltimore, MD, USA

Abstract

Assessment of attentional processes via computerized assessment is frequently used to quantify intra-individual cognitive improvement or decline in response to treatment. However, assessment of intra-individual change is highly dependent on sufficient test reliability. We examined the test–retest reliability of selected variables from one popular computerized continuous performance test (CPT)—i.e., the Conners' CPT – Second Edition (CPT-II). Participants were 39 healthy children (20 girls) ages 6–18 without intellectual impairment (mean PPVT-III SS = 102.6), LD, or psychiatric disorders (DICA-IV). Test–retest reliability over the 3–8 month interval (mean = 6 months) was acceptable (Intraclass Correlations [ICC] = .82 to .92) on comparison measures (Beery Test of Visual Perception, WISC-IV Block Design, PPVT-III). In contrast, test–retest reliability was only modest for CPT-II raw scores (ICCs ranging from .62 to .82) and T-scores (ICCs ranging from .33 to .65) for variables of interest (Omissions, Commissions, Variability, Hit Reaction Time, and Attentiveness). Using test–retest reliability information published in the CPT-II manual, 90% confidence intervals based on reliable change index (RCI) methodology were constructed to examine the significance of test–retest difference/change scores. Of the participants in this sample of typically developing youth, 30% generated intra-individual changes in T-scores on the Omissions and Attentiveness variables that exceeded the 90% confidence intervals and qualified as "statistically rare" changes in score. These results suggest a considerable degree of normal variability in CPT-II test scores over extended test–retest intervals, and suggest a need for caution when interpreting test score changes in neurologically unstable clinical populations.

Keywords

Attention; Conners' Continuous Performance Test; Reliable Change Index; Normal development; Serial assessment; Neuropsychology

Address correspondence to: T. Andrew Zabel, Ph.D., Department of Neuropsychology, Kennedy Krieger Institute, 1750 East Fairmount Avenue, Baltimore, Maryland 21231 USA. zabela@kennedykrieger.org .

Publisher's Disclaimer: Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf
This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.
The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

INTRODUCTION

Serial neuropsychological assessment is an important component in the clinical monitoring of children with potentially unstable neurologic conditions. Detection of change in a child's neuropsychological test scores over time is one of several available means for identifying cognitive and neurologic decline in these patients (Matson, Mahone, & Zabel, 2005). Individual comparison standards such as those used in serial assessment are preferred when addressing clinical questions of improvement or deterioration (Lezak, 1995), as they permit direct measurement of rate of change. The utility of serial neuropsychological assessment has been enhanced by the application of statistical methods designed to differentiate *intra-individual chance* test score variation from *statistically rare* changes in performance, most notably via the Reliable Change Index (RCI; Jacobson & Truax, 1991) and variations of RCI (Chelune, Naugle, Lüders, Sedlak, & Awad, 1993). An emphasis on detecting meaningful intra-individual change has extended from forensic neuropsychology to pediatric neuropsychology, with normative data becoming increasingly available for these methods of statistical comparison (Baron, 2004; Strauss, Sherman, & Spreen, 2006).

Among the most frequently used pediatric neuropsychological measures for the repeated assessment of rate of change in attention and related cognitive variables are computerized continuous performance tests. One such instrument, the Conners' Continuous Performance Test – Second Edition (CPT-II; Conners, 2000), is a 14-minute computerized measure of inhibitory control, sustained attention, vigilance, reaction time, and response variability. The CPT-II is marketed for several uses, including as a clinical screening device, as an aid for monitoring treatment/medication effectiveness, and as a research instrument (Conners & Multi-Health System Staff, 2004). Considerable validation support exists for the use of the original CPT, as well as the CPT-II, in the patient screening process and in detecting performance-based differences between clinical/treatment groups and comparison/control groups (e.g., Boro, Vahip, & Akdeniz, 2006; Epstein et al., 2003, 2006; Gruber et al., 2007).

The repeated use of the CPT-II for detecting changes in attention and neurocognitive status has been explored as well. The vast majority of studies employing the CPT and CPT-II in serial assessment have used it to describe within-group performance based differences. For instance, multiple studies have effectively employed the CPT and CPT-II in repeated measures designs to detect positive or negative treatment related *group* effects in response to stimulant medication or radiation therapy (Borgatti et al., 2004; Kiehna, Mulhern, Li, Xiong, & Merchant, 2006; Posey et al., 2006; Schachar et al., 2008). In these repeated measures studies, CPT and CPT-II test–retest intervals have varied widely, ranging from multiple (i.e., up to five) administrations in 1 day (Schachar et al., 2008) to test–retest intervals of 1 year or longer (Borgatti et al., 2004).

Far less is known about the clinical utility of the Conners' CPT-II and other continuous performance tests for quantifying *intra-individual* changes in attention and neurocognitive status over time, but preliminary studies using other CPT measures suggest potential pitfalls associated with this practice. When assessing children with ADHD in non-treatment conditions using 2- and 4-month test–retest intervals, Llorente et al. (2001) demonstrated significant intra-individual test–retest variability and poor individual test–retest score agreement on the omission and commission variables of a continuous performance test (Tests of Variables of Attention, TOVA; Greenberg & Kindschi, 1996). This pattern of poor intra-individual test–retest score agreement was noted despite group indicators of robust internal consistency and "satisfactory" temporal stability (test–retest reliability) of the measure. Llorente et al. (2001) proposed that the high level of intra-individual test–retest score variability noted in their sample may have been attributable to the clinical symptoms of children with ADHD. The possibility remains, however, that intra-individual test–retest variability is a specific psychometric property of CPT measures as well as the assessment of constructs such as sustained attention in general. If such were the case, it would present clinicians and diagnosticians with the challenge of distinguishing meaningful from insignificant changes in test scores within a wide range of both true and error variance.

The present study was designed to further investigate the intra-individual test–retest stability of continuous performance measures in serial neuropsychological assessment. To this end, the current study examined data from a carefully screened sample of 39 typically developing children and adolescents who had been administered a neuropsychological test battery that included the CPT-II on two occasions. While larger than the test–retest interval of the reliability sample reported in the CPT-II manual, the test–retest interval used for the current study (mean = 6 months) was designed to approximate test–retest intervals routinely used in the clinical monitoring of medically involved children with potentially unstable neurologic conditions. Additionally, rather than collapsing CPT-II scores into group mean averages for analysis, we examined the consistency of each participant's scores at Time 2 relative to his or her scores at Time 1. The proportion of participants with statistically rare test score changes was then used in the statistical analysis. In this sample of typically developing children and adolescents we hypothesized that the observed frequency of participants with statistically rare changes in CPT-II test scores would match the frequency of statistically rare scores expected based upon the assumptions of the RCI method employed. Specifically, we hypothesized that approximately 10% of scores at assessment Time 2 would fall outside of the 90% band of confidence created around the corresponding Time 1 scores using RCI methodology.

