Systematic Review and Meta-analysis: Clinical Utility of Continuous Performance Tests for Identification of ADHD

**Running title:** Meta-analysis: utility of CPTs in ADHD

**Objective:** We aimed to quantify the clinical utility of continuous performance tests (CPTs) for the diagnosis of ADHD compared to a clinical diagnosis in children and adolescents. **Method:** Four databases (MEDLINE, PsycINFO, EMBASE, and PubMed) were screened until January 2023. Risk of bias of included results was judged with the QUADAS-2. We statistically pooled the area under the curve, the sensitivity, and the specificity of three commonly used CPTs subscales: omission/inattention, commission/impulsivity, and total number of errors/ADHD subscales (PROSPERO registration: CRD42020168091). **Results:** 19 studies using commercially available CPTs were identified. Results from up to 835 control individuals and 819 cases were combined in the summary receiver operating characteristic (ROC) curve analyses (sensitivity and specificity pooling), and up to 996 cases and 1083 control individuals in the area under the curve (AUC) analyses. Clinical utility as measured by AUCs could be considered as barely acceptable (between 0.7 and 0.8) for the most part, with the best results for the total/ADHD score, followed by omissions/inattention, and poorest for commission/impulsivity scores. A similar pattern was found when pooling sensitivity and specificity: 0.75 (95% CI=0.66 to 0.82) and 0.71 (0.62 to 0.78) for the total/ADHD score; 0.63 (=0.49 to 0.75) and 0.74 (0. 65 to .81) for omissions; and 0.59 (0.38 to 0.77) and .66 (CI=.50, .78) for commissions. **Conclusion:** At the clinical level, CPTs as a stand-alone tool have only a modest to moderate ability to differentiate ADHD from non-ADHD samples. Hence, they should only be used within a more comprehensive diagnostic process.

Gonzalo Arrondo PhD, Melissa Mulraney PhD, Iciar Iturmendi-Sabater MPsych, Hande Musullulu MPsych, Leyre Gambra PhD, Teodora Niculcea MPsych, Tobias Banaschewski MD, PhD, Emily Simonoff, MD, Manfred Doepfner PhD, Stephen P Hinshaw PhD, Dave Coghill MD, Samuele Cortese MD PhD

Drs. Arrondo and Mulraney shared first authorship of this article and contributed equally.

Drs Coghill and Cortese shared authorship of this article and contributed equally

Dr Arrondo, Dr Gambra, Musullulu, and Niculcea are with the University of Navarra, Pamplona, Spain. Dr. Mulraney and Prof. Coghill are with the University of Melbourne, Australia, and Murdoch Children’s Research Institute, Melbourne. Dr. Mulraney is also with the Institute for Social Neuroscience, Ivanhoe, Australia. Iturmendi-Sabater is with the University of Toronto, Canada. Prof. Cortese is with the University of Southampton, Southampton, United Kingdom; Solent NHS Trust, Southampton, United Kingdom; New York University, New York, and the University of Nottingham, United Kingdom. Prof. Banaschewski is with University of Heidelberg, Mannheim, Germany. Prof Simonoff is with King’s College, London, United Kingdom. Prof. Doepfner is with the University of Cologne, Germany and King’s College London, United Kingdom. Prof. Hinshaw is with the University of California, Berkeley, and the University of California, San Francisco, USA.

**Correspondence to:** Gonzalo Arrondo; Mind-Brain group, Institute for Culture and Society (ICS), Edificio Ismael Sánchez Bella, Campus Universitario S/N, Pamplona, 31009, Navarra, Spain; phone: +34 948 42 56 00 ext. 805615; email: [garrondo@unav.es](about:blank)

**ORCID**

Gonzalo Arrondo, PhD: https://orcid.org/0000-0003-3085-8959

Melissa Mulraney, PhD: https://orcid.org/0000-0003-1953-6481

Iciar Iturmendi-Sabater, MPsych: https://orcid.org/0000-0002-7239-3435

Hande Musullulu, MPsych: https://orcid.org/0000-0002-1658-0199

Leyre Gambra, PhD: https://orcid.org/0000-0002-7422-6294

Tobias Banaschewski, MD, PhD: [https://orcid.org/0000-0003-4595-1144](about:blank)

Emily Simonoff, MD https://orcid.org/0000-0002-5450-0823

Manfred Doepfner, PhD, https://orcid.org/0000-0002-7929-0463

Stephen Hinshaw, PhD: [https://orcid.org/0000-0001-6497-1082](about:blank)

David Coghill, MD: https://orcid.org/0000-0003-3017-9737

Samuele Cortese, MD, PhD: https://orcid.org/0000-0001-5877-8075

# Acknowledgments, and Conflicts of Interest/Disclosure

Dr. Arrondo is supported by Grant RYC2020‐030744‐I funded by MCIN/AEI/ 10.13039/501100011033 and by “ESF Investing in your future”; he is also supported by the 2022-2023 Institute for Culture and Society (ICS) challenge on "Youth, relationships and psychological well-being" of the University of Navarra, and has received funding from the Spanish Ministry of Science, Innovation and Universities to facilitate the mobility of researchers to foreign higher education and research centers (Ref. CAS19/00249).

