OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 36 (2021) 940–953

Archives of CLINICAL NEUROPSYCHOLOGY

Internal (Factorial) Validity of the ANAM using a Cohort of Woman High-School Soccer Players

Joseph J. Glutting[1], Adam Davey[2], Victoria E. Wahlquist[3], Marley Watkins[4], Thomas W. Kaminski[3,*]

[1] *School of Education, University of Delaware, Newark, DE 19716, USA*
[2] *Department of Behavioral Health and Nutrition, University of Delaware, Newark, DE 19716, USA*
[3] *Department of Kinesiology and Applied Physiology - Athletic Training Research Lab, University of Delaware, Newark, DE 19716, USA*
[4] *Baylor University, Department of Educational Psychology, University of Delaware, Newark, DE 19716, USA*

*Corresponding author at: Department of Kinesiology and Applied Physiology, 541 S. College Ave, Newark, DE 19716, USA. ORCID#-0000-0003-0924-0978; Twitter=@UD_ATEP. *E-mail address:* kaminski@udel.edu (Thomas W. Kaminski, PhD, ATC)

Received 20 August 2020; revised 21 October 2020; Accepted 13 November 2020

Abstract

Introduction: Computerized neuropsychological testing is a cornerstone of sport-related concussion assessment. Female soccer players are at an increased risk for concussion as well as exposures to repetitive head impacts from heading a soccer ball. Our primary aim was to examine factorial validity of the Automated Neuropsychological Assessment Metrics (ANAM) neuropsychological test battery in computing the multiple neurocognitive constructs it purports to measure in a large cohort of interscholastic female soccer players.
Methods: Study participants included 218 interscholastic female soccer players (age = 17.0±0.7 year; mass = 55.5±6.8 kg; height = 164.7±6.6 cm) drawn from a large (850+) prospective database examining purposeful heading from four area high schools over a 10-year period. The ANAM-2001 measured neurocognitive performance. Three methods were used to identify integral constructs underlying the ANAM: (a) exploratory factor analysis (EFA), (b) first-order confirmatory factor analysis (CFA), and (c) hierarchical CFA.
Results: Neuropsychological phenomena measured by the ANAM-2001 were best reproduced by a hierarchical CFA organization, composed of two lower level factors (*Simple Reaction Time*, *Mental Efficiency*) and a single, general composite. Although the ANAM was multidimensional, only the composite was found to possess sufficient construct dimensionality and reliability for clinical score interpretation. Findings failed to uphold suppositions that the ANAM measures seven distinct constructs, or that any of its seven tests provide unique information independent of other constructs, or the composite, to support individual interpretation.
Conclusions: Outcomes infer the ANAM possesses factorial-validity evidence, but only scores from the composite appear to sufficiently internally valid, and reliable, to support applied use by practitioners.

Keywords: Statistical methods; Test construction; Cognitive enhancement; Executive functions; Head injury/traumatic brain injury; Learning and memory

Introduction

Validity is fundamental to the clinical interpretation of test scores. The reason is because validity makes clear the meaning that can be attached to those scores (Cohen, Swerdlik, & Sturman, 2017; Cronbach & Meehl, 1955; Gregory, 2014; Messick, 1989; Reynolds, Livingston, & Willson, 2009; Salvia, Ysseldyke, & Whitmer, 2017). The *Standards for Educational and Psychological Testing* identify five types of validity evidence that must be considered when evaluating test scores: (a) content indicators, (b) support grounded in response processes, (c) intended/unintended consequences, (d) criterion-related outcomes, and (e) a test's

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permission@oup.com.
https://doi.org/10.1093/arclin/acaa120 Advance Access publication 29 December 2020

internal structure (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Evidence about a test's factor structure is vital to clinicians because factor analysis serves as the internal rationale for determining which scores practitioners should, or should not, interpret (Braden, 2013; Braden & Niebling, 2012; Canivez & Youngstrom, 2019; Furr, 2011). Essentially, factor analysis determines whether a test's actual structure matches its theoretical framework. Structural validity evidence is especially important for tests that are used for high-stakes decisions. For example, the Automated Neuropsychological Assessment Metrics (ANAM), a computer-based battery of 30 cognitive tests, has been extensively applied in military, athletic, and medical contexts for important diagnostic, eligibility, and treatment decisions (Kaminski, Groff, & Glutting, 2009; Reeves, Winter, Bleiberg, & Kane, 2007; Rice et al., 2011). The ANAM offers three scores for each test: speed, accuracy, and throughput (accuracy/speed) as well as a composite score for interpretation.

The ANAM's relationship to other traditional neuropsychological tests (Trail Making, PASAT, HVLT, Stroop Color-Word, COWAT, etc.) used in sport-related concussion assessment has been previously examined and confirmed by several author groups (Bleibert, Kane, Reeves, Garmoe, & Halpern, 2000; Lovell & Collins, 1988; Schretlen, 1997; Woodard et al., 2002). Thus, these scores can assist clinicians in gaining an understanding of the amount of insult to brain processes involving higher order thinking, reaction time, and executive functioning.

A search for validity evidence to support the ANAM scoring structure was conducted across electronic databases including: Medline, PsycINFO, and ERIC using combinations of the paired terms "a" and "b" and their acronyms: (a) ANAM, (b) factor analysis, (b) principal component analysis (PCA), (b) exploratory factor analysis (EFA), (b) confirmatory factor analysis (CFA), (b) construct validity, (b) factorial validity, and (b) internal structure. Primary sources were read and searched for other articles.

We located only eight studies, which explored the ANAM's structure. The finding is surprising given the ANAM's widespread use and longevity—it was first published nearly 30 years ago (Reeves et al., 2007) and has been administered to more than 100,000 examinees (Vincent, Roebuck-Spencer, Gilliland, & Schlegel, 2012). Two investigations were excluded because they were conference/poster presentations and never appeared in a juried journal (Retzlaff & Vanderploeg, 1999; Roskos, Feller, & Chibnall, 2014). One peer-reviewed publication was rejected because its analyses were limited to item scores from the ANAM's Mood Scale (Johnson, Vincent, Johnson, Gilliland, & Schlege, 2008). This left five articles: (a) Bleibert et al. (2000), (b) Jones, Loe, Krach, Rager, and Jones (2008), (c) Kabat, Kane, Jefferson, and DiPino (2001), (d) Naifeh et al. (2016), and (e) Short, Cernich, Wilken, and Kane (2007).

Sample characteristics were highly diverse and included: high school and college students, undergraduates (attending an urban, southwestern university), outpatients seen for neuropsychological evaluations at two VA medical centers, a large cohort of military personnel, and patients with multiple sclerosis and their normal controls. Evidence regarding factorial validity was similarly diverse; the number of accepted constructs varied considerably, with several studies containing multiple analyses. Results produced a one-factor EFA solution using throughput scores (Naifeh et al., 2016), a single PCA construct when either response time or throughput scores functioned as inputs (Kabat et al., 2001), two PCA constructs when accuracy scores were utilized with the second dimension representing a singlet (a single score loading on a factor; Jones et al., 2008), three PCA constructs when accuracy scores were entered (Jones et al., 2008; Kabat et al., 2001), three PCA constructs using accuracy and response time scores combined, with the third factor being a singlet (Bleibert et al., 2000), and five CFA dimensions when one or more ANAM accuracy or response-time scores were paired with one or more variables from a traditional neuropsychological measure (Short et al., 2007).

