Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Validity and Reliability of the Athis Multidimensional Attention Performance Measurement System in School-Aged Children: Evidence From an Unselected Reference Sample and Criterion Groups

Kerem Çetinkaya

keremcet01@gmail.com

Turkish Naval Academy, National Defence University

Research Article

Keywords: attention, ADHD, children, continuous performance test, classification accuracy, measurement invariance, norms, spectrum bias

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11155561/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. KÇ designed the Athis Multidimensional Attention Performance Measurement System evaluated in this study and is a founding partner, executive, and shareholder of Özne Yazılım Teknolojileri Sağlık ve Eğitim Hizmetleri A.Ş., the company that developed the system and holds its commercial rights; KÇ may therefore benefit financially from future commercial use of the system. To limit optimistic bias, model selection and model testing were conducted in separate samples; the primary classification accuracy estimates were based on an unselected reference group rather than a recruited comparison group; cut scores were cross-validated and re-derived from the age norms of the unselected reference group; key reported values were independently recomputed by a second analysis routine; and null and unfavorable results are reported in full. The author declares no other competing interests.

Abstract

Background

Attention difficulties are a common reason for referring children for psychological assessment, yet such decisions rest largely on informant ratings. Computerized performance measures provide an observer-independent record but are usually validated against recruited problem-free controls. This study examined the structural validity, reliability, and classification accuracy of the Athis Multidimensional Attention Performance Measurement System, a computerized three-subtest instrument, in children aged 6–11 against an unselected reference sample.

Methods

The measurement model was estimated in a norm sample of 1,973 Turkish children (split-half replication, confirmatory factor analysis, multi-group invariance). Criterion-related validity was examined in a separately assembled sample of 855 children: 203 with a school-recorded diagnosis of attention-deficit/hyperactivity disorder and no comorbid diagnosis, 97 flagged by teachers but never evaluated, and 555 reported as problem-free, 148 of whom had passed first-stage gifted-education screening. Scores were expressed against age norms from 1,116 unselected children. Classification accuracy was estimated by receiver operating characteristic analysis, cross-validated cut scores, and predictive values at a 7.6% base rate; convergent validity used Conners Teacher Rating Scale scores (n = 114).

Results

Each subtest yielded two replicable dimensions, response control deficit and accuracy. Response control deficit separated recorded diagnostic status in all three subtests (AUC = .903, .681, and .698), whereas accuracy, as expected from the construct definition, did not (AUC = .454–.572). Subtest 1 reproduced teacher judgment (AUC = .888), whereas Subtests 2 and 3 separated reported risk from recorded status (AUC = .757 and .750). Accuracy, not response control deficit, responded to selection for cognitive ability (g = 0.78 versus − 0.22). These estimates are conservative; against a teacher-nominated comparison group, the conventional design, the same contrasts yielded AUC = .992, .813, and .745. At the population base rate the at-least-one-subtest rule had a negative predictive value of 98.0%.

Conclusions

The system yields an age-referenced two-dimensional profile in each of three tasks and works as a rule-out procedure for population-level screening rather than as an identification instrument. Its accuracy values, estimated against an unselected reference sample, apply directly to unscreened populations, and

the comparison-group effect quantified here explains why conventional validation designs report higher values.

Background

Attention difficulties are among the most frequent reasons for which children are referred for psychological assessment, and their consequences are visible from the first school years. Early attention difficulties predict later academic achievement and educational attainment independently of early academic skills and social competence [1]. In addition, childhood attention-deficit/hyperactivity disorder (ADHD) has been associated with a threefold increase in the odds of not completing secondary education [2]. Similar findings have been reported for peer relationships, in that children with higher symptom levels accumulate more unreciprocated and absent friendship ties over a school year [3]. Attention in childhood is therefore a construct whose measurement has consequences, and the instruments used for it should be validated in the population in which they are applied.

In practice such measurement rests almost entirely on teacher and parent rating scales, which are efficient and ecologically anchored but remain bound to the informant who completes them. A meta-analysis of 22 studies comprising 3,947 children reported that teacher ratings converge with structured classroom observation only at $r = .21$ for inattention and $r = .29$ for hyperactivity–impulsivity [4]. Similarly, trifactor modeling of approximately 24,000 teacher–student dyads showed that informant perspective accounts for a substantial share of the variance in behavioral ratings [5]. Such discrepancies are not random error; they carry context-bound information about where a behavior is displayed [6, 7], and latent parent–teacher correlations for inattention remain weak throughout development [8]. This creates a specific need for the practitioner: standardized instruments are preferred over unaided judgment [9], and decisions reached from identical assessment data have proved unreliable across practitioners [10]. An observer-independent record of how a child responds may therefore complement the ratings already in use without replacing them.

Computerized attention performance measures provide such a record, and reviews of their contribution are consistent and cautious: pooled accuracy indices cluster in the .70–.80 band, and these instruments index attention and impulsivity objectively without sufficing for decisions on their own [11–13]. Their indices are also multidimensional. An exploratory analysis of 931 clinical cases concluded that a single total score is indefensible [14], and a confirmatory analysis of 452 children rejected the one-factor solution in favor of response inhibition and sustained attention [15, 16]. The question is therefore which dimension answers which question, and for which assessment decision. It is also a theoretical question. Behavioral inhibition has been positioned as the primary deficit [17] and is among the domains most consistently impaired [18, 19], whereas multiple-pathway accounts hold that inhibitory, motivational, and timing-related impairments separate and distribute heterogeneously [20–22].

A further problem concerns the design of validation studies and is psychometric in nature. Attention traits are distributed dimensionally rather than categorically [23, 24]. A comparison group assembled by

asking teachers to nominate children with no reported problem therefore represents a selected lower tail of the population. A reference sample departing from representativeness yields systematically biased norm scores [25], and sampling cases and controls through separate gates narrows the spectrum and inflates measured accuracy [26, 27]. The consequences are concrete in any screening application, since positive predictive power depends jointly on the screener–criterion relation and on the base rate in the population screened [28, 29], and reporting standards require sampling and recruitment to be stated explicitly [30]. Here this problem was treated as a quantity to be estimated, and the same contrasts were computed twice within one data set.

A final confound concerns general cognitive ability. High ability and attention difficulty co-occur, and the behavioral signal of ADHD is attenuated in high-ability children, whose difficulties are readily overlooked against average-ability comparison groups [31–34]. An instrument used across the full ability range should therefore show which of its dimensions is sensitive to ability and which is sensitive to attention.

The Present Study

The Athis Multidimensional Attention Performance Measurement System (hereafter, the system) is a computerized three-subtest instrument developed for individuals aged 6 years and older; the present study examined it in children aged 6–11. The present study had four objectives. The first was to establish the measurement structure and internal consistency of the system in a broad norm sample, and to test whether that structure holds across age and sex groups, whether the indicators change with age as expected of a developmental measure, and whether the dimension scores are stable over a retest interval of approximately ten weeks. The second was to estimate how well the dimension scores separate school-recorded ADHD status and teacher-reported risk from an unselected general-population reference group, and whether the three subtests carry the same or different information. The third was to examine convergent validity against Conners Teacher Rating Scale scores and discriminant validity against a group selected toward high cognitive ability. The fourth was to quantify, within the same data set and with the clinical group held constant, how far substituting a teacher-nominated comparison group for the unselected reference group changes the accuracy estimates. On the basis of the literature reviewed above, we expected each subtest to yield two dimensions, response control deficit and accuracy; response control deficit to separate recorded status far more strongly than accuracy; and accuracy, rather than response control deficit, to respond to selection for cognitive ability. The study groups are defined in Table 1.

Table 1

Definition, Size, and Analytic Role of the Study Groups

Group (short name)	*n*	Definition	Role in the analysis
Norm Sample	1,973	All records of children aged 6–11 retained after the integrity checks; includes the Labeled Sample	Measurement model, internal consistency, measurement invariance, developmental validity
General-Population Reference Group	1,116	Members of the Norm Sample reached by none of the study's selection routes (teacher flagging, teacher nomination, special-education register); not screened as problem-free	Source of the age norms and percentile cut scores; reference group for all classification accuracy estimates
Labeled Sample	855	Separately assembled sample comprising the groups below	Criterion-related validity and differential roles of the subtests
Clinical ADHD Group	203	Children whose physician-assigned ADHD diagnosis is recorded in the MoNE special-education register and confirmed by teacher and family; ADHD diagnosis only, no comorbid diagnosis; assessed medication-free	Index group; primary contrast against the General-Population Reference Group
Teacher-Reported Risk Group	97	Children in whom the class teacher indicated a possible attention problem; never clinically evaluated and not in any register	Differential roles of the subtests
Selected Control Group	555	Children identified by class teachers as having no psychiatric history and no chronic illness, that is, as problem-free	Not the primary reference; quantification of comparison-group selection effects
Gifted-Screened Group	148	Members of the Selected Control Group who passed the first-stage screening of the national gifted-education identification program	Discriminant validity against general cognitive ability
Non-Screened Group	407	Remaining members of the Selected Control Group	Comparison group matched to the Gifted-Screened Group on teacher nomination
Retest Group	123	Children to whom Subtest 1 was readministered after approximately ten weeks	Temporal stability
Conners Subsample	114	Children rated with the Conners Teacher Rating Scale (95 from the Selected Control	Criterion-related validity against teacher ratings

Note. MoNE = Ministry of National Education. The name General-Population Reference Group indicates that the group passed through no selection process, not that it was randomly sampled. Indented rows are subsets of the row above.

Group (short name)	*n*	Definition	Role in the analysis
		Group, 16 from the Clinical ADHD Group, 3 from the Teacher-Reported Risk Group)	

Note. MoNE = Ministry of National Education. The name General-Population Reference Group indicates that the group passed through no selection process, not that it was randomly sampled. Indented rows are subsets of the row above.

Methods

Participants and Study Groups

In this cross-sectional psychometric study the measurement model was estimated in one sample and criterion-related validity in another, with the classification accuracy components reported according to STARD 2015 [30]. Data were collected on the Anatolian side of Istanbul, Türkiye, through school-based and general-population administration, after ethics approval had been obtained (27 November 2025). All groups are defined in Table 1, and only the short names given there are used hereafter. Children registered in the Ministry of National Education (MoNE) special-education system with intellectual disability, Down syndrome, autism spectrum disorder, physical disability, or hearing or visual impairment were not eligible.