Dr. Mulraney has received consulting income and research funds from the International Consortium for Health Outcomes Measurement, The Royal Children’s Hospital Foundation, and ISN Innovations.

Iturmendi-Sabater receives the support of a fellowship from “La Caixa” foundation (ID100010434) to pursue postgraduate studies. The fellowship code is LCF/BQ/EU21/11890074.

Prof. Hinshaw has received royalties from Oxford University Press and St. Martin’s Press and has received research support from the National Institute of Mental Health.

Prof. Banaschewski served in an advisory or consultancy role for eye level, Infectopharm, Lundbeck, Medice, Neurim Pharmaceuticals, Oberberg GmbH, Roche, and Takeda. He received conference support or speaker’s fee by Janssen, Medice and Takeda. He received royalities from Hogrefe, Kohlhammer, CIP Medien, Oxford University Press.

Prof. Doepfner received consulting income and research support from Lilly, Medice, Takeda, and eyelevel GmbH. He received income as head, supervisor, and lecturer of the School of Child and Adolescent Cognitive Behaviour Therapy at the University Hospital Cologne and as consultant for Child Behaviour Therapy at the National Association of Statutory Health Insurance Physicians (Kassenärztliche Bundesvereinigung). He also received royalties from treatment manuals, books and psychological tests published by Beltz, Elsevier, Enke, Guilford, Hogrefe, Huber, Kohlhammer, Schattauer, Springer, Wiley.

Professor Simonoff receives funding from the National Institute of Health Research, UK, UKRI, EU Innovative Medicines Initiative and the Maudsley Charity. Prof. Coghill has received research support and/or honoraria from Shire/Takeda, Medice, Novartis, and Servier and royalties from Oxford University Press and Cambridge University Press.

Prof Cortese is supported by NIHR awards NIHR203684, NIHR130077, RP-PG-0618-20003, NIHR203035

Musullulu, Gambra and Niculcea have reported no biomedical financial interests or potential conflicts of interest.

None of the named funding sources had any involvement in the study design; the collection, analysis and interpretation of data; writing of the report; and the decision to submit the article for publication.

**Author contributions**

Conceptualization: Arrondo, Mulraney, Banaschewski, Simonoff, Doepfner, Hinshaw Coghill, Cortese

Data curation: Arrondo, Mulraney

Formal Analysis: Arrondo, Mulraney

Methodology: Arrondo, Mulraney, Coghill, Cortese

Project administration: Arrondo, Mulraney, Coghill, Cortese

Investigation: Arrondo, Mulraney, Iturmendi-Sabater, Musullulu, Gambra, Niculcea

Software: Arrondo, Mulraney

Supervision: Arrondo, Mulraney, Coghill, Cortese

Validation: Arrondo, Mulraney

Visualization: Arrondo, Mulraney, Niculcea

Writing – original draft: Arrondo, Mulraney

Writing – review & editing: Arrondo, Mulraney, Iturmendi-Sabater, Musullulu, Gambra, Niculcea, Banaschewski, Simonoff, Doepfner, Hinshaw, Coghill, Cortese

**Keywords**

Attention Deficit Disorder with Hyperactivity; Neuropsychological Tests, Sensitivity and Specificity, diagnosis

# Abstract

**Objective:** We aimed to quantify the clinical utility of continuous performance tests (CPTs) for the diagnosis of ADHD compared to a clinical diagnosis in children and adolescents. **Method:** Four databases (MEDLINE, PsycINFO, EMBASE, and PubMed) were screened until January 2023. Risk of bias of included results was judged with the QUADAS-2. We statistically pooled the area under the curve, the sensitivity, and the specificity of three commonly used CPTs subscales: omission/inattention, commission/impulsivity, and total number of errors/ADHD subscales (PROSPERO registration: CRD42020168091). **Results:** 19 studies using commercially available CPTs were identified. Results from up to 835 control individuals and 819 cases were combined in the summary receiver operating characteristic (ROC) curve analyses (sensitivity and specificity pooling), and up to 996 cases and 1083 control individuals in the area under the curve (AUC) analyses. Clinical utility as measured by AUCs could be considered as barely acceptable (between 0.7 and 0.8) for the most part, with the best results for the total/ADHD score, followed by omissions/inattention, and poorest for commission/impulsivity scores. A similar pattern was found when pooling sensitivity and specificity: 0.75 (95% CI=0.66 to 0.82) and 0.71 (0.62 to 0.78) for the total/ADHD score; 0.63 (=0.49 to 0.75) and 0.74 (0. 65 to .81) for omissions; and 0.59 (0.38 to 0.77) and .66 (CI=.50, .78) for commissions. **Conclusion:** At the clinical level, CPTs as a stand-alone tool have only a modest to moderate ability to differentiate ADHD from non-ADHD samples. Hence, they should only be used within a more comprehensive diagnostic process.

# Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a common yet impairing neurodevelopmental disorder, with a worldwide prevalence in children and adolescents close to 5%.1 ADHD is characterized by inattention and/or impulsivity and high motor activity (hyperactivity). Diagnosis should be made by a trained clinician and is based on the identification of the core symptoms and an evaluation of the extent to which these hinder daily functioning across different contexts and are not better accounted for by other disorders. Diagnosis can be aided by the use of psychometric scales, measuring symptom severity. In adults these are typically completed by the individual being assessed, whereas for children and adolescents parents and/or teachers are the key informants, sometimes supplemented with child/adolescent self-ratings.2 Additionally, cognitive tests of attention and executive control are often used under the assumption that they are more objective markers of the disorder that can guide identification and diagnosis.3 Among these cognitive tests, continuous performance tasks (CPTs) are the most popular.4 The CPT is a cognitive paradigm that aims to measure selective attention to infrequently occurring stimuli over an extended period of time. CPTs were first developed in 1956 and have been used clinically for decades.5 Two main CPT measures have been theoretically and empirically linked to the core characteristics of ADHD: omission errors (i.e., the inability to respond to a target), which is considered a metric of inattentiveness; and commission errors (i.e., the failure to inhibit responses to inappropriate stimuli), which is conceptualized as an estimation of impulsivity.4 While CPTs have never been recommended as a stand-alone measure in clinical practice, the appeal of the test as a supposedly objective, technologically advanced, and cheap tool has led to some health providers to over rely on them in clinical practice. A systematic review of the clinical usefulness of CPTs against clinical diagnoses could put into perspective their real-world properties and applications.

Importantly, the CPT is not a single test, but an overarching category that includes several academic and commercial implementations. These specific implementations differ in the stimuli used, the proportion of targets versus non-targets, and the speed of stimulus presentation. Commonly used commercial versions include the Conners´ CPT (CCPT),6 the Integrated Visual and Auditory (IVA+) CPT,7 and the Test of Variables of Attention (TOVA).8

Previous meta-analyses have demonstrated that, on average, children with ADHD obtain lower scores on CPTs than typically developing controls by comparing the standardized mean difference between these groups. In their 2012 systematic review and meta-analysis,9 Huang-Pollock and colleagues concluded that children with ADHD committed significantly more omission and commission errors than controls, with the between-group effects being characterized as large. The authors of this meta-analysis attempted to correct for measurement error using reliability coefficients from published sources, which greatly changed the estimation of the effects. In fact, without this correction, effect sizes were much lower.

Virtual reality classroom CPTs have become popular in recent years, given the advances and affordability of this type of technology.10 Recently, Parsons et al. compared virtual and traditional CPTs for the identification of ADHD.11 Importantly, they included in their quantitative pooling only studies comparing individuals with ADHD to controls for a virtual CPT measure or for both a virtual and traditional CPT (i.e., they did not include studies that only used traditional CPTs). Their results indicated that participants with ADHD had significantly more omission and commission errors, with similar effect-sizes recorded for both virtual and traditional CPTs.

Despite these findings, it is difficult to appreciate the usefulness of CPTs in the clinical evaluation process, given that pooled data represented through standardized mean differences do not indicate the proportions of individuals correctly and incorrectly identified. The diagnostic utility of a test is better depicted through measures of clinical accuracy, such as the sensitivity and specificity at different thresholds, which allow the development of receiver operating characteristic curves. This information allows for the creation of cut-off scores indicating to clinicians that a given above-threshold patient is likely to have ADHD. Such procedures are particularly important for neuropsychological measures in the field of ADHD, given the extensive literature demonstrating considerable neuropsychological heterogeneity in this disorder.12

A recent systematic review and meta-analysis on the clinical utility of psychometric rating scales in the evaluation of ADHD in children and adolescents showed that none of the included rating scales had sufficient evidence of clinical utility to be used as a stand-alone tool.13 To the best of our knowledge, the only previous systematic review evaluating the clinical utility of CPTs for ADHD found mixed evidence on the utility of commercial CPTs when predicting group membership—but did not conduct any statistical pooling of the results.14 Moreover, that systematic review was conducted in 2015, since then there have been several additional studies.15–17

Our primary aim was to investigate the clinical utility of continuous performance tests, by conducting a systematic review and a meta-analysis quantitatively pooling measures of diagnostic efficacy (i.e., AUC, and sensitivity and specificity) of CPTs as stand-alone tools for the identification of ADHD in children/adolescents. A secondary aim was to compare the clinical usefulness of various commercial implementations of the CPT.