Exploratory methods dominated previous research with the ANAM. PCA/EFA was used in four out-of-the five papers. One of the most critical decisions in a PCA/EFA is to define the correct number of factors to retain (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Preacher, Zhang, Kim, & Mels, 2013). None of the foregoing ANAM studies used state-of-the-art procedures such as parallel analysis (PA) or minimum average partialling (MAP) to identify the correct number of factors to retain during a PCA/EFA. No CFA used scores from the ANAM alone nor is CFA likely to identify the correct number of factors (Garrido, Abad, & Ponsoda, 2016; Heeme, Hilbert, Draxler, Ziegler, & Bühner, 2011). Consequently, it is reasonable to conclude that results from factorial-validity research were equivocal. One, two, three, and five dimensions were suggested to underlie the ANAM. Thus, to date, no firm foundation has been established as an empirical base for CFA.

The current investigation was developed to address the shortcomings in prior factor analyses. We used three methods to identify integral constructs underlying the ANAM: (a) EFA, (b) first-order CFA, and (c) hierarchical CFA. As indicated, past research identified multiple factor models for the ANAM. Rather than rely on such diverse outcomes, the number and nature of factors for the CFAs were based on more modern empirical results from the current EFAs—along with concordant finding from the previous studies. Consequently, the current investigation followed methodological recommendations that "the purpose of EFA is typically to identify the latent constructs or to generate hypotheses about their possible structures [the nature and number of factors], whereas the purpose of CFA is to evaluate hypothesized structures of the latent constructs and/or to develop a better understanding of such structures" (Bandalos & Finney, 2019, p. 98). That is, the EFAs were used to generate models and the subsequent CFAs were appropriately used to test models (Matsunaga, 2010; Norman & Streiner, 2014). Previous research, along with recommendations by recent textbooks, also indicate that CFA is especially effective in uncovering latent factor structures when used in conjunction with preliminary EFA findings (DeVellis, 2017; Fabrigar et al., 1999; Gerbing & Hamilton, 1996; Selbom & Tellegen, 2019).

The scores clinicians interpret should not be based on factorial-validity evidence alone. Valid interpretation also is dependent upon a test's external validity, and on how well each score reliably reflects its intended construct. With respect to reliability, a clinically meaningful score is defined as "one that is reliable enough for its prospective use and one that has information that is not adequately contained in [just a] total test score" (Wainer & Feinberg, 2015, p. 18). Thus, dimensions uncovered during the CFAs were further evaluated to determine which ANAM scores possess sufficient reliability for clinical score interpretation.

Materials and Methods

Participants

Study participants included 218 interscholastic female soccer players (age = 17.0±0.7 year.; mass = 55.5±6.8 kg.; height = 164.7±6.6 cm) drawn from a large (850+) prospective database examining purposeful heading from four area high schools over a 10-year period. To be considered for inclusion, participants had to have completed at least three full playing seasons. All participants (and their parents if necessary) signed the informed consent agreement, which was approved by the University's human subjects review board (HS IRB 157873–5). Each individual completed a General Health Questionnaire, which included demographic data, concussion history, player position, and number of years playing soccer at the start of each soccer season.

We adapted the commonly recommended strategy of splitting subjects into two samples and performing EFAs on one sample and CFAs on the other (*cf*. Glutting, Youngstrom, Watkins, & Frazier, 2007; Williams et al., 2003). The first sample ($N = 218$) was obtained at the baseline of a large longitudinal study of interscholastic girls' soccer players. Data from wave 1 of this sample were used for EFA. A second set of observations from this sample ($N = 217$) covered the same subjects tested one year later (with the loss of one participant who moved from the area). Data from these observations were used for CFA.

Instruments

Three instruments were utilized in this study. First, a General Health Questionnaire, which included demographic data, concussion history, player position, and number of years playing soccer. Second, a concussion symptom checklist consisting of 17 SRC symptoms (headache, dizziness, blurred vision, poor concentration, fatigued, neck pain, etc.). Each symptom was scored on a spectrum from 0 (none) to 6 (worst ever experienced), and the total symptom score was tallied.

Finally, the ANAM-2001 was utilized to measure neurocognitive performance. Throughput scores measure test performance efficiency, which combine speed and accuracy (number of correct responses per minute) of the particular cognitive domain in question. The higher the throughput score, the more efficient the participant was during a test. Throughput scores served as the dependent variables. The stability of the ANAM test scores was reported previously (Kaminski et al., 2009).

The seven test variables included:

- *Simple Reaction Time (SRT)*—a measure of SRT that measures response time (in ms) to a stimulus (*) presented at various time intervals.
- *Continuous Performance Test (CPT)*—a measure of attention and concentration where the participant must continuously monitor letters and identify whether the current letter displayed is the same or different from the immediately preceding letter.
- *Math Processing (MPH)*—a measure of mental processing speed and mental efficiency where the participant must solve a three-step simple addition or subtraction equation and identify whether the solution is greater or less than 5.
- *Match-to-Sample (MTS)*—a measure of visual memory whereby the participant must recall a checkerboard matrix after 5 s and match it to the original matrix design.
- *Sternberg Memory (ST6)*—a measure of working memory where the participant must memorize a string of six letters and subsequently recall whether or not a presented letter belongs to that six-letter string.
- *SRT2*—a repeat of the SRT test.
- *CPT2*–a repeat of the CPT test.

The software uses a pseudo-randomization procedure to minimize practice effects on repeat testing of SRT2 and CPT2. Therefore, a question arises—Does the ANAM offer users five, or seven, distinct scores? How one answers this question has a direct effect on the number of variables to include in a factor analysis of the ANAM. To be thorough, we factored both five-score and seven-score batteries. Results are presented here for the seven-score battery.

Procedure

All instruments were administered at the beginning and conclusion of each soccer season in a quiet classroom setting, free from distractions. Testing was completed at the end of the school day prior to any soccer activities. The ANAM-2001 protocol was administered on a laptop computer with participants using the handheld mouse to navigate the on-screen prompts. Testing required approximately 30 min to complete and was coordinated through the efforts of the schools' athletic trainer or coach. None of the participants reported any adverse effects from the baseline testing session.

Data Analyses

EFA. Analyses began with a series of EFAs because performing a CFA on the same sample as an EFA results in overfitted models (Flora & Flake, 2017; Osborne & Fitzpatrick, 2012). Indeed, Fokkema and Greiff (2017) randomly generated 25 uncorrelated item scores using 300 observations. They conducted EFA and then CFA on the same random data. The CFA produced deceivingly optimistic model fit indices and parameter estimates. Based on these results, Fokkema and Greiff (2017) strongly recommended against using EFA and CFA with the same sample.

Empirical work has shown that CFA may be a less desirable technique for determining the number of latent dimensions measured by an instrument (Garrido et al., 2016; Heeme et al., 2011). For instance, MacCallum and colleagues found that specification searches in covariance structure modeling often did not uncover the correct population model (MacCallum, 1986; MacCallum, Roznowski, & Nowrwitz, 1992). Likewise, Gorsuch (2003) reported that whereas EFA results nearly always replicate during first-order CFAs; the reverse is not true when CFA is employed to uncover first-order factors and then used to replicate results with a second sample. Therefore, EFAs were employed first because of the uncertainty surrounding the underlying structure of the ANAM (Browne, 2001) and the potential for stronger structural evidence to emerge during second-stage CFAs (Goldberg & Velicer, 2006).

EFAs were conducted with SPSS version 27 (IBM, 2020). Principal axis extraction was applied due to its relative tolerance of multivariate nonnormality and its superior recovery of weak factors (Briggs & MacCallum, 2003). Communalities were initially estimated by squared multiple correlations and were iterated to produce final communalities (Gorsuch, 1983). For both theoretical and empirical reasons, it was assumed retained factors would be correlated (Gorsuch, 1983; Meehl, 1990). Consequently, promax rotations were employed.