Norm Sample and General-Population Reference Group

The raw data file comprised 2,200 records, to which we applied eight pre-specified integrity criteria in fixed order, covering duplicate records, subtests that were not administered, out-of-range values, complete non-response, the absence of any correct response, and weighted timing totals above the algebraic maximum; the criteria and the number of records removed by each are listed in Additional file 1. The Norm Sample then comprised 1,973 children aged 6–11 (52.1% male; n = 252, 462, 521, 509, 175, and 54 at ages 6 through 11; M = 8.03 years, SD = 1.27). No record was removed for a high response count, since excessive responding is the behavior that the response control deficit dimension registers. The consequence of applying such a rule is reported in Additional file 1.

Norm Sample records were then matched to the Labeled Sample on an eight-variable fingerprint of demographic and performance values unique within each file. A total of 1,116 children were reached by none of the study's selection routes: no teacher was asked about them and none appears in a special-education register. These children constitute the General-Population Reference Group (50.4% male; M = 8.06 years, SD = 1.26). The group is unselected with respect to attention and was not screened as problem-free, so it is expected to contain children with unidentified difficulties, which makes estimates against it conservative. It supplies the age norms against which every score in the study is expressed and is the reference in all classification accuracy analyses.

Labeled Sample

The Clinical ADHD Group comprised children whose physician-assigned ADHD diagnosis is recorded in the MoNE special-education register and was independently confirmed by the class teacher and by the family. The group comprised children with an ADHD diagnosis only; children with comorbid diagnoses were not included. Status was taken from official school records without contact with the diagnosing physicians, which removes any influence of the test result on the criterion, and families were asked to withhold medication for one week before administration. This definition corresponds to the information available in the settings where such instruments are used, namely an administratively recorded diagnosis corroborated by the two informants who see the child daily. It is therefore an ecologically grounded criterion, although it is not a research-grade diagnostic confirmation, and its limits are treated below.

The Teacher-Reported Risk Group comprised children in whom the class teacher indicated a possible attention problem without being confident of it; none had undergone psychiatric evaluation and none appears in a special-education register. These are the children for whom the question of further evaluation most often arises. The Selected Control Group was identified by explicit criteria given to class teachers, namely no psychiatric history and no chronic illness, and comprised children for whom no academic or social difficulty was reported. Within this group, 148 children had been nominated for the national gifted-education identification program and had passed its first-stage group screening. Because the second stage involves individual assessment under a limited quota, they should be regarded as selected toward high cognitive ability and not as identified gifted children. These form the Gifted-Screened Group and the remainder the Non-Screened Group, the two differing only in cognitive-ability screening and matched on teacher nomination.

All administrations in the Labeled Sample took place in the children's own schools, in a quiet room, on identical hardware, and with a psychologist present. The administering psychologists were aware of group membership, but all scores are generated by the software without human interpretation, and the reference standard had been recorded before administration. The reference, risk, and clinical groups did not differ in age, Kruskal–Wallis $H = 2.28$, $p = .319$, but they did differ in sex composition, $\chi^2(2) = 22.83$, $p < .001$ (50.4%, 46.4%, and 68.0% boys), which is consistent with the literature [35]. Subtest 1 was readministered to 123 children after approximately ten weeks, and teacher ratings were obtained for 114.

Measures

The Athis Multidimensional Attention Performance Measurement System

The system was developed for individuals aged 6 years and older; the present study examined it in children aged 6–11. It comprises three subtests administered through a computer and headphones in approximately 20 min, all responses being made through a single key; the term multidimensional refers to the six-score profile formed by two dimensions replicated in three task paradigms. In Subtest 1 (visual matching) the child presses when a moving shape matches a stationary silhouette and withholds responding when it does not. In Subtest 2 (awareness–alertness) paired images are presented for a limited time and the child responds as soon as a difference is detected. In Subtest 3 (auditory–verbal attention) an animated story is presented with a spoken narration and the child responds as soon as a mismatch between the two is noticed. The three paradigms are described in full in Additional file 1.

The software records every key press with its time stamp and derives a set of indices for each subtest. These cover the number of targets captured and the total number of responses, the timing of the captures, responses made before a target is reached or to a non-target, responses that fall in no response window, repeated responses within a window, an index of the clustering of errors, and ratios derived from these counts. Several of these indices are by design transformations or subsets of others. Only mutually independent indices therefore entered the model, under two criteria specified in advance: an index that was a transformation or a subset of another was not entered, and where two indices within a dimension correlated above $\rho = .80$ the more derived one was removed.

The indicators retained are defined as follows. In Subtest 1, commission-type responses are presses made before the moving shape has reached the silhouette, together with presses on trials in which the shapes match but do not meet. Extra responses are presses to non-matching shapes and presses that fall in no response window. Hits are matching shapes captured at the moment of superimposition. Timing precision is the weighted timing total per correct response on a 0–100 scale; it is bounded above by construction and independent of how many targets were captured (ρ with hits = .52 in the Norm Sample). In Subtests 2 and 3, net errors are presses that correspond to no target, and extra responses are repeated presses within a response window in which a single press is allowed. Error clustering counts consecutive errors and thus indexes the tendency of errors to occur in runs. Hits are targets detected within the response window, and the fast-hit ratio is the proportion of hits falling in the early half of that window. Commission-type responses, extra responses, net errors, and error clustering form the response control deficit dimension; hits, timing precision, and the fast-hit ratio form the accuracy dimension. Because a weighted timing total above 100 points per correct response is algebraically impossible, records exceeding this bound were treated as invalid (six records; Additional file 1).

Conners Teacher Rating Scale

The short revised form of the scale, which derives from the original Conners scale [36], was used in its Turkish adaptation, whose psychometric properties have been reported as acceptable [37]. Five scores were analyzed: the Total score, the Oppositional, Cognitive Problems/Inattention, and Hyperactivity subscales, and the ADHD Index.

Scoring and Data Analysis

The response control deficit indicators, count variables with marked positive skew, were log(1 + x) transformed, converted to *z* scores against the age-specific mean and standard deviation of the General-Population Reference Group, and averaged within dimension. The age norms and percentile cut scores were derived from that group alone, because it is the only part of the data that passed through no selection route; the measurement model was estimated in the full Norm Sample, whose wider score range identifies the structure more precisely. Because no indicator was normally distributed (D'Agostino K^2, all $p < .001$), we used Spearman's coefficient for correlations and the Mann–Whitney U test for group comparisons, with Hedges' *g* and Cliff's δ as effect sizes and bootstrap confidence intervals from 5,000 within-group resamples.

For structural validity the Norm Sample was split at random into halves stratified by age and sex, the number of dimensions was fixed by parallel analysis in the calibration half [38], and the pattern was tested in the validation half. Extraction was by principal components with varimax rotation, because a common-factor solution is not identified when a dimension has only two indicators, as in Subtest 1. Congruence was evaluated with Tucker's coefficient after Procrustes rotation, and fit was read from CFI and SRMR [39]. Reliability was estimated by standardized Cronbach's alpha [40] and McDonald's omega [41], invariance by multi-group confirmatory factor analysis against the $\Delta CFI \leq .010$ criterion [42–44], and stability by ICC(3,1) with the minimal detectable change at 95% confidence [45]. Classification accuracy was examined by receiver operating characteristic analysis with DeLong confidence intervals and tests for correlated areas [46], cut scores by the Youden index [47] with five-fold cross-validation, and multiplicity by the Benjamini–Hochberg false discovery rate [48]. Dependent correlations were compared with Williams's test [49], and predictive values were computed at a population base rate of 7.6% [50]. The full analytic procedure is given in Additional file 1, and the completed STARD 2015 checklist in Additional file 2.

All analyses used code written for this study in Python 3.11, and key reported values were recomputed by a second routine written independently of the first; all 47 checks agreed. Because the author designed the instrument, four safeguards against optimistic bias were built in: model selection and testing were conducted in separate samples, the age norms and the primary estimates rest on an unselected reference group, cut scores were cross-validated, and null and unfavorable results are reported in full.

Use of generative artificial intelligence

Claude (Anthropic) was used as an assistive tool under the author's instruction and supervision to generate the Python analysis code, to produce tables and figures from the analysis outputs, to translate the Turkish text into English, and to verify bibliographic metadata against publisher records. The study design, objectives, group definitions, interpretation of the findings, and all decisions were made by the author, who reviewed every output, made the necessary revisions, and takes full responsibility for the content.

Ethics

The study was approved by the National Defence University Social and Human Sciences Ethics Committee (reference number E-35592990-050.04-5496359, 27 November 2025), and MoNE permissions were obtained, with written informed consent from parents or legal guardians and verbal assent from the children. Data were anonymized before analysis.

Results

Descriptive Statistics and Structural Validity

Descriptive statistics for the Norm Sample are presented in Additional file 1: Table S5. The response control deficit indicators were markedly right-skewed with common floor values, whereas the accuracy indicators were approximately symmetric. Parallel analysis supported a two-component solution in all three subtests, and the components were labeled response control deficit and accuracy. The calibration pattern replicated in the validation half at Tucker's $\varphi \geq .991$ for every dimension, a level conventionally read as factorial equality, and the two-dimension model fitted better than the one-dimension model in all three subtests (CFI = .959, .969, and .990 against .824, .953, and .900). In Subtest 2, however, the difference was slight (ΔCFI = .016) and the one-dimension model already met conventional criteria. Sampling adequacy was modest and, for Subtest 3, below the conventional floor (KMO = .648, .612, and .544), as is expected when a small set of deliberately non-redundant indicators is factored (Table 2).