# Methods

This systematic review and meta-analysis was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement.18 The study protocol was registered in PROSPERO before the commencement of the screening process (PROSPERO registration: [CRD42020168091](https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020168091))**Inclusion criteria**

We included peer-reviewed studies, recruiting participants between the age of three and eighteen years, and comparing the diagnostic accuracy of a CPT as stand-alone tool to a reference standard. Accepted measures of accuracy were the sensitivity and specificity (or raw numbers from the classification between individuals with ADHD and without ADHD that permitted their calculation) and/or the AUC. Studies had to report a measure of accuracy for an index of: omissions and/or commissions and/or total number of errors—or, alternatively, combination of scores proposed by the authors of the test to identify inattention, impulsivity, ADHD in general, or overall performance in the task. Therefore, studies combining indexes post-hoc (e.g., using discriminant function analysis) were excluded. The accepted reference standards were as follows: a clinical diagnosis with evidence from parent interview, child observation, and independent evidence of pervasiveness (e.g., school reports); a research diagnosis with parent interview; a clinical diagnosis based on codes (ICD/DSM) in medical records/registries; or clinical diagnosis methods not otherwise specified. Hence, studies in which authors indicated that classification of ADHD was solely based on rating scales were excluded. We included studies regardless of study settings (general population or clinical settings) and design (longitudinal, cohort, or case-control studies). Any type of CPT was accepted, including virtual reality ones. Any temporal difference was accepted between the clinical diagnosis and the utilization of the CPT.

**Information source and search strategy**

Searches were conducted in MEDLINE, PsycINFO, EMBASE, and PubMed in three stages. The last search was conducted on January 27th 2023. The full electronic search strategy for each database and search is provided in Table S1 and Table S2, available online. There were no date or language restrictions. Reference lists of potentially relevant studies and related systematic reviews and meta-analyses were also manually searched to detect any possible references missed with our search in electronic databases. The preparation of the search was carried out with the assistance from librarians at the Royal Children’s Hospital, Melbourne.**Study selection and data extraction**

Screening of titles and abstracts was conducted independently by two authors. Full texts of potentially relevant studies were obtained and judged by two authors to determine final inclusion. The data were independently extracted by two authors for all studies using a standardized, pilot-tested extraction sheet. Discrepancies were discussed and resolved by consensus, with a third author acting as an arbitrator in cases of doubt. Extraction fields were basic descriptive study information (e.g., year of publication, sample size, sample type/setting, sample age, sample gender, conflicts of interest declared by authors), CPT implementation, subscales and thresholds, type of diagnosis, and statistical and methodological considerations, including the AUC and data needed to calculate indices of diagnostic accuracy. For articles written in a language other than English, an internal or external collaborator fluent in that language was asked to review and extract the pertinent data.

**Risk of bias in individual studies**

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool was chosen to assess the risk of bias across of included studies across four different domains: case selection, index test, reference standard, flow and timing.19 Risk of bias evaluation was performed independently by two study authors; each included study was rated within each domain as having a low, high, or unclear risk of bias. Table S3, available online, shows the signaling questions used in the current systematic review and meta-analysis to indicate a potential increase in the risk of bias.

**Data analysis**

Commercial versions of CPTs typically provide measures of errors of omission and commission for the identification of inattention and impulsivity, respectively, or some sort of proprietary statistic related to errors of omission and commission. Additionally, the total number of errors or some other proprietary combination of measures can be used as a general measure of ADHD. We aimed to pool two types of accuracy indices (AUC; sensitivity and specificity) of inattention/omissions, impulsivity/commissions, and total number of errors/ADHD/overall performance, yielding six main analyses. The AUC is derived from a receiver operating characteristic (ROC) analysis, whereby the true positives are plotted against the false positives for each cut-off point. The AUC can range from 0 (perfectly inaccurate) to 1 (perfectly accurate), with an AUC of 0.5 indicating the tool performs at chance levels. An AUC of 0.7-0.8 is considered acceptable, 0.8-0.9 excellent, and above 0.9 outstanding.20 Sensitivity and specificity relate to the ability of a specific cut-score on a tool to distinguish between cases and non-cases. Typically, a sensitivity between 0.7 and 0.8 (i.e., 80% of the true cases in the population are identified when the sensitivity is 0.8) is considered acceptable in psychiatric settings, with specificity rates as close to 0.8 or higher (i.e., 80% of non-cases correctly identified) used for selection of the optimal cut-off score.21

***Selection of predictor scores***

In instances where multiple samples and multiple predictor scores were reported in a single study, we followed a set of decision rules. We selected the predictor that best represented inattention/omissions, impulsivity/commissions, and total score/ADHD when additional outcomes were presented.