Defining the number of factors. As mentioned in the introduction, one of the most critical decisions in a PCA/EFA is to determine the correct number of factors to retain (Fabrigar et al., 1999; Preacher et al., 2013). Kaiser's criterion (eigenvalues $\geq$ 1.0) is the most popular method (Fabrigar et al., 1999; Ruscio & Roche, 2012), and it is the default in most statistical packages (Henson & Roberts, 2006). One of the prior ANAM analyses used this procedure (Bleibert et al., 2000). Two ANAM studies used both eigenvalues $\geq$ 1.0 and scree plots (Kabat et al., 2001; Naifeh et al., 2016). The last publication was not explicit about its method(s) for keeping factors (Jones et al., 2008).

Unfortunately, Kaiser's rule performed poorly in Monte Carlo simulations. It tended to under- or overestimate the number of true constructs (Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). Similarly, scree plots were criticized as being too subjective (Crawford & Koopman, 1979; Streiner, 1998). Attempts were made to interpret scree plots more objectively via empirical methods (see Watkins, in press, for a review). Even so, results from simulations were inconsistent and it is not clear if any of the empirical scree methods was superior (Nasser, Benson, & Wisenbaker, 2002; Raiche, Walls, Magis, Riopel, & Blais, 2013).

As also presented in the introduction, none of the foregoing ANAM studies used state-of-the-art procedures such as parallel analysis (PA) or minimum average partialling (MAP). Findings from several Monte Carlo simulations demonstrated PA and MAP were the two best methods for determining the correct number of factors and that scree was a useful adjunct (Garrido, Abad, & Ponsoda, 2011; Ruscio & Roche, 2012; Timmerman & Lorenzo-Seva, 2011; Zwick & Velicer, 1986). Model-fit criteria (χ^2, the root mean square error of approximation [RMSEA], etc.) were shown to be less accurate than either PA or MAP and, they were more likely to over-factor (select too many factors; MacCallum et al., 1992; Schmitt, 2011).

Each EFA model was assessed against the following five rules: (a) eigenvalues $\geq$ 1.0 (Kaiser, 1974a), (b) scree (Cattell, 1966), (c) mean eigenvalues from a PA (Horn, 1965), (d) Glorfeld's (1995) 95th percentile extension of PA, and (e) interpretability (Fabrigar et al., 1999; Gorsuch, 1983, 2003). In addition, two measures of model fit were examined (f) the chi-square (χ^2) test, and (g) the RMSEA (Steiger, 1990).

Factor loadings (i.e., pattern coefficients) ≥.40 were considered salient because that would explain at least 16% of a variable's variance. The overall goal was to achieve interpretable and theoretically meaningful factors that optimally balanced comprehensiveness (accounting for the most variance) and parsimony (with the fewest factors) in an approximate simple structure pattern (Fabrigar et al., 1999; Gorsuch, 1983, 2003).

CFA. Mplus version 8.3 (Muthén and Muthén (1998-2018) was used for the CFAs. As presented in the Results section, certain variables were kurtotic. Therefore, the robust Huber/Pseudo ML/sandwich correction (Mplus: estimator = MLR) was applied because of its ability to uncover underlying factors when data are not normally distributed.

Numerous fit indices are available for estimating the quality of measurement models. Most were developed under somewhat different theoretical frameworks and/or focus on different aspects of fit (*cf*. Browne & Cudeck, 1993; Hu & Bentler, 1995). For this reason, it is generally recommended that multiple measures be employed (Tanaka, 1993). The Tucker-Lewis index (TLI; Tucker & Lewis, 1973), the comparative fit index (CFI; Bentler, 1990), the RMSEA (Browne & Cudeck, 1993), the standardized root mean square residual (SRMR; Chen, 2007), and the parsimonious normed fit index (PNFI; Mulaik et al., 1989) were reported for each model. The first two measures generally range between .00 and 1.0, with larger values reflecting better fit. Historically, values of .90 or greater have been taken as evidence of good-fitting models (Bentler & Bonett, 1980). However, more recent research suggests this threshold should be closer to .95 for the CFI and TLI (Hu & Bentler, 1995; Marsh, Hau, & Grayson, 2005). Alternatively, smaller RMSEA and SRMR values support better fitting models, with values of .05 or less indicating good fit (Browne & Cudeck, 1993). The PNFI ranges between .00 and 1.00, with higher values indicating a more parsimonious fit. Unlike the other measures of fit, there is no standard regarding how high the PNFI should be. It is used to compare competing models (Trost et al., 2003).

Likewise, three information criteria were calculated for the CFAs: (a) the Akaike Information Criteria (AIC; Akaike, 1987), (b) the Bayesian Information Criteria (BIC; Schwartz, 1978), and (c) the sample-size adjusted BIC (SS-BIC; Sclove, 1987). Model selection is relative for the information criteria. The analysis with the lowest AIC, BIC, and/or SS-BIC value is the best model.

Factor invariance. Factorial invariance has emerged as an important consideration in structural equation modeling across groups and/or occasions (Meredith, 1964; Morin, Arens, & Marsh, 2016; Vandenberg & Lance, 2000). Therefore, invariance was examined across the two timepoints the ANAM was administered.

Reliability. Cronbach's (1951) alpha is the most widely used measure of internal-consistency reliability (Trizano-Hermosilla & Alvarado, 2016). Across the last decade or so, shortcomings of the alpha coefficient have been widely discussed (Cho & Kim, 2015; Revelle & Zinbarg, 2009; Sijtsma, 2009, 2011; Sijtsma & van der Ark, 2015). Measurement experts now recommend omega coefficients when gauging the internal consistency of factor-based models (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Gignac & Watkins, 2013; Schweizer, 2011; Watkins, 2017). This is especially important for multidimensional instruments, which require an understanding of an overall test-score (composite) reliability, as well as reliability for each of the lower (first-order factor) scores underlying the composite (Canivez, 2016; Canivez, Watkins, & Dombrowski, 2017; Watkins, 2017).

The reliability of factors and factor scores can be estimated with both omega coefficients (McDonald, 1999; Rodriguez, Reise, & Haviland, 2016a, 2016b; Watkins, 2017) and the *H* index (Hancock & Mueller, 2001). More specifically, McDonald (1999) developed a family of omega coefficients useful for identifying reliabilities from a factor analysis (either EFA or CFA). We employed three omega coefficients. First, omega total (ω_{T}) produces reliability estimates similar to coefficient alpha when a test is unidimensional, or when a test is composed of only first-order factors (Brunner, Nagy, & Wilhelm, 2012; Maydeu-Olivares, Coffman, & Hartmann, 2007; Raykov, 1997). Second, omega hierarchical (ω_{H}) was used to estimate the reliability of a unit-weighted total score (i.e., the higher-order composite) after removing the influence of the group factors. Third, omega scale (ω_{S}) examined the reliability of a unit-weighted first-order (group) scores after removing the influence of all other factors.

A high ω_{H} coefficient indicates the general factor is the dominant source of variance in a test. Conversely, a low ω_{H} coefficient indicates group factors and/or uniqueness account for the majority of reliable variance (Watkins, 2017). It has been suggested ω_{H} values should be greater than .50, and preferred values greater than .75; but these cutoff values have not been thoroughly investigated (Reise, 2012; Reise, Bonifay, & Haviland, 2013). ω_{T} estimates were obtained using Mplus. The ω_{H} and ω_{S} coefficients were produced using the Omega program developed by Watkins (2013, 2017). Omega ω_{H} can also be seen as a validity measure because it addresses the proportion of variance contributed by latent constructs (Brunner et al., 2012; Gustafsson & Åberg-Bengtsson, 2010).