Table 2
Structural Validity and Internal Consistency of the Three Subtests

Index	Subtest 1	Subtest 2	Subtest 3
Kaiser–Meyer–Olkin	.648	.612	.544
Bartlett χ^2 (*df*)	2,026 (6)	1,554 (10)	1,679 (10)
Components retained by parallel analysis	2	2	2
Tucker's φ — response control deficit	.999	.998	.996
Tucker's φ — accuracy	.997	.991	.998
Two-dimension model: χ^2 (*df*)	42.85 (1)	47.34 (4)	17.87 (4)
Two-dimension model: CFI	.959	.969	.990
Two-dimension model: SRMR	.068	.037	.043
Two-dimension model: TLI	.757	.923	.974
Two-dimension model: RMSEA	.206	.105	.059
One-dimension model: CFI	.824	.953	.900
One-dimension model: SRMR	.122	.057	.111
Response control deficit: mean *r*	.647	.551	.524
Response control deficit: α	.785	.786	.767
Response control deficit: ω	.821	.829	.802
Accuracy: mean *r*	.516	.190	.395
Accuracy: α	.681	.319	.566
Accuracy: ω	.708	.577	.671

Note. Congruence coefficients compare the calibration and validation halves of the Norm Sample after Procrustes rotation. Confirmatory fit is estimated in the validation half. TLI and RMSEA are reported for completeness but are not interpreted for Subtest 1, where *df* = 1. The accuracy dimension of Subtests 2 and 3 does not reach composite-level consistency, and its two indicators are reported separately in all subsequent analyses.

The two accuracy indicators of Subtest 1, hits and timing precision, correlate at ρ = .52 and yield α = .681 and ω = .708. In Subtests 2 and 3 the accuracy indicators remain largely independent (ρ = .19 and ρ = .40) and do not reach composite-level consistency, so for those two subtests the number of hits alone was used as the accuracy indicator in all subsequent analyses.

Measurement Invariance, Developmental Validity, and Stability

Metric invariance was supported in five of six comparisons but not in the sixth, Subtest 1 across age groups, where the change slightly exceeded the criterion (ΔCFI = − .012). Scalar invariance failed in Subtest 1 across age and sex and in Subtest 2 across sex, and it held in the other three comparisons (Additional file 1: Table S6). Equality of loadings indicates that the same construct is measured in each group, while the failure of intercept equality shows that raw scores cannot be compared across ages and that age norms are obligatory.

All five accuracy indicators and both response control deficit indicators of Subtest 1 changed monotonically across the six age levels with no reversal. Between ages 6 and 11, commission-type responses in Subtest 1 fell from 38.2 to 20.7 and extra responses from 20.9 to 4.4, while hits rose from 34.2 to 50.7 and timing precision from 86.2 to 89.7; hits and the fast-hit ratio rose in Subtests 2 and 3 as well (|ρ| = .163–.462, all $p < .001$; Additional file 1: Table S2). The response control deficit indicators of Subtests 2 and 3 behaved differently: their age trends were weak (ρ between − .120 and + .013) and non-monotonic, which indicates that response control is measured with limited precision there.

The dimensions are related across subtests without being interchangeable (accuracy ρ = .13–.25, response control deficit ρ = .19–.29) and are weakly to moderately related within a subtest (Additional file 1: Table S1). Over approximately ten weeks, stability in Subtest 1 was moderate, ICC(3,1) = .570 [.437, .678] for accuracy and .534 [.394, .649] for response control deficit, with small practice effects and minimal detectable change values of 1.32 and 1.52 *z* units (Additional file 1: Table S3). Subtests 2 and 3 were not readministered.

Criterion-Related Validity Against School-Recorded Status

The primary contrast compares the Clinical ADHD Group with the General-Population Reference Group, against whose age norms all scores are expressed. In all three subtests the groups separated substantially on response control deficit and only marginally on accuracy. The response control deficit composites yielded AUC = .903 [.883, .924], .681 [.639, .723], and .698 [.655, .740] for Subtests 1, 2, and 3 (Cliff's δ = .81, .36, and .40), whereas the accuracy indicators reached only .572 [.528, .616], .454 [.410, .498], and .466 [.426, .506]. Two of these three contrasts were statistically reliable but negligible in magnitude and one was not reliable ($p = .001$, .035, and .123), and in Subtests 2 and 3 the direction was reversed. Every within-subtest comparison between the two dimensions was significant (DeLong $z = 13.81$, 6.64, and 7.55; all $p < .001$). Subtest 1 differed from the other two ($z = 10.45$ and 8.79, both $p < .001$), which did not differ from one another ($z = -0.62$, $p = .533$) (Tables 3 and 4; Fig. 1).

Table 3
Dimension Scores by Group

Dimension	General-Population Reference (*n* = 1,116)	Teacher-Reported Risk (*n* = 97)	Clinical ADHD (*n* = 203)	Selected Control (*n* = 555)
Subtest 1 — Response control deficit	0.00 (0.84)	1.14 (0.60)	1.40 (0.87)	−0.75 (0.56)
Subtest 1 — Accuracy	0.00 (0.84)	−0.30 (0.91)	−0.19 (0.88)	0.54 (0.64)
Subtest 2 — Response control deficit	0.00 (0.81)	−0.23 (0.61)	0.65 (1.16)	−0.39 (0.48)
Subtest 2 — Accuracy	0.00 (0.76)	0.08 (0.57)	0.13 (0.82)	0.09 (0.65)
Subtest 3 — Response control deficit	0.00 (0.80)	−0.19 (0.50)	0.85 (1.47)	−0.17 (0.60)
Subtest 3 — Accuracy	0.00 (0.82)	0.34 (0.46)	0.14 (0.61)	0.34 (0.58)

Note. Values are means (standard deviations) of *z* scores computed against the age norms of the General-Population Reference Group. Higher response control deficit scores indicate poorer response control; higher accuracy scores indicate better accuracy.

Table 4
Areas Under the Curve With 95% DeLong Confidence Intervals

Measure	Clinical ADHD vs General-Population Reference	Teacher-Reported Risk vs General-Population Reference	Clinical ADHD vs Teacher-Reported Risk	Clinical ADHD vs Selected Control
Subtest 1 — Response control deficit	.903 [.883, .924]	.888 [.863, .913]	.577 [.510, .644]	.992 [.982, 1.000]
Subtest 1 — Accuracy	.572 [.528, .616]	.595 [.534, .656]	.477 [.406, .547]	.747 [.705, .788]
Subtest 2 — Response control deficit	.681 [.639, .723]	.413 [.357, .469]	.757 [.701, .813]	.813 [.775, .851]
Subtest 2 — Accuracy	.454 [.410, .498]	.467 [.413, .520]	.485 [.418, .551]	.487 [.438, .536]
Subtest 3 — Response control deficit	.698 [.655, .740]	.446 [.391, .502]	.750 [.695, .805]	.745 [.702, .789]
Subtest 3 — Accuracy	.466 [.426, .506]	.386 [.341, .432]	.594 [.529, .660]	.598 [.553, .644]

Note. Accuracy rows for Subtests 2 and 3 use the number of hits. All areas are oriented so that a value above .50 means the first-named group is the more impaired: higher response control deficit scores, lower accuracy scores. A value below .50 on an accuracy row therefore indicates that the first-named group was the more accurate. The final column is reported solely to quantify the effect of substituting a teacher-nominated comparison group for the unselected reference group.

The pattern therefore concerns response control and not attentional capture. In Subtest 1 the clinical group captured fewer targets by a small margin (g = − 0.23) beside a difference of g = + 1.65 on response control deficit, and in Subtests 2 and 3 it captured slightly more (g = + 0.17 and + 0.18). This pattern follows from the definition of the two dimensions. Children with a recorded diagnosis capture as many targets as their peers; what distinguishes them is the quality of responding, namely how many responses are emitted in capturing those targets. The criterion-related information therefore lies in the response control deficit dimension by design, and weak separation on accuracy indicates that this dimension measures something else, not that it measures poorly. Sex does not account for the separation: partial η^2 was .229 for group, $F(1, 1315) = 390.86$, $p < .001$, and .007 for sex, $F(1, 1315) = 8.63$, $p = .003$, with no interaction ($p = .293$).

Differential Roles of the Subtests

Against the same reference group the three subtests answered different questions (Table 4; Fig. 2). Two questions are relevant here: whether a subtest separates children whom teachers suspect of an attention problem from the general population, and whether it separates those children from children with a recorded diagnosis. Subtest 1 answered the first question well and the second poorly. On its

response control deficit composite the Teacher-Reported Risk Group scored far higher than the reference group (AUC = .888 [.863, .913]), almost as high as the clinical group. The teacher-flagged and the diagnosed children were therefore barely distinguishable (AUC = .577 [.510, .644]). Subtests 2 and 3 showed the opposite pattern. On their response control deficit composites the teacher-flagged children did not score higher than the reference group but slightly lower, that is, better (AUC = .413 [.357, .469] and .446 [.391, .502]). The diagnosed children, by contrast, scored clearly higher than the teacher-flagged children (AUC = .757 [.701, .813] and .750 [.695, .805]). On the accuracy indicator of Subtest 3 the teacher-flagged children likewise outperformed the reference group (g = + 0.43, p < .001). In short, Subtest 1 registers the difficulty a teacher notices, whereas Subtests 2 and 3 register the difference between a teacher's suspicion and a recorded diagnosis.

Cut Scores and Expected Classification

On the Subtest 1 response control deficit score the Youden-optimal cut score yielded sensitivity .901 and specificity .799, reproduced under five-fold cross-validation (.898 and .789). At a population prevalence of 7.6% the positive predictive value was 27.0% and the negative predictive value 99.0%.

Screening use requires thresholds that are stable across ages and independent of any particular contrast, so cut scores were re-derived for each subtest from the age-specific 90th percentile of the reference group. At least one subtest above threshold yielded sensitivity .818 and specificity .748, at least two .399 and .954, and all three .108 and .993 (Table 5). The number of thresholds exceeded orders the groups as expected. No threshold was exceeded by 93.9% of the Selected Control Group, 74.8% of the reference group, 27.8% of the Teacher-Reported Risk Group, and 18.2% of the Clinical ADHD Group.