METHOD

Participants

Participants were recruited from the greater Baltimore area to serve as controls as part of a larger study of children with neurological disorders. The control group of healthy, typically developing children and adolescents was recruited by advertisement. All participants and parents signed a consent form that met the Institutional Review Board standards of the Johns Hopkins Medical Institutions.

Participants in the control group (hereafter referred to only as "participants") were initially screened via telephone interview with a parent, and were excluded if there was a prior history of psychiatric or neurological disorder, intellectual disability, language disorder, or learning disability. Participants were further screened via structured psychiatric interview (described below) prior to testing, and participants were excluded from the study if they met DSM-IV criteria for any psychiatric disorder. Additionally, participants were also excluded if they had an IQ of <70 as measured by the Peabody Picture Vocabulary Test – Third Edition (PPVT-III; Dunn & Dunn, 1997a). Of the 98 children originally recruited for the control group, 54 healthy volunteers met inclusion and exclusion criteria. In order to examine the reliability of the CPT-II over an extended test–retest interval, the current study included only those participants who underwent reassessment within 3 to 8 months after baseline assessment. A total of 15 participants were excluded due to incomplete data sets (i.e., computer malfunction occurring during administration of the CPT-II or missed Time 2 visits). There were no differences in SES or estimated IQ between the 15 children excluded and those in the final sample. The final sample included 39 healthy participants (19 boys, 20 girls) ranging in age from 6 to 18 years at the time of first assessment (mean = 12.0 ± 3.7 years).

Materials and procedures

Once enrolled, all participants completed a baseline neuropsychological assessment battery that included measures of attention, language, and visual and motor skills. For the purposes of the broader clinical research context (i.e., monitoring late effects of cancer treatment over time), the test battery was intentionally created to be both brief and repeatable. The complete test battery was composed both of individual tests (e.g., PPVT-III) and selected subtests from more comprehensive test batteries (e.g., Block Design subtest from the WISC-IV). Parents of the participants also completed behavior-rating scales at the time of baseline neuropsychological testing. Four screening measures were used, including an indicator of socioeconomic status (Hollingshead, 1975), a semi-structured psychiatric interview (Diagnostic Interview for Children and Adolescents Fourth Edition—DICA-IV; Reich, Welner, & Herjanic, 1997), a parent rating measure of behavior (Child Behavior Checklist—CBCL/6-18; Achenbach & Rescorla, 2001), and an estimate of verbal IQ (PPVT-III).

Several tasks from the broader clinical research battery were selected for investigation in the current study. In addition to the PPVT-III, two additional well-normed and commonly used neuropsychological measures were used to assess the stability of test performance over the test–retest interval, and were included as contrasts to the CPT-II: the Block Design subtest from the Wechsler Intelligence Scale for Children, Fourth Edition—(WISC-IV; Wechsler, 2003) and the Beery Developmental Test of Visual Perception, Fifth Edition (Beery & Beery, 2004). Finally, the Conners' CPT-II (Conners, 2000) was administered to assess the reliability of this measure over an extended test–retest interval using typically developing children. All measures were re-administered an average of 6.4 months following baseline assessment (range = 3 to 8 months).

Screening/descriptive measures

Hollingshead Index—Socioeconomic status for each participant was estimated by a widely used four-factor index (Hollingshead, 1975).

Diagnostic Interview for Children, Fourth Edition (DICA-IV; Reich et al., 1997)—Parents of children deemed eligible via telephone screen were administered the DICA-IV, which is based on the Diagnostic and Statistical Manual of Mental Disorders – Fourth Edition (DSM-IV; American Psychiatric Association, 1994). This is a semi-structured interview that is designed for determining selected current and retrospective psychiatric diagnoses including attention deficit hyperactivity disorder, conduct disorder, oppositional defiant disorder, major depressive disorder, mania/hypomania, dysthymic disorder, separation anxiety disorder, panic disorder, generalized anxiety disorder, specific phobia, and obsessive compulsive disorder. The DICA-IV has been reported to be reliable for DSM-IV diagnoses. Children who met DSM-IV criteria for any psychiatric disorder were excluded from the study.

Conners' Parent Rating Scale – Revised, Long Form (CPRS-R; Conners, 1997)—The revised Conners' Rating Scales are parent reports of child behavior that probe conduct problems, learning problems, psychosomatic problems, impulsivity-hyperactivity, anxiety, and social competence. The standardization samples for the parent scales were drawn from over 2000 parents of children aged 3–17. The scales produced by the revised Conners' Rating Scales correspond with symptoms used in the DSM-IV criteria for ADHD.

Child Behavior Checklist (CBCL/6-18; Achenbach & Rescorla, 2001)—The CBCL/6-18 is a broadband parent rating scale that examines behavioral and adaptive functioning. The scale provides scores on three competence scales (Activities, Social, and School), total competence, eight syndromes, and internalizing, externalizing, and total problems. The syndromes include aggressive behavior, anxious/depressed, attention problems, rule-breaking behavior, social problems, somatic complaints, thought problems, and withdrawn/depressed. The DSM-oriented scales include: Affective Problems, Anxiety Problems, Somatic Problems, Attention Deficit/Hyperactivity Problems, Oppositional Defiant Problems, and Conduct

Problems. The scales are based on factor analyses of parents' ratings of 4994 clinically referred children, and are normed on 1753 children aged 6 to 18, using a representative sample from the 48 contiguous states stratified for SES, ethnicity, region, and urban-suburban-rural residence.

Peabody Picture Vocabulary Test, Third Edition (PPVT-III; Dunn & Dunn, 1997a)—The PPVT-III is a screening test of verbal ability and a measure of receptive (i.e., listening) single-word vocabulary attainment for standard English. The child is shown a page with four pictures and the examiner provides the child with a vocabulary word. The child is asked to identify the picture that best describes the word either by pointing or verbalizing the number of the picture. Thus the test requires little to no motor or expressive language output. The child continues the test until 8 of 12 items are missed in an item set. Standard scores were used to describe the sample. The PPVT-III manual reports that it is highly correlated (.90) with WISC Full Scale IQ. The PPVT-III manual also reports uncorrected test–retest reliability coefficients for groups of selected ages (6–0 to 10–11; 12–0 to 17–11) ranging from .88 to .91 using a test–retest interval of approximately 1 month (Dunn & Dunn, 1997b). Children in the current study were excluded if they had PPVT-III standard scores lower than 70.