The reliability of factors was also examined using the *H* index. *H* is the correlation between a factor and an optimally weighted factor score, and it is considered a measure of construct reliability or replicability that quantifies how well a latent variable is represented by a set of indicators (Hancock & Mueller, 2001). According to Mueller and Hancock (2019), *H* is "an estimate of the correlation that a factor is expected to have with itself over repeated administrations" (p. 455). *H* values lower than .80 suggest that the factor is not well defined and will not replicate across studies nor provide accurate path coefficients if included

Table 1. Distributional statistics for ANAM at first and second testing

	First testing[a]				Second testing[b]			
	M	*SD*	Skew	Kurtosis	*M*	*SD*	Skew	Kurtosis
Variable								
SRT—First administration	213.33	31.46	−0.65	1.93	219.91	29.56	−0.08	0.30
MPH	17.75	5.69	0.69	2.16	21.81	8.11	2.60	14.0
CPT—First administration	93.85	17.68	−0.61	2.27	111.00	16.00	0.14	0.49
Match to sample	35.65	11.22	0.73	0.63	39.94	12.78	0.62	0.12
ST6	75.58	16.31	−0.17	1.56	86.64	15.61	0.33	0.26
SRT—Second administration	212.64	32.74	−0.64	1.72	215.61	34.16	0.01	0.97
CPT—Second administration	101.63	17.62	−0.84	4.09	115.28	18.05	0.15	0.01

Note: All values rounded to second decimal position for convenient presentation.
[a]$N = 218$
[b]$N = 217$

in statistical models (Ferrando & Lorenzo-Seva, 2019; Mueller & Hancock, 2019; Rodriguez et al., 2016b). A tutorial on the use of omega and *H* coefficients is provided by Watkins and Canivez (2020).

Results

Table 1 provides means distributional statistics (*M*s, *SD*s, skew, kurtosis) for the seven ANAM variables across the first and second assessments. Several variables were kurtotic and required uses of the PAF method during the EFAs and the MLR estimator during the CFAs. In addition, reaction time and/or CPT values for four subjects were considerably larger than for the other 213 participants. Mahalanobis distances showed each of the four subjects was a multivariate outlier (Tabachnick & Fidell, 2019). We repeated all analyses excluding the four cases. All substantive conclusions were identical. Results are reported for the full sample.

EFA

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser, 1974) was acceptable (.81). Bartlett's test of sphericity (1950) statistically rejected the hypothesis that the correlation matrix was an identity matrix (χ^2 of 587.098 with 21 degrees of freedom at $p < .001$). Altogether, these measures indicate that factor analysis is appropriate.

One through seven factors (i.e., one factor for each ANAM score that practitioners routinely interpret) were evaluated and compared to criteria for determining the number of factors to retain. None of the three- through seven-factor models satisfied even a single criterion for factor retention. PA (using both 50th and 95th percentile eigenvalues) and MAP pointed to retaining one factor. Alternatively, the scree plot suggested a two-factor solution. Consequently, both one- and two-factor solutions were examined.

Table 2 presents loadings for models with one and two factors. For Model 1, all seven ANAM tests showed salient loadings and this factor accounted for 40.4% of the before-rotation variance. This single, generalized dimension was named *"Neurocognition."* Around 42% of the residuals were $> .05$ and 38% were $> .10$, suggesting that another factor might be extracted (Flora, 2018). Given a one-factor structure, McDonald's (1999) ω_{T} was used to estimate internal consistency. The ω_{T} coefficient ($\omega_{\mathrm{T}} = .76$) indicated the one-factor battery possessed acceptable internal-consistency reliability.

Simple structure was also evident for Model 2. The first, and largest factor, was defined by salient loadings from the CPT (first and second administrations), ST6, MTS, and MPH and accounted for 41.5% of the total variance before rotation. Consequently, the factor was labeled *"Mental Efficiency."* The second factor was characterized by salient loadings from the SRT (first and second administrations) and accounted for 10.4% of the variance before rotation. Therefore, it was named *"SRT."* These two factors accounted for 51.9% of the total variance, which is near the median found during a meta-analysis of 803 EFAs (Peterson, 2000). The interfactor correlation was .60 was not high enough to pose a threat to discriminant validity. Reliability for the two-factor model is presented in the CFA section.

CFA

Given the lack of a clear-cut EFA solution, both one- and two-factor structures were submitted to a CFA. The CFAs used data from the second testing. Fit statistics are reported for all tested models (Table 3). The hierarchical CFA model with two

Table 2. Loadings from the one-factor and two-factor EFAs of the ANAM

Variable	Model 1[a]	Model 2[b]	
	Factor I	Factor I	Factor II
SRT—First admin	**.50**	−.11	**.82**
MPH	**.40**	**.52**	−.12
CPT—First admin	**.87**	**.76**	.15
Match to sample	**.41**	**.56**	−.16
ST6	**.69**	**.76**	−.04
SRT—Second admin	**.53**	−.08	**.83**
CPT—Second admin	**.88**	**.73**	.18
Eigenvalue	2.83	2.91	0.73
% Variance	40.4	41.5	10.4

Note. All values rounded to second decimal position for convenient presentation.
Salient coefficients (i.e., $\geq$.40) in bold.
[a] $N = 218$.
[b] $N = 217$.

Table 3. The goodness-of-fit statistics using robust least squares estimators for CFA models

Fit statistic	Model		
	CFA 1-factor first-order model[a,b]	CFA 2-factor first-order model	CFA[b] hierarchical model with two first-order factors
χ^2	49.498	7.281	7.281
df	14	13	13
p	0.0001	0.8871	0.8871
RMSEA	0.108	0.000	0.000
SRMR	0.059	0.020	0.020
CFI	0.920	1.000	1.000
TLI	0.880	1.000	1.000
AIC	12512.158	12459.547	12459.547
BIC	12583.232	12534.006	12534.006
SS-BIC	12516.685	12464.290	12464.290
PNFI	.894	.984	.984

Notes. CFA = confirmatory factor analysis; *df* = degrees of freedom; RMSEA = root mean square error of approximation; SRMR = standardized root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis Index; AIC = Akaike information criterion; CAIC = constant AIC; BIC = Bayesian information criterion; SS-BIC = sample-size-adjusted BIC. PNFI = Parsimonious Normed Fit Index.
$N = 217$.
[a] EFA Promax rotation, factor correlation, $r = .60$.
[b] Correlation among first-order factors = .66.

first-order factors and the CFA model with two first-order factors showed superior fit statistics. Fit statistics were identical for these two models because the second-order factor is just identified. A first-order CFA with two (or three) correlated factors is mathematically equivalent to a hierarchical CFA including the same two (or three) factors and results in the same fit statistics and degrees of freedom (Hershberger & Marcoulides, 2013; Morin et al., 2016).

There are two compelling reasons for favoring the hierarchical model. First, a substantial association was found between the two first-order factors ($r = .60$ between the promax EFA rotated factors, and $r = .66$ among the two first-order CFA factors). These findings supported the presence of a higher order organization. Second, reliabilities (presented below) were much better for the hierarchical composite than that for the two first-order factors.

Figure 1 provides a visual representation of the accepted, hierarchical CFA model. With the exception of the MPH test (highest loading = .28), the other six tests showed appreciable loadings on both their hypothesized lower order factor, as well as the single higher order dimension (appreciable loadings $\geq$ .45). The first lower order factor was named *Mental Efficiency*, and it was organized according to scores from four tests: CPT (first and second administrations), ST6, and MTS. The second lower order factor was labeled *SRT* factor, and it was characterized by salient loadings from the SRT test (first and second administrations).