Table 5
Expected Classification Under Age-Referenced Cut Scores

Rule	Sensitivity	Specificity	Flagged in the reference group	PPV	NPV
At least one subtest above the 90th percentile	.818 (166/203)	.748 (835/1,116)	25.2%	21.1%	98.0%
At least two subtests above the 90th percentile	.399 (81/203)	.954 (1,065/1,116)	4.6%	41.8%	95.1%
All three subtests above the 90th percentile	.108 (22/203)	.993 (1,108/1,116)	0.7%	55.4%	93.1%
Youden-optimal cut score, Subtest 1	.901	.799	—	27.0%	99.0%

Note. PPV = positive predictive value; NPV = negative predictive value, both at a population base rate of 7.6%. Counts in parentheses are the numerators and denominators from which sensitivity and specificity were obtained. Percentile rules use the age-specific 90th percentile of the General-Population Reference Group for the response control deficit composite of each subtest. Five-fold cross-validation of the Youden cut score reproduced sensitivity .898 and specificity .789.

The Effect of a Teacher-Nominated Comparison Group

All estimates were recomputed with the Selected Control Group substituted for the General-Population Reference Group, holding the clinical group, instrument, scoring rule, and analyst constant. The substitution raised the response control deficit values from .903, .681, and .698 to .992, .813, and .745, an increase of .088, .132, and .048, and it raised the apparent specificity of the Subtest 1 cut score from .799 to 1.000 (Table 4; Fig. 3). The mechanism is visible in the comparison group itself. On the Subtest 1 response control deficit score it lies almost a full standard deviation below the reference group (g = − 0.99; AUC = .796) with a dispersion 66% as wide, and the corresponding differences are smaller in Subtests 2 and 3 (g = − 0.54 and − 0.23).

Criterion-Related Validity Against Teacher Ratings

In the Conners Subsample all five scores separated the Clinical ADHD Group from the remaining children with large effect sizes (Hedges' g = 2.12–2.59; all p < .001). Across the whole rated subsample the Subtest 1 response control deficit composite was moderately associated with every Conners score (ρ = .471–.531), and among the 98 children without a recorded diagnosis these associations fell to ρ = .254–.319. With age partialled, only the coefficient with the Total score survived false discovery rate control (ρ = .303, q = .031), while the one with Cognitive Problems/Inattention did not (ρ = .268, q = .057). The accuracy indicator of Subtest 3 produced a different pattern, its associations being strongest within this range and specific to inattention: with age partialled, ρ = −.442 with Cognitive Problems/Inattention (q < .001) and ρ = −.298 with the Total score (q = .031) (Table 6). It was more strongly associated with

Cognitive Problems/Inattention than with Oppositional behavior, $t(95) = -2.42$, $p = .017$, and more strongly than was the Subtest 1 composite, $t(95) = -5.43$, $p < .001$. Of the three coefficients that survived false discovery rate control, the one specific to inattention therefore belonged to Subtest 3 accuracy and not to the dimension that separates recorded status.

Table 6

Associations Between the Measurement Dimensions and Conners Teacher Rating Scale Scores

Measure	Total	Oppositional	Hyperactivity	Cognitive Problems/Inattention	ADHD Index
Whole Conners Subsample (*n* = 114)					
Subtest 1 — Response control deficit	.53***	.47***	.49***	.52***	.52***
Subtest 1 — Accuracy	−.28**	−.24*	−.29**	−.29**	−.31***
Subtest 2 — Response control deficit	.16	.19*	.24**	.16	.26**
Subtest 2 — Accuracy	−.15	−.11	−.09	−.17	−.11
Subtest 3 — Response control deficit	.09	.09	.16	.14	.14
Subtest 3 — Accuracy	−.34***	−.29**	−.25**	−.43***	−.27**
Children without a recorded diagnosis, age partialled (*n* = 98)					
Subtest 1 — Response control deficit	.30**	.24*	.22*	.27**	.26**
Subtest 1 — Accuracy	−.17	−.13	−.15	−.13	−.18
Subtest 2 — Response control deficit	−.08	−.03	.07	−.03	.06
Subtest 2 — Accuracy	−.17	−.13	−.07	−.18	−.11

Note. Upper panel: Spearman coefficients in the whole rated subsample. Lower panel: age-partialled coefficients among children without a recorded diagnosis. Three coefficients survived Benjamini–Hochberg control of the false discovery rate: the Subtest 3 accuracy indicator with Cognitive Problems/Inattention ($q < .001$) and with the Total score ($q = .031$), and the Subtest 1 response control deficit composite with the Total score ($q = .031$). *$p < .05$. **$p < .01$. ***$p < .001$.

Measure	Total	Oppositional	Hyperactivity	Cognitive Problems/Inattention	ADHD Index
Subtest 3 — Response control deficit	−.13	−.13	−.02	.00	−.05
Subtest 3 — Accuracy	−.30**	−.23*	−.20	−.44***	−.22*

Note. Upper panel: Spearman coefficients in the whole rated subsample. Lower panel: age-partialled coefficients among children without a recorded diagnosis. Three coefficients survived Benjamini–Hochberg control of the false discovery rate: the Subtest 3 accuracy indicator with Cognitive Problems/Inattention ($q < .001$) and with the Total score ($q = .031$), and the Subtest 1 response control deficit composite with the Total score ($q = .031$). *$p < .05$. **$p < .01$. ***$p < .001$.

Discriminant Validity Against Cognitive Ability

The Gifted-Screened Group ($n = 148$) and the Non-Screened Group ($n = 407$) differ only in selection toward high cognitive ability, teacher nomination being held constant. In Subtest 1 the two dimensions dissociated: the difference on accuracy was substantial ($g = 0.78$, 95% CI [0.63, 0.95]), while that on response control deficit, although also reliable, was small ($g = -0.22$ [− 0.41, − 0.03], $p = .009$). The difference between the two effect sizes, tested by bootstrap in the same participants, was significant ($\Delta|g| = 0.56$ [0.31, 0.81], $p < .001$). In Subtests 2 and 3 no difference reached a meaningful magnitude ($|g| \leq 0.16$) (Table 7).

Table 7
Discriminant Validity Against General Cognitive Ability

Dimension	Gifted-Screened Group (*n* = 148)	Non-Screened Group (*n* = 407)	Hedges' *g*	95% CI	*p*
Subtest 1 — Accuracy	0.88	0.41	0.78	[0.63, 0.95]	< .001
Subtest 1 — Response control deficit	−0.84	−0.72	−0.22	[− 0.41, − 0.03]	.009
Subtest 2 — Accuracy	0.15	0.07	0.13	[− 0.06, 0.31]	.262
Subtest 2 — Response control deficit	−0.34	−0.40	0.12	[− 0.05, 0.32]	.048
Subtest 3 — Accuracy	0.41	0.31	0.16	[− 0.02, 0.34]	.132
Subtest 3 — Response control deficit	−0.20	−0.16	−0.06	[− 0.21, 0.12]	.671

Note. Values are means of age-standardized *z* scores. The difference between the two absolute effect sizes in Subtest 1, computed in the same participants, was Δ|*g*| = 0.56, 95% CI [0.31, 0.81], *p* < .001 by bootstrap.

The groups order along two largely separable axes of the Subtest 1 profile: on accuracy they follow cognitive selection (Gifted-Screened 0.88; Non-Screened 0.41; General-Population Reference 0.00; Clinical ADHD − 0.19; Teacher-Reported Risk − 0.30), and on response control deficit they follow attention status (− 0.84, − 0.72, 0.00, + 1.40, and + 1.14), with the reference group defining the origin of both axes (Fig. 4).

Discussion

The present study established the measurement structure of the system in a broad norm sample, expressed every score against the age norms of an unselected reference group, and estimated what a teacher-nominated comparison group does to the resulting values. The findings are discussed below in the order of the study objectives and in light of the relevant literature.

The measurement structure replicated. Parallel analysis returned two components in each of three tasks that share no content, the pattern replicated across independent halves at Tucker's φ ≥ .991, and the two-dimension model outperformed the one-dimension model. The advantage was clear in Subtests 1 and 3 and marginal in Subtest 2, where a one-dimension solution is also defensible. It is consistently reported that performance-test indicators are not unidimensional [14, 15]. These results add that the same two dimensions recur across paradigms differing in modality and material, which suggests that the

distinction is a property of response organization and not of a particular task. Timing precision, referred to the number of correct responses, is independent of hit rate, so the accuracy dimension of Subtest 1 combines two distinct indicators.

Invariance, developmental validity, and stability were only partly established. Loadings were equivalent in five of six comparisons, while intercept equality failed in Subtest 1 across both groupings and in Subtest 2 across sex. Performance-based child batteries have likewise been reported to reach configural and metric but not strict invariance [15, 51]. Raw scores may therefore not be compared across ages. Age norms are all the more necessary because development is rapid in the early school years, where a difference of one year separates children whose scores differ measurably. Every accuracy indicator and both response control deficit indicators of Subtest 1 changed monotonically and substantially with age, which is consistent with the maturation of executive functions and of sustained, selective, and divided attention across the school years [52, 53]. The response control deficit indicators of Subtests 2 and 3 did not behave in this way. Stability over ten weeks was moderate in Subtest 1, with a minimal detectable change exceeding most of the group differences reported here, and this limit has been documented for computerized tasks in children generally [51, 54]. Two features of the design make this level of stability expected. First, ten weeks is not a short interval for a six- or seven-year-old, in whom the same indicators change measurably within a single year, so part of the gain in accuracy at retest reflects development rather than error. Second, a performance task, unlike a rating scale, is open to retest effects because the child has learned the task at the first administration [55]. This effect is likely to be more pronounced in Subtests 2 and 3, whose rules are discovered during the task, so the Subtest 1 estimate should not be extended to them. The consequence is a boundary on use and not a defect of the instrument: the scores locate a child relative to age peers at one point in time and do not track small individual change.