Additional neuropsychological measures

Wechsler Intelligence Scale for Children – Fourth Edition: Block Design Subtest (Wechsler, 2003)—The WISC-IV was normed on 2200 children, aged 6–16 years, representing a nationally representative stratified standardization sample with 100 boys and 100 girls included in each of the 11 age levels. The test–retest reliability of the WISC-IV has been examined using intervals ranging from 13 to 63 days, and a mean interval of 32 days. The average corrected stability coefficient of the Block Design Subtest for the standardization sample was .81.

Beery Developmental Test of Visual Perception (Beery & Beery, 2004)—On the visual perception test, the child is shown one geometric form and is asked to choose the geometric form that is exactly the same from a group of forms within a 3-minute time limit. For example, the child is shown a target stimulus of a Necker cube and is asked to select (by pointing or circling) the exact match of the cube among five choices, four of which may be smaller, missing parts, rotated, etc. Thus, the test is designed as a measure of motor-free visual-perceptual skills. The Beery VMI-5 manual reports test–retest raw score reliability coefficient of .85 for the Developmental Test of Visual Perception using an average of a 10-day test–retest interval. The sample was composed of 115 children aged 5–11 "in regular public school classrooms with full ranges of student abilities and proportionate numbers of children with disabilities" (Beery & Beery, 2004, p. 102).

Neuropsychological measure of sustained attention

Conners' Continuous Performance Test – Second Edition (CPT-II; Conners, 2000)—This is a computerized measure of vigilance/attentional control and response inhibition for children aged 6 and older. The CPT-II requires respondents to press the space bar or click a mouse button when any letter appears on the screen, but to not press the space bar or button when the target letter "X" appears. Stimuli are presented in six blocks with three sub-blocks, each containing 20 trials (i.e., letter presentations). Inter-stimulus intervals (ISIs) vary between 1, 2, and 4 seconds, while the display time is held constant at 250 milliseconds. Altogether, a CPT-II administration takes 14 minutes to complete.

Guidelines presented in the CPT-II test manual indicate that *ideally* the CPT-II should be administered twice (or more) to establish a baseline before treatment is initiated (Conners & MHS, 2004). These guidelines are proposed due to the potential for regression to the mean in the serial assessment process, in which improved test scores may occur simply because the initial (non-treatment) exasperating conditions "could only get better." This guideline (i.e., two test administrations at baseline) was not employed for the current study because this was not a treatment response study, and the screening process of this cohort of typically developing youth did not reveal any overt behavioral or cognitive conditions that would have increased the likelihood of regression to the mean.

In the published CPT-II technical manual, test–retest reliability is reported from a sample of 23 individuals, 10 non-clinical and 13 "with a variety of clinical diagnoses" (Conners & MHS, 2004, p. 56). The mean age for the test–retest sample at assessment time 1 was 27.7 years. Test–retest interval was approximately 3 months, and correlations between time 1 and time 2 parameters for the variables considered in the current study ranged from .55 (Hit Reaction Time) to .84 (Omissions). The CPT-II technical manual reports that the significance of observed test score changes (based on Jacobson-Truax criteria) is presented in the multiple administrations computerized score printout, although limited information is presented regarding how this is accomplished.

While the CPT-II yields a number of variables, the current study focused on five parameters: *Omissions* indicates the number of targets to which the individual did not respond; *Commissions* refers to the number of times that the individual responded to a non-target; *Hit Reaction Time* is the mean response time for all correct target hits over the CPT-II administration; *Variability* is a measure of how much an individual's hit reaction time varied over the entire length of the test, thus providing a measure of intra-individual variability. *Attentiveness* (*d*′) indicates how well the individual discriminates between targets and non-targets (Conners & MHS, 2004). These selected variables are well represented in the literature as measuring different aspects of attention (Ackerman et al., 2008; Gruber et al., 2007; Homack & Riccio, 2006; Kiehna et al., 2006; Molteni, Bianchi, Butti, Reni, & Zucca, 2008; Ogg et al., 2008; Olsson, von Scheele, & Panossian, 2008), and have better test–retest reliability than most other CPT-II variables (Conners & MHS, 2004). In addition to these five variables of interest, the current study also included the *ADHD Confidence Index* as a qualitative means of describing the study sample. This index is based on a discriminate function analysis calculated by the CPT-II software, with a best-fit classification of the respondent into one of three categories—i.e., Clinical [ADHD], Non-Clinical, or No Decision (Conners & MHS, 2004).

Of note, for the purposes of this study the CPT-II was administered on three different PC computers, all of which met the published CPU/RAM requirements identified in the CPT-II manual. These were multi-use computers that were not strictly designated/reserved for this study. Thus sources of error that are specific to computerized assessment measures may have impacted the results (Cernich, Brennana, Barker, & Bleiberg, 2007).

Data analysis

Test score validity considerations were made according to a guideline proposed in the CPT-II manual. Specifically, the CPT-II manual suggests that extreme T-scores (T-score > 100) on the Omissions scale may indicate an invalid protocol (Conners & MHS, 2004). Participant CPT-II data from *both* assessment Time 1 and Time 2 were not included in the statistical analysis if T-scores from the Omissions scale were greater than 100 at Time 1 *or* Time 2.

Two-way random-effects intraclass correlational coefficients (ICC) were calculated to determine the test–retest reliability of each CPT-II variable and each neuropsychological comparison measure. The ICC is a univariate measure that estimates the level of agreement between scores on the same test at two points in time (Bland & Altman, 1986).

Skewness and kurtosis were calculated to determine if data for each test variable were normally distributed. For those variables in which data were not normally distributed, non-parametric analyses (Wilcoxon signed ranks tests) were used to determine if statistically significant differences existed between means from Time 1 and Time 2 for each variable. Paired *t*-tests were used for the same purpose on variables without abnormal skewness or kurtosis.

Further analysis was conducted using a modified Reliable Change Index (RCI). RCI analysis was used to determine whether an unexpected number of statistically rare changes in CPT-II scores occurred from baseline to follow-up assessment. The RCI is a means for determining the amount of change in each individual's test score that might be expected if no actual systematic treatment, injury, or unintended threat to validity occurred in the inter-assessment interval.