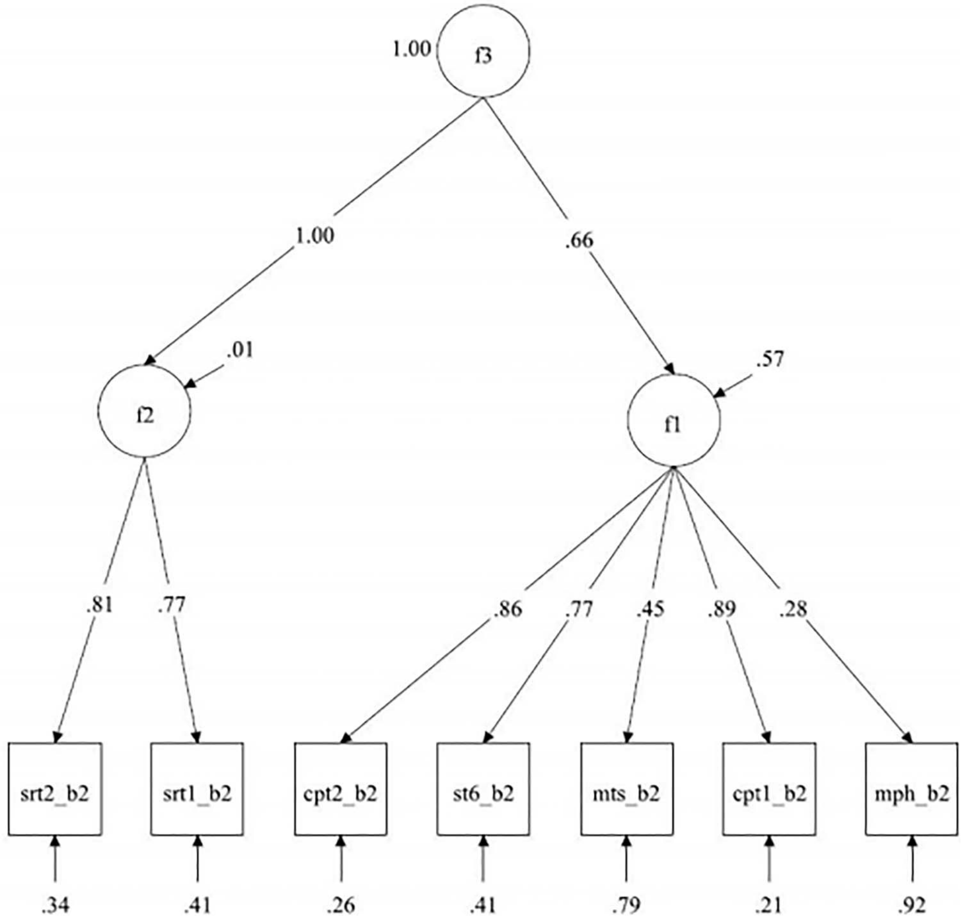

Fig. 1. Hierarchical CFA model.

Factor Invariance

Under most circumstances, factor invariance requires that an established factor structure be in place to establish the level of invariance (e.g., configural, scalar, metric) across groups and/or times (Meredith, 1964; Vandenberg & Lance, 2000). In the current case, where the factorial structure has not yet been established, we sought to compare pertinent elements of the covariance matrix between the first occasion and the second occasion in our data. Because our approach is factor analytic (in contrast to principal component analysis), modeling of covariances is the central consideration. Using procedures recommended in Jöreskog & Sörbom (1996), beginning with a freely estimated covariance matrix (i.e., the saturated model), we constrained the corresponding elements of the ANAM to be equal across occasions, resulting in a 28 degree of freedom test (χ^2 (28) = 71.8, $p < .0001$, CFI = .97, TLI = .89, RMSEA = .088). In order to allow for the possibility that subscale variances could vary across occasions, we eliminated cross-occasion constraints on subscale variances, resulting in a 21 degree of freedom test (χ^2 (21) = 15.5, $p = .796$, CFI = 1.00, TLI = 1.00, RMSEA = .001). Follow-up analyses on these results suggested that the variance of the MPH subscale was greater at the second occasion that of the first, but that there were no other differences in the covariance matrices on the first and second occasions, suggesting that whatever underlying structure applies to the data, it can be considered comparable across timepoints. We used this information in order to construct our approach to assessment of the factor structure of the ANAM using data from the first occasion to determine a structure and data from the second occasion to cross-validate it.

Reliability

Table 4 presents internal-consistency estimates. The ω_T) column supplies internal-consistencies for the two factors from the first-order CFA. Reliability was acceptable for the first factor ($\omega_T = .83$), but the second factor showed insufficient internal consistency for score interpretation ($\omega_T = .54$).

Table 4. Omega and H coefficients for the hierarchical

Score	Omega ω_T for two first-order factors	Omega ωh	Omega ω_S	*H*
Hierarchical score (overall composite)		.85	.73	.86
Factor 1 (SRT)	.83		.20	.41
Factor 2 (mental efficiency)	.54		.03	.01

Notes: Omega ω_T = Omega total, Omega ωh = omega hierarchical; Omega ω_S = omega scale, H = H index.

Omega hierarchical (ω_H) and omega scale (ω_S) were estimated for the hierarchical CFA.

The omega coefficient was high for the overall composite (.85), and it was sufficient for confident score interpretation. By contrast, omegas were much lower for the two lower level factors. Values ranged from a high of .20 for the "*SRT*" factor to a low of .03 for the *"Mental Efficiency"* factor. Both internal-consistency coefficients were below the recommended minimum of .50.

Equally important, the *H* index demonstrated that both of the two first-order ANAM factors were not well defined and were unlikely to replicate in future studies (i.e., $H < .80$). The general factor was well defined and should replicate ($H =$ .86). Therefore, results across both sets of factor analyses (EFAs, CFAs), and the reliability examinations, were consistent in supporting interpretation of the overall composite score from the ANAM.

Discussion

The ANAM offers practitioners numerous test scores and a composite to interpret. But valid interpretation is dependent on whether these scores provide unique information independent of other constructs and on how reliably each score reflects its intended dimension (Brunner et al., 2012; Canivez & Youngstrom, 2019; Chen et al., 2012; Ferrando & Lorenzo-Seva, 2019; Ferrando & Navarro-González, 2018; Reise et al., 2013, 2018; Rodriguez et al., 2016a, 2016b; Wainer & Feinberg, 2015). This study evaluated the internal structure (factorial validity) of the ANAM using a sample of high-school, female soccer players. Findings failed to uphold suppositions that the ANAM measures seven distinct constructs. Thereby, results make clear clinicians cannot validly interpret *any* of the ANAM's seven test scores independently.

Instead, neuropsychological phenomena measured by the ANAM-2001 were best reproduced by a hierarchical CFA organization composed of only three scores: a general composite and two lower level factors (*SRT*, *Mental Efficiency*). The lower level *SRT* factor was characterized by salient loadings from the SRT test (first and second administrations). The lower-level *Mental Efficiency* factor was organized according to scores from four tests: CPT (first and second administrations), ST6, and MTS. The MPH test consistently failed to make a meaningful contribution to either lower-level factor or the composite. Therefore, *MPH* may be measuring something apart from the other six tests in the ANAM.

Reliability is another essential component of assessment. It sets the upper limits of a test's validity, and it establishes confidence limits for interpreting test scores (Glutting, McDermott, & Stanley, 1987). Consequently, adequate reliability is a minimum requirement for clinical score interpretation. The reason is because low scores from each of the ANAM's seven test are often taken to be indicators of neuropsychological weaknesses and/or the possibility of a concussion (Wahlquist, Glutting, & Kaminski, 2019).