The main substantive finding concerns which dimension separates recorded status. Against an unselected reference group, response control deficit separated school-recorded status in all three subtests at effect sizes several times larger than those of accuracy, and in two subtests the clinical group captured slightly more targets rather than fewer. This is the pattern the two dimensions were defined to capture: the information lies in how many responses it takes a child to capture the targets rather than in how many targets are captured. Response inhibition is among the domains most consistently impaired in ADHD [18, 19]. Indeed, in a consecutively referred clinical sample the only performance variable separating children who did and did not receive a diagnosis was the number of commission errors [56]. These data add that the pattern replicates across three tasks with unrelated content, which makes a task-specific explanation implausible. The finding should not be read as evidence for a single-deficit account, however, since the accuracy dimension carried independent information about cognitive ability and about teacher ratings, which fits multiple-pathway models more closely [20–22].

The three subtests did not carry the same information, although the pattern was not the one expected. Subtest 1 separated teacher-flagged children from the general population almost as sharply as it separated children with a recorded diagnosis, but it distinguished the two flagged groups only weakly.

Subtests 2 and 3 placed the teacher-flagged children slightly below the reference group while separating them clearly from children with a recorded diagnosis. Two explanations are available. Either these subtests grade severity, so that children whose difficulty a teacher has noticed but no one has evaluated are genuinely closer to the population than to the clinical group on the processes these tasks tax, or the reversal is a selection artifact, a teacher-flagged but never-referred child being one whose difficulty is behavioral rather than performance-based. It is not possible to decide between them here, and the same ambiguity attaches to these children also exceeding the reference group on the accuracy indicator of Subtest 3. The pattern nevertheless establishes that the two functions are carried by different subtests and would mask one another in a total score, which is the concrete form of the argument that a single performance-test score entails a loss of information [11, 12].

The associations with teacher ratings were modest and dimension-specific. Among children without a recorded diagnosis, the Subtest 1 response control deficit composite retained an association with the Conners Total score but not with the inattention subscale. The accuracy indicator of Subtest 3, on the other hand, produced its strongest association precisely there and specifically with inattention rather than with oppositional behavior, and two dimensions showed no reliable association with any Conners score. Weak associations between performance measures and rating scales are the norm in this literature. A review of twenty studies found a median correlation of .19 [57], and later work confirmed that the two load on separate constructs [58, 59]. The finding obtained here is more specific: the indicator that separates recorded status and the one that tracks teacher-perceived inattention among unrecorded children belong to different dimensions of different subtests, so identification and sensitivity to subclinical variation are functions that a single total score would confound.

Selection for cognitive ability affected the two dimensions differently. With teacher nomination held constant, selection toward high cognitive ability was associated with a difference of more than three-quarters of a standard deviation on the accuracy dimension of Subtest 1 and of about a fifth on response control deficit, and the difference between these effects was significant. High ability and attention difficulty co-occur, and the behavioral signal of ADHD is attenuated in high-ability children, so that their difficulties are readily overlooked against average-ability comparison groups [31–33]. This observation is given a measurement form here: the dimension that responds most to ability is not the one that responds to attention status, so for a practitioner deciding whether an able child's reported difficulty reflects ability or attention the two dimensions answer largely separately. The dissociation appears only when teacher nomination is held constant, because nomination preselects on attention as well as on ability.

The fourth objective concerned what a teacher-nominated comparison group does to the estimates. Holding data set, instrument, clinical group, and analyst constant, substitution of the Selected Control Group raised the primary value from .903 to .992 and the apparent specificity of the cut score from .799 to 1.000. The source of this increase is the composition of that group, which lies almost a full standard deviation below the reference group with a dispersion 66% as wide. It has been shown that sampling cases and controls through separate gates narrows the spectrum and inflates measured accuracy [26,

27], and that norm scores from a non-representative sample are systematically biased [25]. The contribution here is that the inflation was measured within the study's own data instead of being estimated from outside, so published accuracy values for instruments of this kind should be read together with the method by which the comparison group was assembled. Deriving the age norms from an unselected reference group converts the problem into a correction: the two-subtest rule flags 4.6% of that group against a population prevalence estimate of 7.6%, but only 0.5% of the teacher-nominated comparison group.

Implications for practice

The findings suggest three principles for psychologists and other professionals who use performance-based attention measures with children.

First, the result should be read as an age-referenced profile of two dimensions in each of three tasks, and not as a single score that averages dimensions answering different questions. A child may stand above the age norm on accuracy and at the same time above it on response control deficit. This combination is informative precisely because the axes are largely separable, and it is what an able child with reported difficulty is likely to produce.

Second, the subtests belong at different points in a multi-stage procedure. Informant report, here teacher judgment, remains the efficient first stage, since multi-stage models identify risk more effectively than single screening or universal retesting [60] and each additional administration carries a cost per child correctly classified [61]. Subtest 1 largely reproduces teacher judgment, since it separated teacher-flagged children from the reference group almost as sharply as it separated children with a recorded diagnosis (AUC = .888 and .903), so it confirms a teacher's flag rather than adding to it. Subtests 2 and 3, by contrast, separated teacher-flagged children from children with a recorded diagnosis (AUC = .757 and .750). For a teacher-flagged child the system therefore answers a second question: whether the difficulty is at the level typically seen in children already in the register.

Third, the rule-out value of the system belongs to the general population and not to children a teacher has already flagged. At a population base rate of 7.6% the at-least-one-subtest rule has a negative predictive value of 98.0% and the Subtest 1 cut score one of 99.0%, and 74.8% of an unselected reference group exceeds no threshold. A child screened without prior suspicion who exceeds none is therefore unlikely to hold a recorded diagnosis. For a teacher-flagged child the prior is higher and these values are upper bounds: 27.8% of the Teacher-Reported Risk Group exceeded no threshold against 18.2% of the Clinical ADHD Group. A positive result is in no case a diagnosis; at the population base rate the positive predictive value rises from 21.1% for at least one threshold to 55.4% for all three (Table 5), and referral to a clinical evaluation is warranted at the higher levels. Finally, that children with a recorded diagnosis capture as many targets as their age peers while failing to withhold responses gives teachers and families a more accurate frame than the interpretation that the child is not listening.

Limitations and future research

The present study has several limitations. First, general cognitive ability was not measured with an individually administered test, so ability differences could be examined only through a group selected toward high ability. Second, the criterion for the Clinical ADHD Group was an administratively recorded diagnosis corroborated by the class teacher and the family rather than a structured interview, and severity and subtype were not recorded; because comorbid cases were excluded, the estimates apply to uncomplicated ADHD presentations. Whether a high response control deficit score is specific to attention disorders therefore remains open, and a study with structured diagnostic interviews and a second clinical group would settle it. For etiological research this is a genuine constraint, but it is also the criterion available in applied settings: a recorded diagnosis confirmed by the two informants who see the child daily, which is the profile of the children who are referred for assessment in practice. Third, the General-Population Reference Group is not clinically verified and is expected to contain children with unidentified difficulties, which makes the estimates conservative rather than unbiased; its residual undetected rate is below the population prevalence, since register-recorded and teacher-flagged children were removed from it, but cannot be quantified here. Fourth, the measurement model was estimated in the full Norm Sample, which includes the labeled samples, whereas the age norms and cut scores were derived from the unselected reference group alone; probability sampling would remove the need for this separation. Fifth, temporal stability was examined for Subtest 1 only, retest effects in Subtests 2 and 3 remain to be quantified, and the minimal detectable change values indicate that the scores are not suited to individual monitoring. Finally, all samples came from one metropolitan region by a non-probability method and the Conners Subsample was small, so the cut scores must be re-derived before use elsewhere.

Conclusions

The Athis Multidimensional Attention Performance Measurement System is a computerized instrument that records, in three tasks of different modality and content, how a child responds to targets and withholds responses to non-targets. The result is expressed as an age-referenced profile of two dimensions per task. In a norm sample of 1,973 children this structure replicated across independent halves and across the three tasks, and the accuracy indicators followed the expected developmental course. Against an unselected reference sample, response control deficit separated children with a recorded ADHD diagnosis in every task, whereas accuracy responded to selection for cognitive ability rather than to attention status. The three subtests assumed distinct roles: Subtest 1 registers the difficulty a teacher notices, whereas Subtests 2 and 3 separate that difficulty from a recorded diagnosis. The subtests are therefore complementary measures that should be read as a profile and not summed into a single score. Because the classification values rest on an unselected reference sample, they are conservative and apply directly to unscreened populations. At the population base rate the system functions as an effective rule-out procedure, with negative predictive values of 98–99%, rather than as an identification instrument. The comparison-group effect quantified here explains why conventional validation designs report higher values for instruments of this kind, and it argues for reporting the sampling route of the comparison group alongside any accuracy estimate.

Abbreviations

ADHD
attention-deficit/hyperactivity disorder
AUC
area under the receiver operating characteristic curve
CFI
comparative fit index
CI
confidence interval
ICC
intraclass correlation coefficient
KMO
Kaiser–Meyer–Olkin measure of sampling adequacy
MoNE
Ministry of National Education
NPV
negative predictive value
PPV
positive predictive value
RMSEA
root mean square error of approximation
SD
standard deviation
SRMR
standardized root mean square residual
STARD
Standards for Reporting of Diagnostic Accuracy Studies
TLI
Tucker–Lewis index

Declarations

Ethics approval and consent to participate

The study protocol was approved by the National Defence University Social and Human Sciences Ethics Committee (approval transmitted by the Turkish Military Academy Deanship, reference number E-35592990-050.04-5496359, dated 27 November 2025). Permission for data collection in schools was granted by the Istanbul Provincial Directorate of National Education. The study was performed in accordance with the Declaration of Helsinki and its later amendments. Written informed consent to

participate was obtained from the parents or legal guardians of all participating children, and verbal assent was obtained from the children before administration.

Consent for publication

Not applicable.

Competing Interests

KÇ designed the Athis Multidimensional Attention Performance Measurement System evaluated in this study and is a founding partner, executive, and shareholder of Özne Yazılım Teknolojileri Sağlık ve Eğitim Hizmetleri A.Ş., the company that developed the system and holds its commercial rights; KÇ may therefore benefit financially from future commercial use of the system. To limit optimistic bias, model selection and model testing were conducted in separate samples; the primary classification accuracy estimates were based on an unselected reference group rather than a recruited comparison group; cut scores were cross-validated and re-derived from the age norms of the unselected reference group; key reported values were independently recomputed by a second analysis routine; and null and unfavorable results are reported in full. The author declares no other competing interests.