If rate of omissions, commissions or total error rate was reported we used them. Alternatively, if there was a score named “overall score” or “total score” “ADHD score” or similar names, they were selected to represent the CPT Total Score. Alternatives for omission rates were scores named “inattentiveness” or “distraction scores”. Alternatives for commission rates were scores named “impulsivity” or “hyperactivity” If multiple thresholds were used, we selected the one with a sensitivity closest to 0.8 and considered it our main predictor. Additionally, we also extracted the predictor with a specificity closest to 0.8 for a sensitivity analysis. If sensitivity and specificity could not be obtained numerically, but the study included ROC curves, we extracted data using the online tool WebPlotDigitizer ([https://apps.automeris.io/wpd/](about:blank)). If multiple independent samples were included (e.g., male and female samples reported separately) we extracted them separately but combined them before statistical pooling. Group sizes were summed to obtain overall sensitivity and specificity whenever possible, otherwise a fixed-effect analysis was carried out to obtain a single measure per study that could be meta-analyzed with the other studies.

***Meta-analysis of the Area Under the Curve (AUC)***

The AUC was meta-analyzed for our three main CPT predictor measures (omissions/inattention score, commissions/impulsivity-hyperactivity score, total errors/ADHD score) using the ‘meta’ package version 5.2-0 in R-4.2.2. We used the ‘metagen’ command to conduct a generic inverse variance meta-analysis with a random-effects model where the AUC and standard error (SE) of AUC were available.22 We estimated sensitivity following the standard normal distribution whenever they were not reported for a study.23

We also carried out two sensitivity analyses: (a) excluding any study with a high risk of bias in any of the domains of the QUADAS-2; and (b) limiting the analyses to studies on clinical or high-risk samples, as opposed to studies including community samples. AUCs were also pooled for specific CPT versions whenever more than one study reported on the same measures.

Heterogeneity was explored through visual inspection of forest plots and the I2 statistic. I2 was used to indicate the percentage of variation across studies due to heterogeneity rather than chance.24 As a general rule of thumb for interpretation, the Cochrane guidelines indicate that I2 between 0-40% might not be important, I2 between 30-60% may represent moderate heterogeneity, 50-90% may represent substantial heterogeneity, and 75-100% considerable heterogeneity. 25 Although we intended to evaluate publication bias, tests are not meaningful when the N of studies is < 10.

***Meta-analysis of sensitivity and specificity***

Diagnostic accuracy coefficients of sensitivity and specificity were obtained from the study reports or where these were not available, calculated via the construction of a 2 x 2 contingency table providing the true positives, true negatives, false positives, and false negatives, when using the CPT score of omissions/inattention score, commissions/impulsivity-hyperactivity score, and total errors/ADHD score to identify ADHD. Sensitivity and specificity were meta-analyzed using the ‘MADA’ package version 1.0.3 in R-4.2.2, including a bivariate model that jointly analyzed pairs of sensitivity and specificity estimates whilst accounting for their correlation.26,27 Pooled sensitivity and specificity along with 95% CIs were derived from a random-effects model.

Four sensitivity analyses were performed: (a) using the specificity closest to 80% when multiple values were provided; (b) excluding any study with a high risk of bias in any of the domains of the QUADAS-2; (c) limiting the analyses to studies on clinical or high-risk samples; and (d) excluding Breaux et al.16 (due to being the only study on pre-school children and which had a time-lag between the CPT and the clinical evaluation). SROC curves were pooled for specific CPT versions whenever possible. Heterogeneity was explored through visual inspection of forest plots and the I2 statistic.

# Results

The flow chart in Figure 1 depicts the searching and screening process. Overall, we screened 8507 references, from which 433 were assessed for eligibility checking the full text. Reasons for the exclusion of articles moved to the full text evaluation phase can be found in Table S4, available online. Finally, nineteen studies were included in the current review.15–17,28–42

**Characteristics of included studies**

Included studies are described in Table 1. Eleven studies (58%) were published in the last 10 years. Samples were recruited from the US (k = 7, 37%), East Asia (k = 6, 32%), Israel (k = 3, 16%) and Sweden (k = 3, 16%), mostly in clinical settings (n=15, 79%). Most studies included only or mostly children under the age of 13 (n=13, 72%), whereas two recruited adolescents and four studies included both children and adolescents. As expected from the distribution of ADHD in the population, there were more male than female individuals in most studies.

Different implementations of the CPT were employed across the studies: the CCPT was the most frequently chosen measure (n=6), followed by the IVA+ (n=4), TOVA (n=3), MOXO (n=2), the Quantified Behavioral Test (qbTest; n=2), and other or unnamed versions (n= 2). We were not able to find any study on virtual CPTs reporting clinical utility data. Only a single study compared two different CPTs.32 The characteristics of the CPTs used in the included studies are described in Table S5, available online.

