Journal of Clinical and Experimental Neuropsychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ncen20

What can we learn about performance validity from TOVA response profiles?

Beth Pollock, Allyson G. Harrison & Irene T. Armstrong

What can we learn about performance validity from TOVA response pro□les?

Beth Pollock, Allyson G. Harrison and Irene T.

[illegible]

validity of most AD/HD symptom self-report measures, their lack of validity subscales (Quinn, 2003), and the fact that the symptoms associated with AD/HD are well known (Conti, 2004). Base rates for feigned or exaggerated AD/HD symptoms in the post-secondary setting are estimated to range from 14.6 to 47.6% (Harrison & Edwards, 2010; J. Suhr et al., 2008; Sullivan et al., 2007). Individuals may be motivated to be diagnosed with AD/HD by external incentives, such as access to academic accommodations and stimulant medication (Garnier-Dykstra et al., 2012; Jasinski & Ranseen, 2011; Young & Gross, 2011). There are strong societal interests in preventing false-positive diagnoses of AD/HD including: the substantial costs of unnecessary assessments and treatments; unjustified use of limited medical resources; unjustified access to disability-related funding or grants; passive support of drug use; disadvantages for individuals who do not feign AD/HD (and thus are not granted access to academic accommodations); and damage to public confidence in clinical diagnostic practices (Frazier et al., 2008; Harrison, 2006; Harrison et al., 2007, 2012; J. Suhr et al., 2008; Sullivan et al., 2007; Tucha et al., 2014).

Thus, researchers have searched for assessment strategies and/or measures to distinguish between feigned and genuine AD/HD. Research has demonstrated that patients suspected of feigning AD/HD could easily be diagnosed with AD/HD based on their responses to several important clinical interview questions (P. S. Marshall et al., 2016), responses on self-report questionnaires of AD/HD (Fisher & Watkins, 2008; Harrison et al., 2007; Jasinski et al., 2011; Quinn, 2003; P. S. Marshall et al., 2016; Sollman et al., 2010; Tucha et al., 2009; Williamson et al., 2014), self-reported impairments (Bryant et al., 2018; Fuermaier et al., 2018), or performance on commonly used tests of executive functioning, processing speed, and/or academic achievement (Booksh et al., 2010; Frazier et al., 2008; Harrison et al., 2007; Musso & Gouvier, 2012; Sollman et al., 2010).

In an attempt to obtain more reliable, objective, and standardized evidence of AD/HD symptomology, continuous performance tests (CPTs) are often used by clinicians to aid with the diagnosis of AD/HD. Studies have found that adults with AD/HD perform differently on CPTs when compared to controls (e.g., Advokat et al., 2007; Fasmer et al., 2016), though findings have been mixed (see Baggio et al., 2020). However, research has also demonstrated that post-secondary students can easily feign symptoms of AD/HD on these tests. Specifically, J. A. Suhr et al. (2010) demonstrated that students who responded in a non-credible manner on

from those diagnosed with AD/HD in terms of general CPT performance, but performed worse than the AD/HD group on reaction time (RT) variability and RT change. More recently, Nicholls et al. (2020) reported that children who failed a performance validity test (PVT) committed significantly more errors of omission on the TOVA than those who passed a PVT. Harrison and Armstrong (2020

a participant's group assignment depends on the investigator's application of an independent criterion, misclassifications and misestimates of accuracy are always possible (Mossman et al., 2012).

The current study takes M66sificati OcYKcFKc&kMkMk8NMis

Checklist-90 Revised: SLC-90-R- Global Severity Index, 1994) were compared.