Using standard error of measurement (*SEM*) values published in the CPT-II manual for each CPT-II variable (Conners & MHS, 2004), we calculated standard error of the difference (SE_{diff}) values, and from these calculated RCI values ($1.64 \times SE_{\text{diff}}$). The RCI values were then used to create 90% change score confidence intervals, within which we would expect chance variations in score to occur 90% of the time. Random changes (either declines or increases) in test score at follow-up assessment that exceeded the RCI confidence interval created around the baseline test score were only expected to occur 10% of the time in a typically developing sample of children and adolescents such as the one used in the present study.

Several clarifications are in order regarding our use of RCI. First, the first phase of examination of test score stability using RCI was conducted using *raw* test scores from the CPT-II. This was done because the published age-referenced *SEM* values found in the CPT-II manual are only reported in raw score form, and *SEM* values for CPT-II standard score variables are not commercially provided by the test publisher. Use of raw score data to calculate RCI is potentially problematic, as the CPT-II computerized scoring program conducts logarithmic transformations of all raw data involving reaction times (Conners & MHS, 2004, p. 17). Raw score data were nonetheless employed for RCI calculations, as age-referenced *SEM* values for CPT-II (post-transformation) standardized scores are not published in the CPT-II manual or related publications.

Second, as a follow-up procedure, we calculated *SEM* values for each of the variables of interest using the reliability coefficients (based on T-scores) provided in the CPT-II test manual (Conners & MHS, 2004, p. 57). Reliability coefficients reported in the CPT-II manual are reported for the entire test–retest sample (mean age 27.7 years) rather than according to discrete age categories, thus reducing their utility for our developmental sample. Nonetheless, the resulting *SEM* values were then used in a second phase of RCI examination of intra-individual stability of CPT-II T-scores in our sample of typically developing children. This follow-up procedure permitted analysis of standardized CPT-II scores as well as raw scores (described above).

Third, for the RCI calculations using *raw* CPT-II test scores, we used a modified RCI method described by Chelune (2002) that utilized discrete *SEM* values corresponding to the age of each individual at the time of each assessment. This is a departure from standard RCI calculation as first described by Jacobson and Truax (1991), which calculates the SE_{diff} using the *SEM* available for the age of the participant at the time of baseline assessment (depicted below):

$$SE_{\text{diff}} = (2(SEM)^2)^{1/2}$$

In contrast, the modified RCI (Chelune, 2002) accounts for different *SEM* values that correspond to the changing age of the participant. For instance, if a participant advanced in age from 9 to 10 during the test–retest interval, the SE_{diff} of modified RCI would be calculated using the test's *SEM* for 9-year olds (SEM_1) and 10-year olds (SEM_2) (depicted below):

$$SE_{\text{diff}}=[(SEM_1)^2+(SEM_2)^2]^{1/2}$$

Finally, it should be noted that we did not make an adjustment to account for potential practice effects in the repeated administration of the CPT-II. While different types of practice adjustment have been proposed for use with RCI (Chelune, 2002), practice adjustment was not implemented in the design of the present study, as the CPT-II manual indicates that the CPT-II is "relatively unaffected by practice effects" (Conners & MHS, 2004, p. 58).

In the first phase of RCI analysis we determined if each individual participant's change in raw score—i.e., difference score from baseline (Time 1) to follow-up (Time 2)—for each CPT-II variable of interest exceeded the 90% RCI confidence interval. For each CPT-II variable we then used chi-square analysis to determine if the observed number of participants with "statistically rare" changes in raw scores exceeded the number expected (i.e., n = 3.7; ~10% of 37 typically developing participants) based on RCI assumptions.

We repeated this procedure for the second phase of RCI analysis, this time examining the change in standardized (T) scores. As noted earlier, RCI confidence intervals for this phase of RCI analysis were established using *SEM*s calculated from the broad age-range reliability coefficients reported in the CPT-II technical manual.

RESULTS

Demographic information

Demographic characteristics of the 39 participants are summarized in Table 1. The mean age of the study sample at the time of baseline assessment was 12.0 years (standard deviation = 3.7; range = 6.3–18.4 years); 49% of the sample was male; the racial composition was 44% Caucasian, 41% African-American, 13% biracial, and 3% Asian; and 80% of the sample was right-handed. The mean T-scores on the CBCL Anxious/Depressed and Attention Problems scales were 51.9 and 51.6, respectively, with none of the participants' CBCL scores exceeding a score of 65 on either scale. The average T-score on the CPRS ADHD Index was 45.0, with none of these scores exceeding a T-score of 56. These scores are shown in Table 1.

Eight of the participants were between the ages of 16 and 18 at Time 1 and were outside of the age parameters of the WISC-IV Block Design subtest. As such, reliability coefficients for Block Design are based only on the performance of the 31 participants in the WISC-IV age range.

General validity considerations

The majority of standardized CPT-II Omissions scores fell well below a T-score of 100, with only 1 of 39 participants generating a T-score over 100 (T = 100.58) at assessment Time 1, and only one (different) participant generating such as score (T = 129) at assessment Time 2. As these scores did not meet commonly used standards of test score validity, the CPT-II data (collected at assessment Time 1 and Time 2) from these two participants were not included in subsequent statistical analysis of CPT-II variables. As such, data analysis involving CPT-II variables were conducted using a more restricted sample of 37 children and adolescents.

Qualitative description of CPT-II performance

At Times 1 and 2 the number and percentage (in parentheses) of various clinical classifications assigned to participants using the CPT-II ADHD Confidence Index was as follows: Non-Clinical at baseline: 19 (51.4%); Non-Clinical at follow-up: 16 (43.2%); No Decision at baseline: 6 (16.2%); No Decision at follow-up: 6 (16.2%); Clinical at baseline: 12 (32.4%); Clinical at follow-up: 15 (40.5%).

A total of 23 participants (62%) had no change in ADHD Confidence Index classification from Time 1 to Time 2, 6 participants (16.2%) changed from No Decision or Non-Clinical to Clinical, 3 participants (8.1%) changed from No Decision or Clinical to Non-Clinical, and 5 participants (13.5%) changed from Clinical or Non-Clinical to No Decision. At Times 1 and 2 (Table 2), the number and percentage (in parentheses) of T-scores in the elevated range (T > 60) for each of the variables of interest was as follows: Omissions baseline: 3 (8.1%); Omissions follow-up: 9 (24.3%); Commissions baseline: 2 (5.4%); Commissions follow-up: 7 (18.9%); Variability baseline: 8 (21.6%); Variability follow-up: 11 (29.7%); Hit Reaction Time baseline: 8 (21.6%); Hit Reaction Time follow-up: 8 (21.6%); Attentiveness baseline: 1 (2.7%); and Attentiveness follow-up: 8 (21.6%).