Funding

The development of the Athis Multidimensional Attention Performance Measurement System was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the 1812 (BİGG) Program (project number 2240520). The present validation study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The funder had no role in the design of the study, in the collection, analysis, or interpretation of the data, or in writing the manuscript.

Author Contribution

KÇ conceived and designed the study, developed the measurement system, supervised data collection, analysed the data, and wrote the manuscript. The author read and approved the final manuscript.

Acknowledgement

The author thanks the school administrators, class teachers, families, and children who took part in the study, and the psychologists who administered the assessments.

Data Availability

The datasets generated and analysed during the current study are not publicly available because the parental consent obtained did not cover the open deposition of child assessment data, but they are available from the corresponding author on reasonable request. The analysis code written for this study is available from the corresponding author on request.

References

1. Rabiner DL, Godwin J, Dodge KA. Predicting academic achievement and attainment: The contribution of early academic skills, attention difficulties, and social competence. School Psychol Rev. 2016;45(2):250–67. https://doi.org/10.17105/SPR45-2.250-267.
2. Erskine HE, Norman RE, Ferrari AJ, Chan GCK, Copeland WE, Whiteford HA, et al. Long-term outcomes of attention-deficit/hyperactivity disorder and conduct disorder: A systematic review and meta-analysis. J Am Acad Child Adolesc Psychiatry. 2016;55(10):841–50. https://doi.org/10.1016/j.jaac.2016.06.016.
3. Lee Y, Mikami AY, Owens JS. Children's ADHD symptoms and friendship patterns across a school year. Res Child Adolesc Psychopathol. 2021;49(5):643–56. https://doi.org/10.1007/s10802-021-00771-7.
4. Staff AI, van der Oosterlaan J, Hoekstra PJ, Vertessen K, de Vries R, et al. The validity of teacher rating scales for the assessment of ADHD symptoms in the classroom: A systematic review and meta-analysis. J Atten Disord. 2021;25(11):1578–93. https://doi.org/10.1177/1087054720916839.
5. n der Embse N, Kim ES, Kilgus S, Dedrick R, Sanchez A. Multi-informant universal screening: Evaluation of rater, item, and construct variance using a trifactor model. J Sch Psychol. 2019;77:52–66. https://doi.org/10.1016/j.jsp.2019.09.005.
6. De Los Reyes A, Augenstein TM, Wang M, Thomas SA, Drabick DAG, Burgers DE, et al. The validity of the multi-informant approach to assessing child and adolescent mental health. Psychol Bull. 2015;141(4):858–900. https://doi.org/10.1037/a0038498.
7. De Los Reyes A, Cook CR, Gresham FM, Makol BA, Wang M. Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. J Sch Psychol. 2019;74:74–89. https://doi.org/10.1016/j.jsp.2019.05.005.
8. Narad ME, Garner AA, Peugh JL, Tamm L, Antonini TN, Kingery KM, et al. Parent–teacher agreement on ADHD symptoms across development. Psychol Assess. 2015;27(1):239–48. https://doi.org/10.1037/a0037864.
9. Lockwood AB, Farmer RL, Krach SK. Examining school psychologists' attitudes toward standardized assessment tools. J Psychoeducational Assess. 2022;40(3):311–26. https://doi.org/10.1177/07342829211057642.
10. Sullivan AL, Sadeh S, Houri AK. Are school psychologists' special education eligibility decisions reliable and unbiased? A multi-study experimental investigation. J Sch Psychol. 2019;77:90–109. https://doi.org/10.1016/j.jsp.2019.10.006.

11. Arrondo G, Mulraney M, Iturmendi-Sabater I, Musullulu H, Gambra L, Niculcea T, et al. Systematic review and meta-analysis: Clinical utility of continuous performance tests for the identification of attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry. 2024;63(2):154–71. https://doi.org/10.1016/j.jaac.2023.03.011.
12. Bellato A, Hall CL, Groom MJ, Simonoff E, Thapar A, Hollis C, et al. Practitioner Review: Clinical utility of the QbTest for the assessment and diagnosis of attention-deficit/hyperactivity disorder — A systematic review and meta-analysis. J Child Psychol Psychiatry. 2024;65(6):845–61. https://doi.org/10.1111/jcpp.13901.
13. Hall CL, Valentine AZ, Groom MJ, Walker GM, Sayal K, Daley D, et al. The clinical utility of the continuous performance test and objective measures of activity for diagnosing and monitoring ADHD in children: A systematic review. Eur Child Adolesc Psychiatry. 2016;25(7):677–99. https://doi.org/10.1007/s00787-015-0798-x.
14. Lund O, Raudeberg R, Johansen H, Myhre ML, Walderhaug E, Poreh A, et al. Factor structure of the Conners Continuous Performance Test Third Edition (CCPT-3): Exploratory factor analysis in a mixed clinical sample. J Atten Disord. 2025;29(13):1163–76. https://doi.org/10.1177/10870547251341928.
15. Clark CAC, Cook K, Wang R, Rueschman M, Radcliffe J, Redline S, et al. Psychometric properties of a combined go/no-go and continuous performance task across childhood. Psychol Assess. 2023;35(4):353–65. https://doi.org/10.1037/pas0001202.
16. Egeland J, Kovalik-Gran I. Measuring several aspects of attention in one test: The factor structure of Conners's Continuous Performance Test. J Atten Disord. 2010;13(4):339–46. https://doi.org/10.1177/1087054708323019.
17. Barkley RA. Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. Psychol Bull. 1997;121(1):65–94. https://doi.org/10.1037/0033-2909.121.1.65.
18. Lipszyc J, Schachar R. Inhibitory control and psychopathology: A meta-analysis of studies using the stop signal task. J Int Neuropsychol Soc. 2010;16(6):1064–76. https://doi.org/10.1017/S1355617710000895.
19. Pievsky MA, McGrath RE. The neurocognitive profile of attention-deficit/hyperactivity disorder: A review of meta-analyses. Arch Clin Neuropsychol. 2018;33(2):143–57. https://doi.org/10.1093/arclin/acx055.
20. Kofler MJ, Irwin LN, Soto EF, Groves NB, Harmon SL, Sarver DE. Executive functioning heterogeneity in pediatric ADHD. J Abnorm Child Psychol. 2019;47(2):273–86. https://doi.org/10.1007/s10802-018-0438-2.
21. Sonuga-Barke EJS. Causal models of attention-deficit/hyperactivity disorder: From common simple deficits to multiple developmental pathways. Biol Psychiatry. 2005;57(11):1231–8. https://doi.org/10.1016/j.biopsych.2004.09.008.

22. van Hulst BM, de Zeeuw P, Durston S. Distinct neuropsychological profiles within ADHD: A latent class analysis of cognitive control, reward sensitivity and timing. Psychol Med. 2015;45(4):735–45. https://doi.org/10.1017/S0033291714001792.
23. Arildskov TW, Thomsen PH, Sonuga-Barke EJS, Lambek R, Østergaard SD, Virring A. Is attention-deficit/hyperactivity disorder (ADHD) a dimension or a category? What does the relationship between ADHD traits and psychosocial quality of life tell us? J Atten Disord. 2024;28(7):1035–44. https://doi.org/10.1177/10870547231222228.
24. Haslam N, McGrath MJ, Viechtbauer W, Kuppens P. Dimensions over categories: A meta-analysis of taxometric research. Psychol Med. 2020;50(9):1418–32. https://doi.org/10.1017/S003329172000183X.
25. Gary S, Lenhard A, Lenhard W, Herzberg DS. Reducing the bias of norm scores in non-representative samples: Weighting as an adjunct to continuous norming methods. Assessment. 2023;30(8):2491–509. https://doi.org/10.1177/10731911231153832.
26. Rutjes AWS, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PMM. Case–control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51(8):1335–41. https://doi.org/10.1373/clinchem.2005.048595.
27. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
28. Edwards AA, Van Dijk W, White CM, Schatschneider C. Screening screeners: Calculating classification indices using correlations and cut-points. Ann Dyslexia. 2022;72(3):445–60. https://doi.org/10.1007/s11881-022-00261-5.
29. Klingbeil DA, Latham AD, Kim JS, Schmitt MC. Base rates of non-proficiency in reading and math: Implications for universal screening. Psychol Sch. 2025;62(3):721–33. https://doi.org/10.1002/pits.23350.
30. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. https://doi.org/10.1136/bmj.h5527.
31. Cadenas M, Hartman C, Faraone S, Antshel K, Borges Á, Hoogeveen L, et al. Cognitive correlates of attention-deficit hyperactivity disorder in children and adolescents with high intellectual ability. J Neurodevelopmental Disorders. 2020;12:6. https://doi.org/10.1186/s11689-020-9307-8.
32. Cornoldi C, Giofrè D, Toffalini E. Cognitive characteristics of intellectually gifted children with a diagnosis of ADHD. Intelligence. 2023;97:101736. https://doi.org/10.1016/j.intell.2023.101736.
33. Gomez R, Stavropoulos V, Vance A, Griffiths MD. Gifted children with ADHD: How are they different from non-gifted children with ADHD? Int J Mental Health Addict. 2020;18(6):1467–81. https://doi.org/10.1007/s11469-019-00125-x.
34. Rommelse N, Antshel K, Smeets S, Greven C, Hoogeveen L, Faraone SV, et al. High intelligence and the risk of ADHD and other psychopathology. Br J Psychiatry. 2017;211(6):359–64.

https://doi.org/10.1192/bjp.bp.116.184382.