Test of Variables of Attention

The TOVA is a computer-based continuous performance test, measuring attention and impulse control (Leark et al., 2008). Measures of performance on the TOVA include the following: (1) errors of omission: failure to respond to the target,

Concept (5 items), Social (9 items), and Risk (14 items). Each item employs a four-point Likert scale scored from 0 to 3 (0 = *never, not at all*; 1 = *sometimes, somewhat*; 2 = *often, much*; 3 = *very often, very much*). A scale score per domain is calculated by summing the response values to all items per domain and dividing this sum by the number of endorsed items. The WFIRS has been found to have excellent internal consistency for the total score (.96), and good to excellent (.85 to .94). for the subscales (Canu et al., 2020). The Canadian AD/HD Resource Alliance (CADDRA) recommends that clinicians consider any domain with a mean score > 1.5 as suggesting impairment (CADDRA, 2011).

The Symptom Checklist-90-Revised

The Symptom Checklist-90-Revised (SCL-90-R; Derogatis, 1994) is a 90-item multidimensional questionnaire developed to screen for a range of psychological symptoms and psychopathological features. In psychotherapy outcome research, the SCL-90-R has been best conceptualized as a one-dimensional measure of symptom load and the Global Severity Index (GSI) has become a widely used measurement of psychological distress (e.g., Osterberg et al., 2002; Skydsbjerg et al., 2001).

Procedure

Participants underwent a one-hour semi-structured interview, were asked to provide report cards from childhood, and supplied rating scales completed by collateral informants to provide information regarding past and present symptoms of AD/HD. During the session, participants also completed the TOVA and MSVT, as well as the self-report version of the

agreement and profile similarity, respectively. Significant ICCs were found for both solutions ($p < .001$). Kappa values were also significant ($p < .001$), with substantial agreement (based on a Kappa interpretive system suggested by Landis & Koch, 1977). Therefore, both solutions were deemed reasonably stable and were retained for multiple-method reliability assessment.

Multiple-method reliability

The prospective solutions were then subjected to three additional hierarchical clustering algorithms (Complete Linkage, Average Linkage Between Groups, and Average Linkage Within Groups), followed by a K-means pass through the data. The level of agreement between cluster solutions generated using the various methods was then calculated. Kappa values for both solutions were significant ($p < .001$) and suggested Fair to Almost Perfect agreement in the three-cluster solution and Moderate to Almost Perfect for the five-cluster solution. All ICCs in the three- and five- cluster solutions were significant ($p < .001$). Thus, both the three- and five- cluster solutions were deemed adequately replicable and were subjected to split-half reliability analyses.

Split-half reliability

To determine the extent to which the derived cluster solutions could be replicated in di□erent samples, the initial sample was randomly split in half and each subsample was subjected to a two-stage Ward's analysis specifying the number of clusters to be recovered. The split-half profiles associated with the three-cluster solutions had good visual agreement, and all ICCs were significant ($p < .01$). Conversely, although the ICCs were all significant ($p < .01$), the split-half profiles from the five-cluster solutions were di□cult to match. Based on these findings, the three-cluster solution was considered representative of the data and was selected as the final cluster solution.

Description of clusters

The three clusters generated on the basis of the initial two-stage Ward's analysis were assigned descriptive labels reflecting the most salient features of each mean TOVA profile. Mean TOVA Index scores for each cluster are presented in Table 2. There were significant di□erences in gender distribution, $\chi^2_{(2)} = 6.87$, $p = .032$, and clinical classification of cases, $\chi^2_{(10)}$ [illegible]

normal scores on the scales assessing for commission errors, a borderline normal score on the first half and a normal score on the second half of the scale assessing for RT, and scores not within normative limits on measures of RT variability and omissions. Their ACS score was similar to individuals with AD/HD (−3.99) but not as profoundly low as the Low group. Their mean SEI score was low (0.67), with 18% of students in this cluster attaining an SEI score of 2 or above. Of those classified to this cluster, 18.1% were ultimately diagnosed with AD/HD and 1 student admitted to malingering. This group was labeled Mixed TOVA