Distribution of scores

None of the test score distributions for Block Design, PPVT-III, or Beery Test of Visual Perception tests was found to have significant skew or kurtosis. In contrast, the T-score distributions of four of the five CPT-II variables of interest (all but Variability) were characterized by significant skew and/or kurtosis, either at Time 1 and/or at Time 2. Similarly, the raw score distributions of four of the five CPT-II variables of interest (all but Commissions) were characterized by significant skew and/or kurtosis at Time 1 and/or Time 2.

Practice effect

There were no statistically significant differences between mean group performances at Time 1 and Time 2 for any of the variables examined for this study, using standard scores from the CPT-II, Block Design, PPVT-III, and Visual Perception (Table 3), or raw scores from the CPT-II. Moreover, there was no evidence of systematic effects between individual change scores (absolute values) and age at time of assessment or length of test–retest interval.

Test–retest reliability

Results of test–retest reliability analyses for standardized scores are listed in Table 3. The test–retest reliability ICC values for standard scores from the neuropsychological comparison tests (PPVT-III, WISC-IV Block Design, and Visual Perception) were all strong, ranging from .82 to .92. In contrast, test–retest reliability ICC values for most of the variables (T-scores) of the CPT-II were below 0.6, with the exception of Hit Reaction Time (0.65). For CPT-II *raw* scores, reliability coefficients were as follows: Variability (0.62), Attentiveness (0.64), Omissions (0.69), Hit Reaction Time (0.75), and Commissions (0.82).

Reliable change indices for CPT-II raw scores

RCI analyses using CPT-II raw scores are listed in Table 4. Using a 90% confidence interval in our modified RCI calculations, we anticipated that ~10% of our sample (i.e., approximately four individuals) would display a statistically rare change in raw test score for each of the five selected CPT-II variables. Chi-square analysis indicated a significantly higher than expected number of statistically rare changes in *raw* test score for three of the five CPT-II variables (Omissions, Variability, Attentiveness). Using this method, Hit Reaction Time and Commission errors were the only CPT-II variables in which the number of statistically rare changes in *raw* test score did not significantly exceed that which was expected.

Reliable change indices for CPT-II standardized (T) scores

Of this sample of 37 children and adolescents, 32.4% had one statistically rare change in T-score, 18.9% had two, 8.1% had three, 0% had four, and 2.7% (1 participant) had five. Using the RCI procedure noted above, we investigated whether there were an unusually high number of participants with statistically rare changes in CPT-II T-scores for each individual scale. RCI analyses using CPT-II standardized (T) scores are listed in Table 4. Chi-square analysis indicated a significantly higher than expected number of statistically rare changes in T-score values for two of the five CPT-II variables, i.e., Omissions and Attentiveness.

Of the 11 individuals with statistically rare changes in Omissions T-score, seven had score increases from baseline (Time 1) to follow-up (Time 2). Six of these Omissions score increases involved changes in score from the average/normal range to the Moderately atypical (T-score = 60–64) or Markedly atypical (T-score > 65) ranges, and represented change in test score *away* from the mean. Of the four statistically rare reductions in Omissions T-score, three scores were in the clinically elevated range (T-score > 60) at baseline, and two of these declined back to normal limits during follow-up assessment. All three of these noted T-score reductions represented change in test score *in the direction of the mean*. In summary, 8 of the 11 changes in Omission T-scores would likely have been interpreted as both statistically and clinically meaningful (e.g., decline or improvement in score) in a clinical setting.

Of the 11 individuals with statistically rare changes in Attentiveness T-score, 6 had score increases from baseline (Time 1) to follow-up (Time 2). Three of these Attentiveness score increases involved changes in scores from average/normal to Moderately or Markedly atypical, and represented change in test score *away* from the mean. Of the five statistically rare reductions in Attentiveness T-score, only one of these scores was in the clinically elevated range (T-score > 60) at baseline, and subsequently declined back to normal limits during follow-up assessment. In summary, 4 of the 11 changes in Attentiveness T-scores would likely have been interpreted as both statistically and clinically meaningful (e.g., decline or improvement in score) in a clinical setting.

DISCUSSION

The present study examined a popular computerized measure of attention and investigated its test–retest reliability when used in the serial assessment of children. While the reliability and related psychometric properties of the CPT-II and other continuous performance measures have periodically been investigated using clinical populations, this study sought to examine reliability using a typically developing pediatric cohort. Doing so allowed an examination of “normal” variations in test performance without the presence of overt clinical symptoms (e.g., inattention, impulsivity) thought to disrupt the consistency of neuropsychological test performance in general. Moreover, this study was designed to allow investigation of commonly held assumptions regarding the rate of aberrant changes in the test scores of typically developing children.

In our sample of typically developing youth, mean average raw scores and T-scores on the CPT-II variables of interest were comparable at baseline and follow-up, suggesting no evidence of a practice effect. Despite the general stability of mean scores over time, the observed test–retest reliability coefficients of the CPT-II scores over this extended test–retest interval (~6 months) were low. Specifically, while time-linked performance on the comparison neuropsychological measures (i.e., WISC-IV Block Design subtest, PPTV-III, and Beery Test of Visual Perception) resulted in acceptable reliability coefficients (ICC = .83 to .92), the reliability coefficients of the CPT-II variables of interest were considerably more modest (raw score ICCs ranging from .62 to .82; T-score ICCs ranging from .33 to .65). These modest CPT-II reliability coefficients were, to a degree, expected given differences in the types of tests used in the study. Specifically, the comparison measures were primarily power tests, i.e., tests of traits/content that provide time limits sufficient for the potential completion of all test items (Anastasi & Urbina, 1997). In contrast, the CPT-II maintains a combination of both power and speed components in its composition. Constructs such as attention that are measured by speed tests may involve more day-to-day or moment-to-moment "normal" variability, thus contributing to lower reliability estimates in general. As such, it was not unexpected to find that test–retest reliability coefficients for the CPT-II were considerably lower than those of the more stable power-based comparison tests.