35. Faraone SV, Banaschewski T, Coghill D, Zheng Y, Biederman J, Bellgrove MA, et al. The World Federation of ADHD International Consensus Statement: 208 evidence-based conclusions about the disorder. Neurosci Biobehavioral Reviews. 2021;128:789–818. https://doi.org/10.1016/j.neubiorev.2021.01.022.
36. Conners CK. A teacher rating scale for use in drug studies with children. Am J Psychiatry. 1969;126(6):884–8. https://doi.org/10.1176/ajp.126.6.884.
37. Kaner S, Büyüköztürk Ş, İşeri E. Conners öğretmen derecelemeölçeği-yenilenmiş kısa: Türkiye uyarlama çalışması [Conners Teacher Rating Scale–Revised Short: Turkish adaptation study]. Eğitim ve Bilim. 2013;38(167):81–97. https://doi.org/10.15390/ES.2013.1104.
38. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;30(2):179–85. https://doi.org/10.1007/BF02289447.
39. Kenny DA, Kaniskan B, McCoach DB. The performance of RMSEA in models with small degrees of freedom. Sociol Methods Res. 2015;44(3):486–507. https://doi.org/10.1177/0049124114543236.
40. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334. https://doi.org/10.1007/BF02310555.
41. McNeish D. Thanks coefficient alpha, we'll take it from here. Psychol Methods. 2018;23(3):412–33. https://doi.org/10.1037/met0000144.
42. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equation Modeling: Multidisciplinary J. 2007;14(3):464–504. https://doi.org/10.1080/10705510701301834.
43. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equation Modeling: Multidisciplinary J. 2002;9(2):233–55. https://doi.org/10.1207/S15328007SEM0902_5.
44. Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Dev Rev. 2016;41:71–90. https://doi.org/10.1016/j.dr.2016.06.004.
45. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
46. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988;44(3):837–45. https://doi.org/10.2307/2531595.
47. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–5. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3.
48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

49. Steiger JH. Tests for comparing elements of a correlation matrix. Psychol Bull. 1980;87(2):245–51. https://doi.org/10.1037/0033-2909.87.2.245.
50. Salari N, Ghasemi H, Abdoli N, Rahmani A, Shiri MH, Hashemian AH, et al. The global prevalence of ADHD in children and adolescents: A systematic review and meta-analysis. Ital J Pediatr. 2023;49(1):48. https://doi.org/10.1186/s13052-023-01456-1.
51. Karalunas SL, Bierman KL, Huang-Pollock CL. Test–retest reliability and measurement invariance of executive function tasks in young children with and without ADHD. J Atten Disord. 2020;24(13):1891–904. https://doi.org/10.1177/1087054715627488.
52. Best JR, Miller PH. A developmental perspective on executive function. Child Dev. 2010;81(6):1641–60. https://doi.org/10.1111/j.1467-8624.2010.01499.x.
53. Talalay IV. The development of sustained, selective, and divided attention in school-age children. Psychol Sch. 2024;61(6):2223–39. https://doi.org/10.1002/pits.23160.
54. Syväoja HJ, Tammelin TH, Ahonen T, Räsänen P, Tolvanen A, Kankaanpää A, et al. Internal consistency and stability of the CANTAB neuropsychological test battery in children. Psychol Assess. 2015;27(2):698–709. https://doi.org/10.1037/a0038485.
55. Scharfen J, Peters JM, Holling H. Retest effects in cognitive ability tests: A meta-analysis. Intelligence. 2018;67:44–66. https://doi.org/10.1016/j.intell.2018.01.003.
56. Johnsen KH, Kvitland LR, Sollie H, Reiestad M, Jonsbu E, Hagen K, et al. Is the Conners' continuous performance test helpful for assessing attention deficit hyperactivity disorder in a clinical setting? Nord J Psychiatry. 2024;78(2):120–7. https://doi.org/10.1080/08039488.2023.2279640.
57. Toplak ME, West RF, Stanovich KE. Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? J Child Psychol Psychiatry. 2013;54(2):131–43. https://doi.org/10.1111/jcpp.12001.
58. Soto EF, Kofler MJ, Singh LJ, Wells EL, Irwin LN, Groves NB, et al. Executive functioning rating scales: Ecologically valid or construct invalid? Neuropsychology. 2020;34(6):605–19. https://doi.org/10.1037/neu0000681.
59. Ten Eycke KD, Dewey D. Parent-report and performance-based measures of executive function assess different constructs. Child Neuropsychol. 2016;22(8):889–906. https://doi.org/10.1080/09297049.2015.1065961.
60. VanMeveren K, Hulac D, Wollersheim-Shervey S. Universal screening methods and models: Diagnostic accuracy of reading assessments. Assess Effective Intervention. 2020;45(4):255–65. https://doi.org/10.1177/1534508418819797.
61. Paly BJ, Klingbeil DA, Clemens NH, Osman DJ. A cost-effectiveness analysis of four approaches to universal screening for reading risk in upper elementary and middle school. J Sch Psychol. 2022;92:246–64. https://doi.org/10.1016/j.jsp.2022.03.009.

Figures

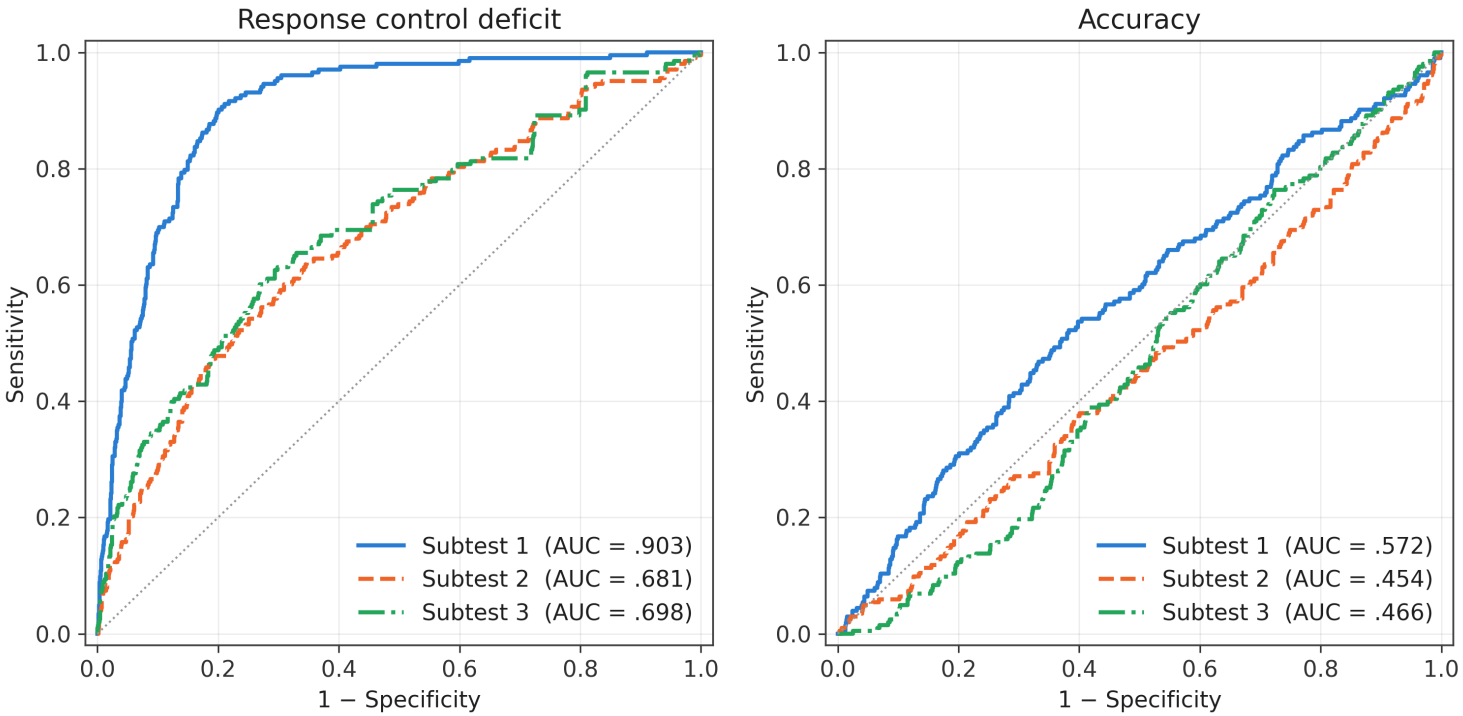


Figure 1

Receiver operating characteristic curves for the Clinical ADHD Group against the General-Population Reference Group. Left panel: response control deficit composites. Right panel: accuracy indicators. The diagonal marks chance performance.