from the CAARS: Self-Report, WFIRS impairment mean scores, and the SLC-90 Global Severity Index. Analysis of variance (ANOVA) and chi-square tests were conducted to determine if the derived subgroups differed on these external variables. In response to significant ANOVA findings, subsequent post hoc comparisons (Games-Howell procedure) were conducted to

a significant difference among the groups on the CII, $\chi^2_{(2)} = 7.97$, $p = .019$, with similar rates of failure (both at 12%) in the High and Mixed groups, but a significantly higher percentage of individuals attaining a score greater than 21 in the Low group (29%). There was no significant difference in failure rates among the groups on the EI, $\chi^2_{(2)} = 3.20$, $p = .202$, with 27% in the High group, 42% in the Low group, and 29% of the Mixed group attaining scores of 3 or above.

Mean impairment scores from the WFIRS were compared across the groups. There was a significant difference

group. They did not di□er from the other two groups in terms of reported level of psychological distress.

Conclusions

In agreement with Robinson and Rogers (2018), the results suggest that good faith assumptions that all AD/HD referrals will put forth their best e□ort appear unwarranted. By using cluster analysis to identify patterns of performance on the TOVA, a group of participants emerged (about one-quarter of the sample) with globally low TOVA performance, high levels of AD/HD

sample was limited, as it included only post-secondary students with attentional concerns. In addition, although using retrospectively gathered data enables researchers to carry out studies that may not be possible otherwise, such investigations are constrained by available data. Another limitation relates to changes in score reporting in different versions of the TOVA. Specifically, index scores below 40 were given as 0 in older versions of the test and provided as < 40 (which were represented as 40 in our database) in the newer versions of the test. As the majority of individuals attaining these extreme scores fell in the Low cluster group, mean Index scores on the TOVA in this cluster were likely affected by the changing representation of the lowest score that could be attained. Other limitations of the present investigation relate to the use of cluster analytic methodology. Despite attempts to ensure the reliability and validity of the derived typology, the fact remains that cluster analysis represents a relatively subjective research tool (Lange et al., 2002). Although efforts were made to ensure that selections regarding the similarity coefficient, grouping algorithm and association indexes followed conventional standards and were empirically derived, in the end, a somewhat subjective decision is required by the researcher to determine the metrics to be used. Additionally, with the use of cluster analysis, all participants in a sample are forced into clusters on the basis of relative similarity to other participants without consideration of similarity in an absolute sense (Hair & Black, 2000). Thus, the clusters generated in this investigation likely include some individuals who bear only a minimal similarity to the mean profile derived for that cluster. Furthermore, although Squared Euclidean Distance, the measure of similarity used in the current investigation, is the most commonly used similarity index in taxonomic research, it has been argued that the methodology that maximizes the influence of profile shape and minimizes the influence of profile magnitude may derive clusters that provide more meaningful information (Lange, 2007). Finally, from a clinical standpoint, the final decision about diagnosis of AD/HD was based on the clinical judgment of specific clinicians within the clinic along with results from measures included in the analyses. Future studies may wish to consider using methods to obtain an independent judgment regarding diagnosis. Considering that this investigation represents the first empirical attempt to delineate patterns of performance using the TOVA, it is necessary to through replication and cross-validation. Nevertheless, this study is the first to demonstrate that cluster analysis may be a useful alternative to simulation and known-groups approaches when conducting research on performance measures.

In conclusion, our investigation confirms that cluster analysis can identify reliable and clinically meaningful groups of young adults seeking initial assessment for possible AD/HD. Three profiles emerged, including one cluster who demonstrated exceptionally low performance on the TOVA and exceptionally high reporting of AD/HD symptomology. The implication from our analysis is that this group likely represents individuals who were exaggerating or magnifying their difficulties to obtain an AD/HD diagn VA,

Baggio, S., Hasler, R., Giacomini, V., El-Masri, H., Weibel, S., Perroud, N., & Deiber, M.-P. (2020). Does the Continuous Performance Test predict

Harrison, A. G. (2006). Adults faking AD/HD: You must be