What *was* more unexpected, however, was the rate at which statistically rare changes in CPT-II test score occurred in our sample of typically developing children. Using the RCI methodology and assumptions described above, we expected error variance and chance variation to result in statistically rare changes in raw and standardized test score approximately 10% of the time (based on a 90% confidence interval). The modified RCI calculations were made using SEM values (raw score) reported in the CPT-II manual, or from SEM values (T-score) derived from test–retest reliability data published in the CPT-II manual. As such, the reported or calculated SEM values were presumed to already reflect the increased variability associated with test–retest performance and the speed components of the CPT-II. In fact, an unusually high proportion of statistically rare changes were identified in the CPT-II Omissions and Attentiveness scores within this typically developing pediatric sample. Proportions of statistically rare changes in score were considerably higher when RCI calculations were conducted using raw CPT-II scores, but remained much higher than expected (i.e., ~30% of sample) when *T-score* changes were examined as well.

The findings from this study have potential implications for clinical and forensic practice, particularly for those clinicians who employ an "$n = 1$" empirical approach to neuropsychological assessment. While information necessary for RCI analysis is becoming increasingly available, many clinicians continue to employ "fixed standards" of significance (e.g., test score changes of 10 T-score points or higher) when interpreting test score changes in serial assessment. While problematic in general, this "fixed standard" approach could potentially lead to the false identification of cognitive skill increases or declines (Type I error) when using scores from the CPT-II and/or less reliable speed-based tests. To illustrate this point, 90% confidence intervals were calculated using RCI for the CPT-II variables of interest and the comparison neuropsychological measures. The broad age range (6 to 18) reliability coefficients (based on T-scores) from the present study were used in the RCI calculations, rather than the published CPT-II test–retest reliability information. This resulted in a large range of RCI confidence intervals between different tests, as well as wide variations in related criteria for determining the statistical significance of test score changes. For instance, using the reliability data from the current study to assess change in the test scores of typically developing youth, a standard score change comparable to two-thirds of a standard deviation on the PPVT-III would be considered statistically rare, while a T-score change comparable to one and two thirds standard deviations on the CPT-II Attentiveness (*d*′) variable would not. For the CPT-II variables of interest, the current study suggests that the "thresholds" of statistically rare changes in standardized test scores for typically developing youth range from ~1.3 standard deviations (Hit Reaction Time) to approximately 1.8 standard deviations (Attentiveness).

While the thresholds for significant change on the CPT-II noted above demonstrate the broad range of "normal" test score variation in typically developing children, it is unclear if this finding can be generalized to clinical populations. Interestingly, the CPT-II reliability findings from the current study were somewhat lower than expected given available reliability estimates for another continuous performance measure obtained using a sample of children with ADHD (i.e., TOVA; 2- and 4-month test–retest intervals; $r = .51$ to $.75$; Llorente et al., 2001). These reliability distinctions may reflect a difference between CPT tests, but may also suggest *increased* stability of CPT test scores *in clinical populations* (relative to typically developing controls) when scores are obtained over extended test–retest intervals. Given the potential for restriction of range in score, it is quite possible that individuals from clinical populations who obtain deficient scores on attentional measures at baseline will continue to do so when tested at later follow-up assessment points, thus contributing to increased (rather than decreased) stability of their test scores over time. If such were the case, reliability coefficients derived from samples of typically developing children may contribute to RCI confidence intervals that are actually prohibitive (excessively large) for detecting “true” cognitive changes in neurologically involved clinical populations.

To illustrate this point, consider the case of a child with shunted hydrocephalus and past white matter injury. If his or her scores on the CPT-II are already elevated at the time of his or her non-acute baseline assessment, change in CPT-II scores associated with acute shunt failure may not be of the same magnitude as those of a typically developing child with new onset hydrocephalus. While we espouse the use of RCI methodologies in pediatric neuropsychology, separate clinical group norms may be necessary in order to establish RCI confidence intervals that are sensitive to true neurologically based changes in cognitive status. To this end, further follow-up studies in this area will be necessary to address this question, with direct comparison of clinical patients and typically developing children over more extended test–retest intervals.

The potentially high base-rates of CPT-II test score changes suggested by the current study could certainly complicate clinical interpretation of test score gain or loss. This could be particularly the case when using the CPT-II to address questions of cognitive change associated with medical conditions and/or procedures such as temporal lobectomy, chemotherapy/radiation, shunt failure/revision, etc. To address these assessment-related concerns, we propose the following next steps:

1. *Special population norms for clinical groups:* The current study demonstrates the wide range of “normal” test score change in typically developing youth, and raises empirical questions regarding CPT-II test stability when it is administered to clinically involved populations. We suggest examining CPT-II test stability in children with potentially unstable neurologic conditions in order to determine the extent to which patterns of test performance are more, less, or equally stable when compared to those of typically developing youth. Publication of special population norms for tests like the CPT-II will provide neuropsychologists with psychometric information necessary for a more refined detection of clinically and statistically meaningful changes in intra-individual test score performance.

2. *Reliability and standard error of measurement* (SEM) *values for standard scores* (*T-scores*): While the CPT-II manual provides *SEM* values for discrete age groupings (e.g., ages 6–7, 8–9, etc.), these values were reported in *raw* score terms, and could only be used for the test–retest comparison of *raw* CPT-II scores rather than standardized T-scores. In our RCI analysis, there was little support for the intra-individual comparison of *raw* score values over time, as we found that this practice contributed to an extremely high (and likely spurious) base rate of statistically rare intra-individual changes in test score. In contrast, establishment of reliability and *SEM* values for discrete age-groupings of T-score values will likely have considerable value to clinicians and researches alike.

3. *Additional investigation into intra-individual reaction time and variability*: It is noteworthy that T-scores from three of the five CPT-II variables of interest *did* conform to the test–retest assumptions of the RCI method employed. If intra-individual test–retest stability of the Hit Reaction Time and Variability variables can

be demonstrated in clinical populations and/or replicated in typically developing youth, these may prove to be key variables for detecting meaningful intra-individual declines or improvements in the neurocognitive functioning of children. These noted variables are of particular interest and potential utility, as they have been found to be strongly/reliably linked to the symptomatology of ADHD and sensitive to stimulant medication effects (Epstein et al., 2003; Riccio, Reynolds, & Lowe, 2001).