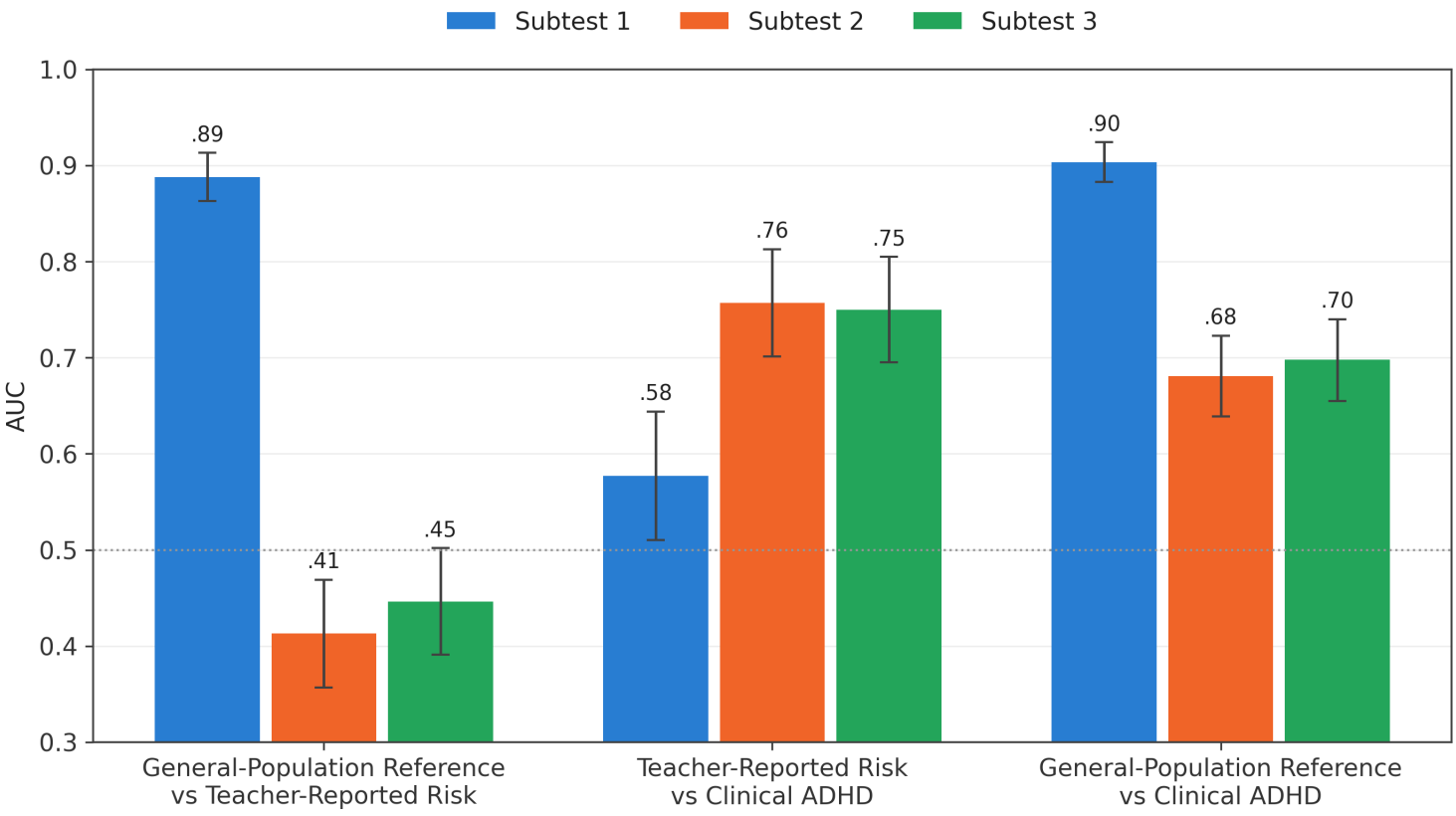


Figure 2

Classification accuracy of the response control deficit composites across three contrasts. Error bars are 95% DeLong confidence intervals. The dotted line marks chance performance.

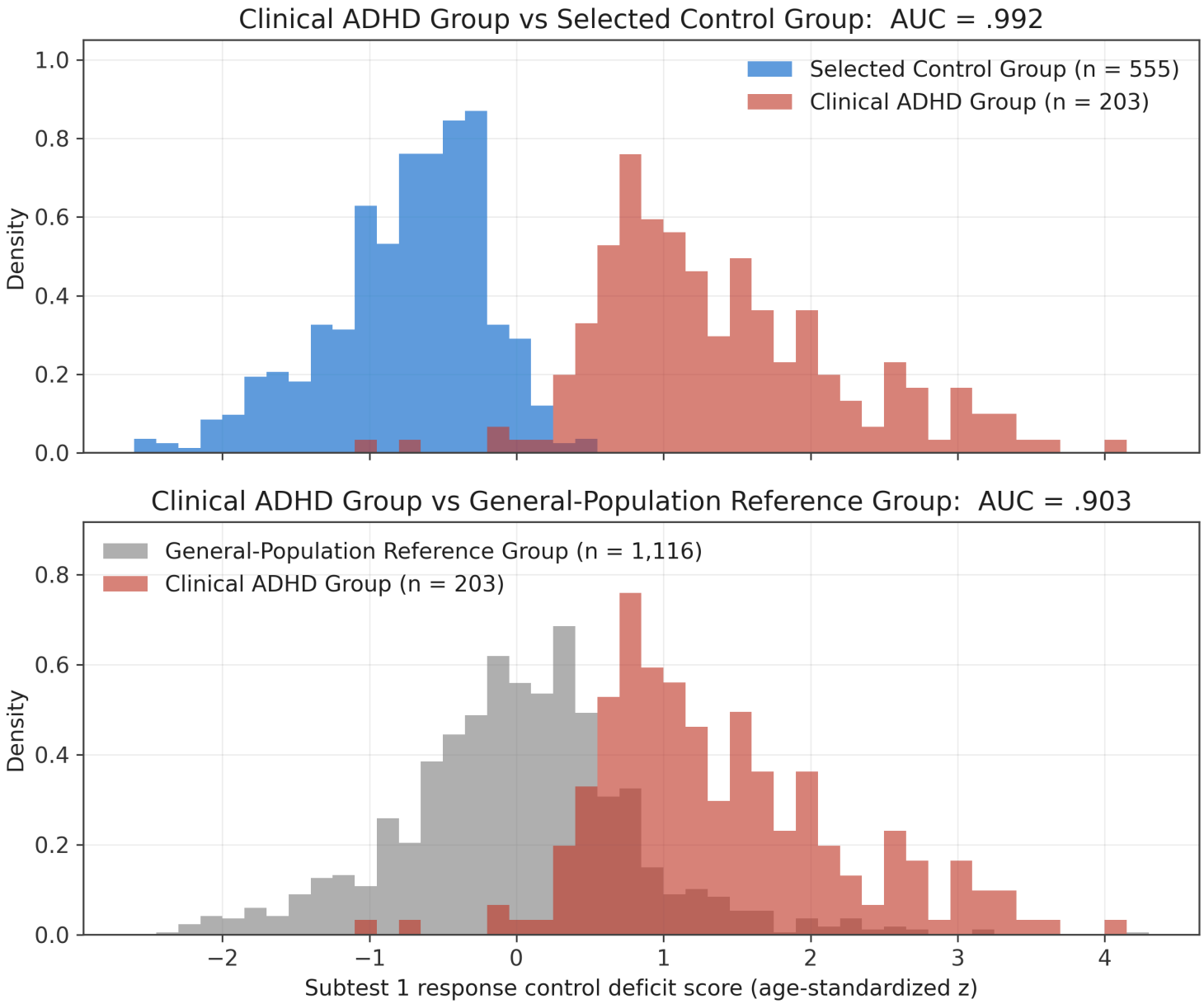


Figure 3

Distribution of the Subtest 1 response control deficit score against two comparison groups. Upper panel: the Clinical ADHD Group against the teacher-nominated Selected Control Group. Lower panel: the same Clinical ADHD Group against the unselected General-Population Reference Group.

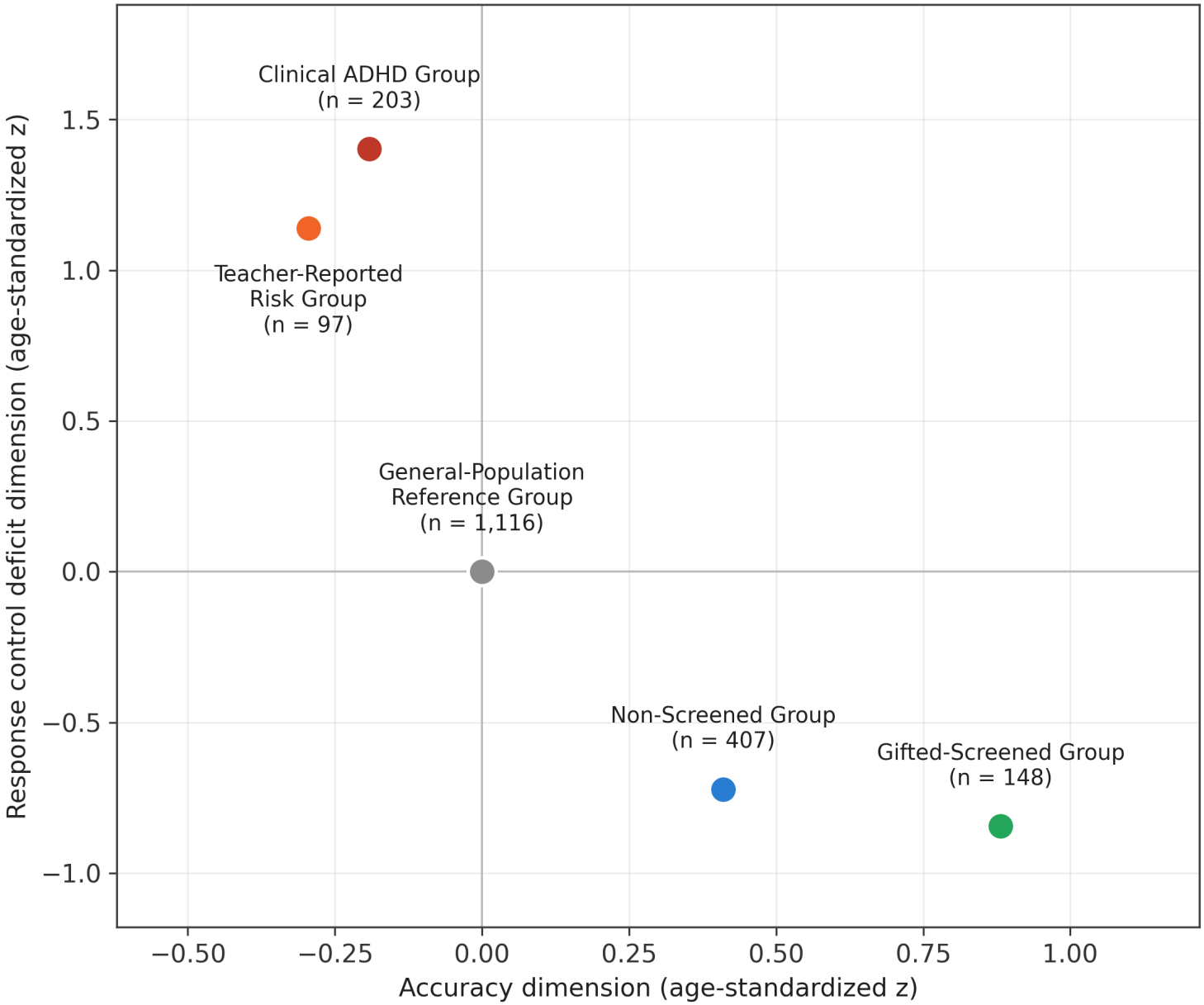


Figure 4

Group means on the two dimensions of Subtest 1. Values are means of age-standardized *z* scores. Higher response control deficit scores indicate poorer response control.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AthisBMCAdditionalfile1.docx
- AthisBMCAdditionalfile2.docx