**Risk of bias**

The results of the assessment via the QUADAS-2 can be found in Figure 2 and Table S6, available online. Additional information can be derived from Table 1. Most studies were judged to have a high risk of bias in at least one domain. The most frequent was related to participant selection and, more specifically, to designs involving cases and controls. Over 50% (n=10) of the studies employed a non-consecutive case-control design. This can lead to biased estimates of sensitivity and specificity because more typical examples of individuals with and without the disorder might be selected for participation. Additionally, eight studies recruited controls from the general population, hence reducing the real world applicability of their results, as CPTs are typically used to identify ADHD among individuals who have identified mental health concerns (i.e., referred to a mental health clinic).

Eleven studies reported AUCs, and 17 reported sensitivity and specificity. Out of the latter, two were likely to have introduced bias in their estimations as the threshold used was the one that best identified ADHD in the study sample. Regarding the applicability of the studies in relation to the CPT predictors used, Jarrett et al.28 used as outcomes of the CCPT the Hit Reaction Time (HRT) and its standard error (HRT-SE). Because no other study used these measures in isolation, it was decided to exclude this study from the quantitative synthesis.

Nine studies (47%) employed a ‘gold standard’ reference combining information from interviews and multiple informants to diagnose ADHD, whereas four (21%) used parent interview only. Importantly, a third of the samples did not provide details on how the clinical diagnosis was carried out. One study was deemed to be at a high risk of bias in relation to the diagnostic process, as it included the CPT as one of the tests used as reference-standard for diagnosis. Finally, all studies but one evaluated a cross-sectional relationship between the diagnosis of ADHD and the results in the CPT (without any time lag), whereas Breaux et al.16 used the CPT with three-year-old children to determine whether CPT scores predicted ADHD diagnoses at age six.

CPTs´ computerized implementations can be commercially viable, especially when their use is supported scientifically. Hence, a reasonable source of potential bias can be the economic ties among authors of the studies and the companies selling the CPTs. Authors in three studies reported a direct financial conflict of interest or to have received funding from the companies selling CPTs. Conversely, authors in ten studies reported no conflicts of interest or potentially problematic funding. Seven studies did not provide a conflict of interest or funding section, precluding analysis of this potential moderator variable (Table S5).

**Quantitative synthesis**

Threshold-based statistics of clinical utility are reported alongside AUCs in Table S7, available online.

***Area Under the Curve***

Table 2, Tables S9-S10, and Figures S1 to S14, available online, summarize the results of pooling the AUCs. AUC was reported in ten samples for the CPT total score or ADHD index. However, six of these were derived from the study by Berger et al.15 and corresponded to different age ranges within the same study. To account for the non-independence of these data and the excessive weight that would be given to this study in a random effects analysis, we first conducted a fixed-effect analysis including solely the samples within the same study, the outcome of which was used in the random effects along the effect sizes from the other studies. Hence, the final analysis included five effect sizes.

A similar strategy was followed in the case of omissions, commissions, and sensitivity analyses. The AUC for the total score was 0.73 (95% CI=0.60, 0.87) derived from pooling the results over 632 cases and 837 controls. The AUC for omissions was similar, while it was obtained from a greater number of individuals (1083 controls and 966 cases from 8 different samples): 0.71 (CI=0.63, 00.80). The AUC for commissions was calculated by pooling results of 5 samples comprising 635 cases and 865 controls; it was somewhat lower, with a pooled AUC of 0.63 (0.53, 0.72). Heterogeneity was above 90% in all cases. When differentiating between CPT versions, MOXO obtained higher pooled AUCs than the IVA+ and CCPT, although all effect sizes were obtained by pooling data from only two samples or even a single sample. The AUC dropped to between .57 and .64 when only studies with low risk of bias were included. Heterogeneity was reduced when comparing CPT versions or including solely studies with low risk of bias. Limiting the analyses to studies on clinical and high-risk samples resulted in reduced AUCs (0.62 for total score, 0.65 for omissions, and 0.58 for commissions) (Table S10).

***Sensitivity and Specificity***

Table 2, Tables S11 to S14, and Figures S15 to S33, available online, summarize the results of pooling sensitivity and specificity.