4. *Further empirical investigation of test administration practices:* The procedures used in this study introduced several potential sources of error and threats to validity that resulted in several important limitations. While each of these may have detracted from the reliability of specific CPT-II variables under the experimental conditions, the procedures used are thought to be consistent with those under which many clinicians and researchers practice. As such, additional empirical questions are raised regarding the negative impact of each of these noted practices. First, the test administration procedures did not conform to suggested guidelines presented in the CPT-II test manual indicating that the CPT-II should *ideally* be administered twice (or more) to establish a baseline. While this practice (i.e., testing until asymptote is reached) is suggested in the manual, it is often considered impractical for use in clinical practice. Further empirical investigation will be useful for determining if collection of multiple baseline assessments results in improved intra-individual CPT-II performance stability over time, and if this procedure increases the sensitivity of the measure for detecting neurologically based changes in cognitive status. Second, different computers were used in this study, and in some instances, baseline and follow-up CPT-II administration for the same individual occurred on different computers. Similarly, the computers that operated the CPT-II in the study were multi-use computers (rather than exclusively designated for CPT-II use), which may have reduced the capacity of the program to operate with millisecond timing. It is possible that variables (i.e., Hit Reaction Time and Variability) thought to be most vulnerable to this type of instrumentation-linked threat to validity would be even more stable under stricter assessment conditions.

In summary, this study raises a number of questions regarding currently used methods for assessing changes in attentional functioning over time. Using the less-than-optimal yet *realistic* conditions under which the CPT-II was administered, the current study provides evidence of considerable variability in some aspects of intra-individual test performance, and general stability in others. Further investigation of this issue is of considerable importance, and will be necessary to refine the ability of neuropsychologists to contribute to the care and monitoring of children with potentially unstable neurological conditions.

Acknowledgments

The authors wish to thank Alena Horska, Ph.D., for her support of this work. Supported by R01 NS04285 and HD-24061 (Developmental Disabilities Research Center).

REFERENCES

Achenbach, TM.; Rescorla, LA. Manual for the ASEBA preschool forms and profiles. Burlington: University of Vermont, Research Center for Children, Youth, & Families; 2001.

Ackerman JP, Llorente AM, Black MM, Ackerman CS, Mayes LA, Nair P. The effect of prenatal drug exposure and caregiving context on children's performance on a task of sustained visual attention. Journal of Developmental and Behavioral Pediatrics 2008;29(6):467–474. [PubMed: 19047916]

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Washington, DC: APA; 1994.

Anastasi, A.; Urbina, S. Psychological testing. 7th ed. Upper Saddle River, NJ: Prentice-Hall; 1997.

Baron, IS. Neuropsychological evaluation of the child. New York: Oxford University Press; 2004.

Beery, KE.; Beery, NA. Developmental Test of Visual Perception. 5th ed. Minneapolis, MN: Pearson, Inc.; 2004.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1(8476):307–310. [PubMed: 2868172]

Borgatti R, Piccinelli P, Montirosso R, Donati G, Rampani A, Molteni L, et al. Study of attentional processes in children with idiopathic epilepsy by Conners' Continuous Performance Test. Journal of Child Neurology 2004;19(7):509–515. [PubMed: 15526955]

Boro E, Vahip S, Akdeniz F. Sustained attention deficits in manic and euthymic patients with bipolar disorder. Progress in Neuropsychopharmacology & Biological Psychiatry 2006;30(6):1097–1102.

Cernich AN, Brennana DM, Barker LM, Bleiberg J. Sources of error in computerized neuropsychological assessment. Archives of Clinical Neuropsychology 2007;22 Suppl 1:S39–S48. [PubMed: 17097851]

Chelune GJ, Naugle RI, Lüders H, Sedlak J, Awad IA. Individual change after epilepsy surgery: Practice effects and base-rate information. Neuropsychology 1993;7(1):41–52.

Chelune, GJ. Assessing reliable neuropsychological change. In: Franklin, R., editor. Prediction in forensic and neuropsychology: Sound statistical practices. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 2002. p. 223-248.

Conners, CK. Conners' Rating Scales – revised. Niagara Falls, NY: Multi-Health Systems; 1997.

Conners, CK. Conners' Continuous Performance Test. 2nd ed. Toronto, Canada: Multi-Health Systems, Inc.; 2000.

Conners, CK. Multi Health Systems. Conners' Continuous Performance Test II: Technical guide for software manual. New York: Multi-Health Systems; 2004.

Dunn, LM.; Dunn, LM. Peabody Picture Vocabulary Test. 3rd ed. Bloomington, MN: Pearson Assessments; 1997a.

Dunn, LM.; Dunn, LM. Peabody Picture Vocabulary Test (3rd ed.): Examiner's manual. Bloomington, MN: Pearson Assessments; 1997b.

Epstein JN, Conners CK, Hervey AS, Tonev ST, Arnold LE, Abikoff HB, et al. Assessing medication effects in the MTA study using neuropsychological outcomes. Journal of Child Psychology and Psychiatry 2006;47(5):446–456. [PubMed: 16671928]

Epstein JN, Erkanli A, Conners CK, Klaric J, Costello JE, Angold A. Relations between Continuous Performance Test (CPT) performance measures and ADHD behaviors. Journal of Abnormal Child Psychology 2003;31(5):543–554. [PubMed: 14561061]

Greenberg, LM.; Kindschi, CL. Test of Variables of Attention: Clinical guide. Los Alamitos, CA: Universal Attention Disorders; 1996.

Gruber R, Grizenko N, Schwartz G, Bellingham J, Guzman R, Joober R. Performance on the Continuous Performance Test (CPT) in children with ADHD is associated with sleep efficiency. Sleep 2007;30 (8):1003–1009. [PubMed: 17702270]

Hollingshead, AB. Four factor index of social status. New Haven, CT: Yale University, Department of Sociology; 1975.

Homack S, Riccio CA. Conners' Continuous Performance Test (2nd ed.; CCPT-II). Journal of Attention Disorders 2006;9(3):556–558. [PubMed: 16481673]

Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Clinical and Consulting Psychology 1991;59(1):12–19.

Kiehna EN, Mulhern RK, Li C, Xiong X, Merchant TE. Changes in attentional performance of children and young adults with localized primary brain tumors after conformal radiation therapy. Journal of Clinical Oncology 2006;24(33):5283–5290. [PubMed: 17114662]

Lezak, MD. Neuropsychological Assessment. 3rd ed. New York: Oxford University Press; 1995.