Reliability was a major drawback for the accepted, hierarchical CFA model. Internal-consistency reliability was appreciable for the composite ($\omega_H = .85$, $H = .86$), but it was deficient for the two lower-level factors ($\omega_S = .20$, $H = .41$ for *SRT* & $\omega_S =. 03$ & $H = .01$ for *Mental Efficiency*). Neither first-order factor possessed sufficient reliability to support clinical score interpretation. Results, therefore, reveal that practitioners should not interpret *either* of the two first-order factor scores or, *any* the seven test scores from the ANAM individually. Instead, clinicians should simplify interpretations to the one composite score.

The current study is notable in several respects. Its sample ($N = 218$ first testing, $N = 217$ on retesting) is the second largest among factor analyses of the ANAM. One study was superior. It included 9,883 soldiers (Naifeh et al., 2016). Comrey and Lee (1992) posited that a sample of 100 subjects is poor for a factor analysis, 200 is fair, 300 is good, 500 is very good, and 1,000 or more is excellent. By their standards, the current sample size is fair, the Naifeh et al. (2016) investigation is excellent, and the other four studies are somewhere between poor and fair.

In contrast to the global advice of Comrey and Lee (1992), more recent sample-size guidelines are based on quality. The recommendations are derived from Monte Carlo simulations that included variables such as sample size, the number of variables, magnitude of the factor loadings, and amount of factor overdetermination (number of variables per factor; Guadagnoli & Velicer, 1988; Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005; MacCallum, Widaman, Zhang, & Hong, 1999; Mundfrom, Shaw, & Ke, 2005; Velicer & Fava, 1998; Wolf, Harrington, Clark, & Miller, 2013). Watkins (in press) summarized the findings and offered useful look-up tables. The current investigation included a minimum of 217 subjects. We had seven tests and coefficients $\geq .40$ were considered appreciable. Accordingly, simulation guidelines revealed we needed a sample of 140 subjects.

The current study also is noteworthy with regard to the breadth of its variables. We evaluated a seven-test battery. Only one other investigation encompassed seven tests (Kabat et al., 2001). No prior study comprised more variables. To date, only 12 of the ANAM's 30 tests were submitted to factor analysis. The tests include the seven here, as well as: Code Substitution, Code Substitution Delayed, Spatial Processing, Logical Memory, and Delayed Memory.

Limitations

Our sample was the primary limitation. The size of our two, repeated samples was sufficient to conduct exploratory and confirmatory analyses. It also is unique among the youth soccer population who have reason to be tested using computerized neuropsychological test batteries based on the fact that they are involved in a sporting activity with a high incidence of sports related concussion, as well as the propensity to encounter RHI from heading a soccer ball. However, we must note that our sample was restricted to female soccer players at the high school level in one specific region of the Northeast United States and may not reflect the entire population of interscholastic soccer players across the country.

A second limitation is that analyses were conducted using an older version of the ANAM—as were at least three of the other factor analyses (One used a newer version of the test and it was difficult to ascertain which version was used in another.) A third drawback was our use of throughputs. As mentioned in the introduction, the ANAM offers three scores: speed, accuracy, and throughput (accuracy/speed).

Implications

Outcomes support inferences that scores from the ANAM should be parsimoniously interpreted as measuring a single dimension—one, umbrella construct. Luckily, newer versions of the ANAM offer a composite score. As found here, examining scores from the seven individual ANAM tests, or its first-order factors, is very risky and unlikely to be clinically relevant.

Future Directions

We recommend that our factor analyses be replicated and extended to a wider variety of subjects, age levels, ANAM tests, and score types. Equally, if not more important, studies need to build upon current findings and use them as a guide on how to examine the ANAM's external validity. Future criterion-related outcomes need to investigate two types of evidence: incremental validity and diagnostic validity.

Incremental validity is the "extent to which a measure adds to the prediction of a criterion beyond what can be predicted with other data" (Hunsley, 2003, p. 443). This type of validity is rooted in the scientific law of parsimony, which states: "what can be explained by fewer principles is needlessly explained by more" (Jones, 1952, p. 620). There is a loss of parsimony when clinicians switch from interpreting the ANAM's composite to its seven individual test scores. Accordingly, there should be compelling external validity supporting assumptions that these seven discrete scores provide criterion-related validity not already available in the factorially supported composite score—information that is pertinent to the increased risk for concussion as well as exposures to repetitive head impacts from heading a soccer ball.

Statistically significant group differences have been the traditional benchmark for determining whether a test has discriminant and/or criterion-related validity (*cf*. Glutting, Youngstrom, Ward, Ward, & Hale, 1997; Watkins, 2009). However, as noted by Elwood (1993), "significance alone does not reflect the size of the group differences nor does it imply the test can discriminate subjects with sufficient accuracy for clinical use" (p. 409, original italics). Diagnostic validity statistics such as sensitivity, specificity, and receiver operator curve statistics directly examine the individual aspects of clinical decision-making (Streiner, 2018; Watkins, 2009). Here too, future studies must demonstrate that individual test scores from the ANAM offer significantly greater diagnostic validity than that offered by the ANAM's parsimonious, and factorially valid, composite score.

Conflict of Interest

No potential conflict of interest was reported by the authors.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332. doi: 10.1007/bf02294359.

American Educational Research Association, American Psychological Association., National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, District of Columbia. doi:10.1037/e487712008-001

Bandalos, D. L., & Finney, S. J. (2019). Factor analysis: Exploratory and confirmatory. In Hancock, G. R., Stapleton, L. M., & Mueller, R. O. (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (2nd ed., pp. 98–122). Boston: Hogrefe.

Bentler, P. (1990). Comparative fit in structural models. *Psychological Bulletin*, *107*, 238–246. doi: 10.1037/0033-2909.107.2.238.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606. doi: 10.1037/0033-2909.88.3.588.

Bleibert, J., Kane, R. L., Reeves, D. L., Garmoe, W. S., & Halpern, E. (2000). Factor analysis of computerized and traditional tests used in mild brain injury research. *The Clinical Neuropsychologist*, *14*(*3*), 287–294. doi: 10.1076/1385-4046(200008)14:3;1-P;FT287.

Braden, J. P. (2013). Psychological assessment in school settings. In Graham, J. R., & Naglieri, J. A. (Eds.), *Handbook of psychology: Assessment psychology* (2nd ed., Vol. *10*, pp. 291–314). Hoboken, New Jersey: Wiley.

Braden, J. P., & Niebling, B. C. (2012). Using the joint test standards to evaluate the validity evidence for intelligence tests. In Flanagan, D. P., & Harrison, P. L. (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 739–757). Hoboken, New Jersey: Guilford.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, *38*(*1*), 25–56. doi: 10.1207/S15327906MBR3801_2.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. doi: 10.1207/s15327906mbr3601_05.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K., & Long, J. (Eds.), *Testing structural equation models* (, pp. 136–162). Newbury Park, California: Sage.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*, 796–846. doi: 10.1111/j.1467-6494.2011.00749.x.

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for multidimensionality and test interpretation. In Schweizer, K., & DiStefano, C. (Eds.), *Principles and methods of test construction: Standards and recent advancements* (, pp. 247–271). Göttingen, Germany: Hogrefe.

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler intelligence scale for children–fifth edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment*, *29*, 458–472. doi: 10.1037/pas0000358.

Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Carroll theory: Empirical and policy implications. *Applied Measurement in Education*, *32*(*3*), 232–248. doi: 10.1080/08957347.2019.1619562.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276. doi: 10.1207/s15327906mbr0102_10.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(*3*), 464–504. doi: 10.1080/10705510701301834.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*, 219–251. doi: 10.1111/j.1467-6494.2011.00739.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*, 207–230. doi: 10.1177/1094428114555994.