Summary ROC curves yielded results in line with the previous analysis. When predictors dealing with total number of errors or a general ADHD scale were pooled, the AUC was 0.78, the pooled sensitivity was 0.75 (CI=0.66, 0.82) and the specificity was 0.71 (0.62, 0.78), with 819 and 835 combined cases and controls respectively from a total of eleven samples. For omissions (690 cases and 597 controls from 9 samples), AUC was 0.75, sensitivity 0.63 (CI=0.49, 0.75) and specificity 0.74 (CI=0.65, 0.81). Once again, the lowest numbers were obtained for commissions (444 cases and 451 controls from 7 samples), with AUC at 0.66, sensitivity 0.59 (CI=0.38, 0.77) and specificity 0.66 (CI=0.50, 0.78). Heterogeneity ranged between 76% for omissions to 86% for commissions. Importantly, thresholds chosen to obtain sensitivity and specificity varied between studies. Using the data point with a specificity closest to 0.8 (as opposed to the sensitivity closest to 0.8) as the selection rule for studies with multiple, similarly valid, predictors did not change overall results. When only studies with no risk of bias were pooled, AUC dropped to between 0.34 (commissions) to 0.72 (total errors); total errors had a sensitivity of 0.73 (CI=0.49, 0.88) and a specificity of 0.69 (CI=0.60, 0.76), omissions had a sensitivity of 0.58 (CI=0.37, 0.77) and a specificity of 0.71 (CI=0.39, 0.90); and commissions had a pooled sensitivity of 0.33 (CI=0.18, 0.53) and a specificity of 0.49 (0.12, 0.87). Limiting the analyses to studies on clinical and high-risk samples did not substantially change the results in relation to total scores, however it reduced the sensitivity and specificity for errors of omission and commission (see Table S10, available online, and figures S25-S30, available online). Eliminating Breaux et al.16 did not substantially change the results (Table S14, available online, Figures S31-S33, available online). MOXO again appeared as an outlier compared to the other CPT versions, with higher sensitivity and specificity for all measures. Conversely, heterogeneity was lower when comparing CPT versions or including solely studies with low risk of bias.

# Discussion

In this systematic review and meta-analysis, we have provided overall estimations of the Area Under the Curve, sensitivity, and specificity for three key predictor measures from CPTs: overall number of errors, omissions, and commissions, judged in relation to ADHD clinical diagnoses. To our knowledge, this is the first meta-analysis of measures of clinical utility of the CPT in ADHD. These measures of clinical utility specifically address the question of how well a tool distinguishes between two groups: in our case, children and adolescents with ADHD vs. control youth. This approach differs from a previous meta-analysis that pooled the means across groups on the CPT, which can only estimate the average difference between individuals with ADHD and controls in the measure of interest. Hence, our results are more easily translatable into practice in mental health settings and should help develop clearer best practice guidelines. Moreover, studies reporting on group averages do not provide measures of clinical utility for the most part. Hence, the present meta-analysis can be also seen as a separate but convergent line of evidence on the utility of the CPT for the identification of ADHD.

Our results indicate that performance of the CPT against a clinical diagnosis is not strong. Indeed, in most cases the AUC, sensitivity, and specificity of the CPT are in the lower range of what is typically considered acceptable, hovering around a value of 0.7. Such values preclude the usage of the CPT as a stand-alone screening tool in a population setting. Indeed, even in a clinical setting where the prevalence of the disorder is likely to be considerably higher, the CPT would still misidentify many subjects and should not be considered a valid diagnostic tool when used independently. Also, it important to note that there are no agreed cut-offs that were used across different studies, leaving clinicians uncertain in relation to how to implement CPT, even if metrics were better. There was one case when the AUC dropped to .49 (commission errors AUC in Edwards et al. 40). The other AUC indices were also under 0.6 for this study. This means that the CPT in Edwards et al. basically performed at chance level. While it is the result of a single study, it was the study judged to have the lowest risk of bias among those in the meta-analyses. If the results of this study were replicated, they would question even the value of the CPT in combination with other measures, as AUCs around 0.5 will hardly improve sensitivity or specificity at any given threshold.

If CPTs are used, even as supplementary measures, as part of an ADHD assessment, the present results indicate that ADHD is best identified by the overall scores provided by the different CPT implementations (either the total number of errors or a specifically tailored score of ADHD). Closely following, however, were measures of the number of omissions and inattention. The lowest clinical utility was found in the impulsivity/commission measures (with respect to AUCs or the sensitivity/specificity analyses), where all values dropped under the 0.7 threshold—considered unacceptable. This conclusion is in line with previous meta-analyses revealing greater differences between individuals with ADHD and controls in omission than in commission errors.9,11 Clinicians should be aware of these differences in accuracy between the different indices when considering using the CPT as a part of a diagnostic process.

Although we aimed to compare different commercial implementations of the CPT, we were unable, due to the nature of the available data, to obtain a clear picture in this regard. Indeed, CPT implementations vary in their duration, type of stimuli, and visual design all of which could be affecting performance. When taken at face-value, our analyses suggest a higher accuracy for the MOXO compared to other CPTs. However, our confidence in this difference is low and might not replicate in future studies. On the one hand, there were very few studies per implementation, on the other, the studies on the MOXO had a high risk of bias. Specifically, the threshold for which sensitivity and specificity were reported in Berger 2017 was the one that best separated cases from typically developing population controls, and hence, was likely to have inflated the clinical utility measures. Additionally, authors of MOXO studies were involved in its development and commercialization, suggesting a direct conflict of interest. Of note, we were only able to find a single head-to-head study comparing two different CPTs; specifically the TOVA and CCPT, with better results for the CCPT.32 Relatedly, we were not able to find any study on virtual reality CPTs that met our inclusion criteria and reported measures of clinical utility. This was a surprising finding considering the popularity of such measures in recent years. A future avenue of research should be studies directly comparing the clinical efficacy of different versions of the CPT, including virtual reality CPTs.