Llorente AM, Amado AJ, Voigt RG, Berretta MC, Fraley JK, Jensen CL, et al. Internal consistency, temporal stability, and reproducibility of individual index scores of the Test of Variables of Attention (TOVA) in children with attention-deficit/hyperactivity disorder. Archives of Clinical Neuropsychology 2001;16(6):535–546. [PubMed: 14590152]

Matson M, Mahone EM, Zabel TA. Serial neuropsychological assessment and evidence of shunt malfunction in spina bifida: A longitudinal case study. Child Neuropsychology 2005;11(4):315–332. [PubMed: 16051561]

Molteni E, Bianchi AM, Butti M, Reni G, Zucca C. Combined behavioral and EEG power analysis in DAI improve accuracy in the assessment of sustained attention deficit. Annals of Biomedical Engineering 2008;36(7):1216–1227. [PubMed: 18452058]

Ogg RJ, Zou P, Allen DN, Hutchins SB, Dutkiewicz RM, Mulhern RK. Neural correlates of a clinical continuous performance test. Magnetic Resonance Imaging 2008;26(4):504–512. [PubMed: 18068933]

Olssan EM, von Scheele B, Panossian AG. A randomised, double-blind, placebo-controlled, parallel-group study of the standardised extract SHR-5 of the Roots of Rhodiola rosea in the treatment of subjects with stress-related fatigue. Planta Medica 2009;75:102–112.

Posey DJ, Wiegand RE, Wilkerson J, Maynard M, Stigler KA, McDougle CJ. Open-label atomoxetine for attention-deficit/hyperactivity disorder symptoms associated with high-functioning pervasive developmental disorders. Journal of Child and Adolescent Psychopharmacology 2006;16(5):599–610. [PubMed: 17069548]

Reich, W.; Welner, Z.; Herjanic, B. The Diagnostic Interview for Children and Adolescents – IV. North Tonawanda: Multi-Health Systems; 1997.

Riccio, CA.; Reynolds, CR.; Lowe, PA. Clinical applications of continuous performance tests: Measuring attention and impulsive responding in children and adults. New York: John Wiley & Sons, Inc.; 2001.

Schachar R, Ickowicz A, Crosbie J, Donnelly GA, Reiz JL, Miceli PC, et al. Cognitive and behavioral effects of multilayer-release methylphenidate in the treatment of children with attention-deficit/hyperactivity disorder. Journal of Child and Adolescent Psychopharmacology 2008;18(1):11–24. [PubMed: 18294084]

Strauss, E.; Sherman, E.; Spreen, O. A compendium of neuropsychological tests: Administration, norms, and commentary. 3rd ed. New York: Oxford University Press; 2006.

Wechsler, D. Wechsler Intelligence Scale for Children – Fourth Edition. Minneapolis, MN: Pearson, Inc.; 2003.

Table 1

Study sample characteristics

	Assessment Time 1		
Variable	**Frequency**	**%**	
Male	19	48.7	
Female	20	51.3	
	Mean	*SD*	Range
Demographics			
Age	12.1	3.70	6.3–18.4
Hollingshead Index	46.45	12.79	16–66
CPRS ADHD Index	45.03	4.28	40–56
CBCL Anxious/Depressed T-score	51.91	3.65	50–65
CBCL Attention Problems T-score	51.63	2.32	50–57
Estimated VIQ (PPVT-III)	102.62	16.82	72–150

SD = standard deviation, CBCL = Child Behavior Checklist; CPRS = Conners’ Parent Rating Scale; PPVT-III = Peabody Picture Vocabulary Test – Third edition; VIQ = Verbal IQ.

Table 2

Number of participants with clinically elevated CPT-II scores

	Assessment Time 1	Assessment Time 2
ADHD Confidence Index*	12 (32.4%)	15 (40.5%)
CPT-II Variables**		
• Omissions	3 (8.1%)	9 (24.3%)
• Commissions	2 (5.4%)	7 (18.9%)
• Variability	8 (21.6%)	11 (29.7%)
• Hit Reaction Time	8 (21.6%)	8 (21.6%)
• Attentiveness	1 (2.7%)	8 (21.6%)

*Confidence Index % > 50.

**T-scores > 60. CPT-II = Continuous Performance Test – Second edition.

Table 3

Neuropsychological test performance

	Assessment Time 1			Assessment Time 2				
	$Mean_1$	**SD_1**	**$Range_1$**	**$Mean_2$**	**SD_2**	**$Range_2$**	**$Mean_1$–$Mean_2$ Diff.**	**Test–retest Reliability (ICC)**
Age (years)	12.0	3.7	6–18	12.5	3.7	6.7–19	.5	n/a
Comparison tests								
PPVT-III (*n* = 39)	102.6	16.8	72–150	104.7	18.0	70–154	2.1	.92
WISC-IV: Block Design (*n* = 31)	9.4	3.3	4–17	9.8	3.8	3–17	.4	.87
Beery Test of Visual Perceptual (*n* = 39)	100.3	20.6	59–139	99.0	18.2	58–145	−1.3	.82
Conners CPT-II (*T-scores*) (*n* = 37)								
Errors of Omission	51.8	10.8	41–93	53.0	10.6	41–74	1.2	.39
Errors of Commission	46.7	10.8	11–62	48.3	9.8	32–65	1.6	.57
Variability	53.7	9.1	37–74	53.0	12.0	29–81	−.7	.48
Hit Reaction Time	52.3	12.1	33–85	51.4	12.3	33–77	−.9	.65
Attentiveness	48.9	9.8	16–66	48.3	11.9	12–69	−.6	.33

SD = standard deviation, CPT-II = Continuous Performance Test – Second edition; PPVT-III = Peabody Picture Vocabulary Test – Third edition; WISC-IV = Wechsler Intelligence Scale for Children – Fourth edition; ICC = Intraclass correlation.

Table 4

Changes in CPT-II raw and standardized (T) scores

		Statistically rare *RAW* score changes	**Statistically rare *T-score* changes**
Conners' CPT-II T-Scores	**Expected % (# of participants)**	**Observed % (# of participants)**	**Observed % (# of participants)**
Omissions	10 (3.7)	59.4 (22***)	29.7 (11***)
Commissions	10 (3.7)	0 (0)	16.2 (6)
Variability	10 (3.7)	94.6 (35***)	16.2 (6)
Hit Reaction Time	10 (3.7)	18.9 (7)	16.2 (6)
Attentiveness	10 (3.7)	62.1 (23***)	29.7 (11***)

n = 37. Table depicts the expected versus observed proportion of statistically rare changes in CPT-II raw scores and T-scores occurring between Times 1 and 2. For each variable it was expected that 10 of participants (*n* = 3.7) would generate CPT-II score discrepancies between Times 1 and 2 that exceeded the RCI-derived confidence interval (90). The two columns on the right depict the observed proportions of participants with statistically rare changes in raw score and T-score.

$^{***}\chi^2 = p <.001$. CPT = Continuous Performance Test, Attentiveness = Detectability.