Cohen, R. J., Swerdlik, M. E., & Sturman, D. E. (2017). *Psychological testing and assessment: An introduction to tests and measurement*. New York: McGraw-Hill Education.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Crawford, C. B., & Koopman, P. (1979). Note: Inter-Rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills*, *49*(*1*), 223–226. doi: 10.2466/pms.1979.49.1.223.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi: 10.1007/bf02310555.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. doi: 10.1037/h0040957.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Newbury Park, California: Sage.

Elwood, R. W. (1993). Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review*, *13*, 409–419. doi: 10.1016/0272-7358(93)90012-b.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299. doi: 10.1037/1082-989x.4.3.272.

Ferrando, P. J., & Lorenzo-Seva, U. (2019). An external validity approach for assessing essential unidimensionality in correlated-factor models. *Educational and Psychological Measurement*, *79*(*3*), 437–461. doi: 10.1177/0013164418824755.

Ferrando, P. J., & Navarro-González, D. (2018). Assessing the quality and usefulness of factor analytic applications to personality measures: A study with the statistical anxiety scale. *Personality and Individual Differences*, *123*, 81–86. doi: 10.1016/j.paid.2017.11.014.

Flora, D. B. (2018). *Statistical methods for the social and behavioural sciences: A model-based approach*. Newbury Park, California: Sage.

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science*, *49*(*2*), 78–88. doi: 10.1037/cbs0000069.

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it. *European Journal of Psychological Assessment*, *33*(*6*), 399–402. doi: 10.1027/1015-5759/a000460.

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology*. Newbury Park, California: Sage.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement*, *71*(*3*), 551–570. doi: 10.1177/0013164410389489.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*(*1*), 93–111. doi: 10.1037/met0000064.

Gerbing, D., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling-a Multidisciplinary Journal*, *3*, 62–72. doi: 10.1080/10705519609540030.

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model based reliability in the WAIS-IV. *Multivariate Behavioral Research*, *48*, 639–662. doi: 10.1080/00273171.2013.804398.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*, 377–393. doi: 10.1177/0013164495055003002.

Glutting, J. J., McDermott, P. A., & Stanley, J. C. (1987). Resolving differences among methods of establishing confidence limits. *Educational and Psychological Measurement*, *47*, 607–614. doi: 10.1177/001316448704700307.

Glutting, J. J., Youngstrom, E. A., Watkins, M. W., & Frazier, T. W. (2007). ADHD and achievement: Meta-analysis of the child, adolescent, and adult literatures and a concomitant study with college students. *Journal of Learning Disabilities*, *40*, 49–65. doi: 10.1177/00222194070400010401.

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. (1997). Incremental efficacy of WISC-III factor scores in predicting achievement: What do they tell us? *Psychological Assessment*, *9*, 295–301. doi: 10.1037/1040-3590.9.3.295.

Gregory, R. J. (2014). *Psychological testing: History, principles, and applications* (7th ed.). Chicago, Illinois: Pearson.

Goldberg, L. R., & Velicer, W. F. (2006). Principles of exploratory factor analysis. In Strack, S. (Ed.), *Differentiating normal and abnormal personality* (2nd ed., pp. 209–337). Berlin, Germany: Springer.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Gorsuch, R. L. (2003). Factor analysis. In Shinka, J. A., & Velicer, W. F. (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (, pp. 143–164). Mahwah, New Jersey: John Wiley.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*, 265–275. doi: 10.1037/0033-2909.103.2.265.

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97–121). Washington, District of Columbia: American Psychological Association. doi:10.1037/12074-005

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In Cudek, R., duToit, S. H. C., & Sorbom, D. F. (Eds.), *Structural equation modeling: Present and future* (, pp. 195–216). Chicago: Scientific Software International.

Heeme, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(*3*), 319–336. doi: 10.1037/a0024917.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(*3*), 393–416. doi: 10.1177/0013164405282485.

Hershberger, S. L., & Marcoulides, G. A. (2013). The problem of equivalent structural models. In Hancock, G. R., & Mueller, R. O. (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 13–42). Charlotte, North Carolina: Information Age Publishing.

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, *65*, 202–226. doi: 10.1177/0013164404267287.

Horn, J. (1965). A rational and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In Hoyle, R. H. (Ed.), *Structural equation modeling: Concepts, issues, and applications* (, pp. 76–99). Newbury Park, California: Sage.

Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. *Psychological Assessment*, *15*, 443–445. doi: 10.1037/1040-3590.15.4.443.

IBM. (2020). *IBM SPSS statistics*. https://www.ibm.com/products/spss-statistics

Jones, W. T. (1952). *A history of western philosophy*. San Diego, California: Harcourt-Brace.

Jones, W. P., Loe, S. A., Krach, S. K., Rager, R. Y., & Jones, H. M. (2008). Automated neuropsychological assessment metrics (ANAM) and woodcock-Johnson III tests of cognitive ability: A concurrent validity study. *The Clinical Neuropsychologist*, *22*, 305–320. doi: 10.1080/13854040701281483.

Johnson, D. R., Vincent, A. S., Johnson, A. E., Gilliland, K., & Schlege (2008). Reliability and construct validity of the automated neuropsychological assessment metrics (ANAM) mood scale. *Archives of Clinical Neuropsychology*, *23*, 73–85. doi: 10.1016/j.acn.2007.10.001.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, Illinois: Scientific Software International.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36. doi: 10.1007/bf02291575.

Kabat, M. H., Kane, R. L., Jefferson, A. L., & DiPino, K. (2001). Construct validity of selected automated neuropsychological assessment metrics (ANAM) battery measures. *The Clinical Neuropsychologist*, *15*(*4*), 498–507. doi: 10.1076/clin.15.4.498.1882.

Kaminski, T. W., Groff, R. M., & Glutting, J. J. (2009). Examining the stability of automated neuropsychological assessment metric (ANAM) baseline test scores. *Journal of Clinical and Experimental Neuropsychology*, *31*(*6*), 689–697. doi: 10.1080/13803390802484771.

Lovell, M. R., & Collins, M. W. (1988). Neuropsychological assessment of the college football player. *J Head Trauma Rehabil*, *13*(*2*), 9–26. doi: 10.1097/00001199-199804000-00004. PMID: 9575253.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107–120. doi: 10.1037/0033-2909.100.1.107.

MacCallum, R., Roznowski, M., & Nowrwitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalizations on chance. *Psychological Bulletin*, *111*, 490–504. doi: 10.1037/0033-2909.111.3.490.

MacCallum, R., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99. doi: 10.1037/1082-989x.4.1.84.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*, 157–176. doi: 10.1037/1082-989x.12.2.157.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In Maydeu-Olivares, A., & McArdle, J. (Eds.), *Contemporary psychometrics: A Festschrift for Roderick P. McDonald* (, pp. 275–340). Mahwah, New Jersey: Erlbaum.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, *3*, 97–110. doi: 10.21500/20112084.854.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, New Jersey: Lawrence Erlbaum.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(*1*), 195–244. doi: 10.2466/pr0.1990.66.1.195.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177–185. doi: 10.1007/bf02289699.

Messick, S. (1989). Validity. In Linn, R. L. (Ed.), *Educational measurement* (, pp. 13–103). Stuttgart, Germany: Macmillan.

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*, 116–139. doi: 10.1080/10705511.2014.961800.

Mueller, R. O., & Hancock, G. R. (2019). *Structural equation modeling*. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (2nd ed., pp. 445–456). Milton Park, Abingdon, England: Routledge. doi:10.4324/9781315755649-33

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(*3*), 430–445. doi: 10.1037/0033-2909.105.3.430.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*, 159–168. doi: 10.1207/s15327574ijt0502_4.

Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Naifeh, J. A., Nock, M. K., Ursano, R. J., Vegella, P. L., Aliaga, P. A., Fullerton, C. S. et al. (2016). Neurocognitive function and suicide in U.S. army soldiers. *Suicide and Life-threatening Behavior*, *47*(*5*), 589–602. doi: 10.1111/sltb.12307.

Nasser, F., Benson, J., & Wisenbaker, J. (2002). The performance of regression-based variations of the visual scree for determining the number of common factors. *Educational and Psychological Measurement*, *62*, 397–419. doi: 10.1177/00164402062003001.

Norman, G. R., & Streiner, D. L. (2014). *Biostatistics: The bare essentials* (4th ed.). People's Medical Publishing.

Streiner, D. L. (2018). Commentary no. 26: Dealing with outliers. *Journal of Clinical Psychopharmacology*, *38*(3), 170–171. https://doi.org/10.1097/jcp.0000000000000865.

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research and Evaluation*, *17*, Article 15. https://scholarworks.umass.edu/pare/vol17/iss1/15/.

Peterson, R. A. (2000). A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters*, *11*(*3*), 261–275.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*, 28–56. doi: 10.1080/00273171.2012.710386.

Raiche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology*, *9*, 23–29. doi: 10.1027/1614-2241/a000051.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*, 329–353. doi: 10.1207/s15327906mbr3204_2.

Reeves, D. L., Winter, K. P., Bleiberg, J., & Kane, R. L. (2007). ANAM genogram: Historical perspectives, description, and current endeavors. *Archives of Clinical Neuropsychology*, *22*(*S1*), S15–S37. doi: 10.1016/j.acn.2006.10.013.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. doi: 10.1080/00273171.2012.715555.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*, 129–140. doi: 10.1080/00223891.2012.725437.

Reise, S. P., Bonifay, W., & Haviland, M. G. (2018). Bifactor modelling and the evaluation of scale scores. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 677–707). New York: Wiley New York.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 121–127. doi: 10.1007/s11336-008-9102-z.

Retzlaff, P., & Vanderploeg, R. D. (1999). *Construct validity study of the Space-flight Cognitive Assessment Tool*. In *Technical report MS80,665, Medical Operations*. Johnson Space Center.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Pearson.

Rice, V. J., Lindsay, G., Overby, C., Jeter, A., Alfred, P. E., Boykin, G. L. et al. (2011). *Automated Neuropsychological Assessment Metrics (ANAM) Traumatic Brain Injury (TBI): Human Factor Assessment* (, pp. 1–42). US Army Research Laboratory.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(*2*), 137–150. doi: 10.1037/met0000045.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(*3*), 223–237. doi: 10.1080/00223891.2015.1089249.

Roskos, P. G., Feller, J., & Chibnall, J. (2014). An exploratory factor analysis of the repeatable battery for the assessment of neuropsychological status and the automated neuropsychological assessment metrics. *Poster Session. Archives of Clinical Neuropsychology*, *2*(*6*), 568. doi: 10.1093/arclin/acu038.171.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*, 282–292. doi: 10.1037/a0025697.

Salvia, J., Ysseldyke, J. E., & Whitmer, S. (2017). *Assessment in special and inclusive education* (13th ed.). Houghton Mifflin.

Schretlen, D. (1997). *Brief Test of Attention professional manual*. Psychological Assessment Resources.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Schweizer, K. (2011). On the changing role of Cronbach's alpha in the evaluation of the quality of a measure. *European Journal of Psychological Assessment*, *27*, 143–144.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304–321. https://doi.org/10.1177/0734282911406653.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343. doi: 10.1007/BF02294360.

Selbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*, 1428–1441. doi: 10.1037/pas0000623.

Short, P., Cernich, A., Wilken, J. A., & Kane, R. L. (2007). Initial construct validation of frequently employed ANAM measures through structural equation modeling. *Archives of Clinical Neuropsychology*, *22S*, 63–77. doi: 10.1016/j.acn.2006.10.012.

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120. doi: 10.1007/s11336-008-9101-0.

Sijtsma, K. (2011). Psychological measurement between physics and statistics. *Theory & Psychology*, *22*(*6*), 786–809. doi: 10.1177/0959354312454353.

Sijtsma, K., & van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research*, *64*(*2*), 128–136. doi: 10.1097/nnr.0000000000000077.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180. doi: 10.1207/s15327906mbr2502_4.

Streiner, D. L. (1998). Factors affecting reliability of interpretations of scree plots. *Psychological Reports*, *83*, 687–694. doi: 10.2466/pr0.1998.83.2.687.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In Bollen, K. S., & Long, J. S. (Eds.), *Testing structural equation models* (, pp. 10–39). CA: Sage.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(*2*), 209–220. doi: 10.1037/a0023353.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, *7*, 769. doi: 10.3389/fpsyg.2016.00769.

Trost, S. G., Sallis, J. F., Pate, R. R., Freedson, P. S., Taylor, W. C., & Dowda, M. (2003). *American Journal of Preventative Medicine*, *25*(*4*), 277–282. doi: 10.1016/s0749-3797(03)00217-4.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. doi: 10.1007/bf02291170.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *3*(*1*), 4–70. doi:10.1177/109442810031002

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helms (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). Guilford. doi:10.1007/978-1-4615-4397-8_3

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231–251. doi: 10.1037/1082-989x.3.2.231.

Vincent, A. S., Roebuck-Spencer, T., Gilliland, K., & Schlegel, R. (2012). Automated neuropsychological assessment metrics (v4) traumatic brain injury battery: Military normaitve data. *Military Medicine*, *177*(*3*), 256–269. doi: 10.7205/MILMED-D-11-00289.

Wahlquist, V. E., Glutting, J. J., & Kaminski, T. W. (2019). Examining neurocognitive performance in interscholastic female football players over their playing careers. *Science and Medicine in Football.*, *3*(*2*), 115–124. doi: 10.1080/24733938.2018.1532104.

Wainer, H., & Feinberg, R. (2015). For want of a nail: Why unnecessarily long tests may be impeding the progress of western civilization. *Significance*, *12*(*1*), 16–21. doi: 10.1111/j.1740-9713.2015.00797.x.

Watkins, M. W. (2009). Errors in diagnostic decision making and clinical judgment. In Gutkin, T. B., & Reynolds, C. R. (Eds.), *Handbook of school Psychology* (4th ed., pp. 210–229). New York: Wiley.

Watkins, M. W. (2013). *Omega [Computer Software]*. Phoenix, AZ: Ed & Psych Associates.

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist.* doi: 10.1080/13854046.2017.1317364.

Watkins, M. W. (in press). *A step-by-step guide to exploratory factor analysis with R and RStudio*. Routledge.

Watkins, M. W., & Canivez, G. L. (2020). Assessing the psychometric utility of IQ scores: A tutorial using the Wechsler intelligence scale for children–fifth edition. *School Psychology Review*. doi: 10.1080/2372966X.2020.1816804.

Williams, J. M., Jerome, G. J., Kenow, L. J., Rogers, T., Sartain, T. A., & Darland, T. A. (2003). Factor structure of the coaching behavior questionnaire and its relationship to athlete variables. *The Sport Psychologist*, *17*, 16–34. doi: 10.1123/tsp.17.1.16.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*, 913–934. doi: 10.1177/0013164413495237.

Woodard, J., Marker, C., Tabanico, F., Miller, S., Dorsett, E., Cox, L. et al. (2002). A validation study of the automated neuropsychological assessment metrics (ANAM) in non-concussed high school players. *Journal of the International Neuropsychological Association*, *8*(*2*), 175.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442. doi: 10.1037/0033-2909.99.3.432.