We found that risk of bias was rather strong in the studies reviewed, especially related to selection of cases. Most studies used a non-consecutive case-control design, which is likely to overestimate the differences between groups as clear-cut cases and controls are more likely to be recruited. Indeed, when studies with any risk of bias were eliminated from the analysis, the estimation of the pooled clinical utility of the CPTs was reduced, as was heterogeneity. We therefore urge that future studies should also aim to utilize more robust research designs. Regarding the reporting of effects, future studies should report the sensitivity and specificity at multiple unbiased cut-offs, in order to facilitate comparisons between studies and CPT implementations.43 Our systematic review also highlights other important limitations of the existing literature. Overall, studies were few and had low sample sizes. There was a greater number of studies in children versus adolescents, and, generally, studies did not report race and ethnicity of participants or included predominantly White non-Hispanic samples. Therefore, future studies should aim to include larger samples and should be carried out in more age, ethnic or geographically diverse populations so that results can be more easily generalized to the overall population of individuals with ADHD. Additionally, studies should strive to use gold-standard diagnoses and/or better describe the diagnostic approach used.

Overall, these findings strengthen the best current practice recommendations and mirror results found in relation to symptom rating scales for ADHD.13,43 Although CPTs and validated questionnaires of symptoms can help in the identification of ADHD, they are far from ideal, especially as stand-alone measures. Indeed, functional impairment and differential diagnostic issues need to be clarified as well in the diagnostic process. However, the specific case of the CPT raises a different, and maybe more epistemological debate. The level of granularity and level of analysis for neuropsychological measures such as the CPT is quite different from that used in the diagnostic manuals, namely overt specific behaviors in real-world settings.44 Indeed, whereas all individuals with ADHD will by definition have several of the core ADHD symptoms, ADHD is cognitively and behaviorally heterogeneous, with only a proportion of those with ADHD actually revealing a deficit in any one neuropsychological domain.45 Studies assessing whether combining the CPT with other measures improves sensitivity and specificity are warranted. However, even if data were available, their combination through meta-analysis would not be feasible. An issue for our current meta-analysis is that primary studies used different thresholds and tools. This problem would be compounded in a meta-analysis combining multiple tools, as there will be a very broad number of combinations of thresholds. Moreover, the weighting of each independent measure is as important as the independent thresholds. In this regard, the utilization of a CPT in the evaluation process is best used to characterize core deficits with potential clinical and prognostic value for an individual rather than providing information that can be used to determine diagnosis.46 Nevertheless, our meta-analysis provides an easy to interpret benchmark that permits the comparison of the clinical usefulness of the CPT against other tools, such as symptom scales, and between specific versions of the CPT.

In short, despite considerable research literature on CPT measures in the domain of ADHD, they show limited clinical utility in the differentiation between individuals with and without ADHD.

## References

1. Polanczyk G V, Willcutt EG, Salum GA, Kieling C, Rohde LA. ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. *Int J Epidemiol*. 2014;43(2):434-442. doi:10.1093/ije/dyt261

2. Faraone S V, Banaschewski T, Coghill D, et al. The World Federation of ADHD International Consensus Statement: 208 Evidence-based conclusions about the disorder. *Neurosci Biobehav Rev*. 2021;128:789-818. doi:10.1016/j.neubiorev.2021.01.022

3. Pievsky MA, McGrath RE. The Neurocognitive Profile of Attention-Deficit/Hyperactivity Disorder: A Review of Meta-Analyses. *Arch Clin Neuropsychol*. 2018;33(2):143-157. doi:10.1093/arclin/acx055

4. Albrecht B, Uebel-von Sandersleben H, Wiedmann K, Rothenberger A. ADHD History of the Concept: the Case of the Continuous Performance Test. *Curr Dev Disord Reports*. 2015;2(1):10-22. doi:10.1007/s40474-014-0035-1

5. Rosvold HE, Mirsky AF, Sarason I, Bransome, E. D., Jr., Beck LH. A continuous performance test of brain damage. *J Consult Psychol*. 1956;20(5):343–350. doi:https://doi.org/10.1037/h0043220

6. Keith Conners C, Sitarenios G, Ayearst LE. Conners’ Continuous Performance Test Third Edition. In: *Encyclopedia of Clinical Neuropsychology*. Cham: Springer International Publishing; 2018:929-933. doi:10.1007/978-3-319-57111-9\_1535

7. Sandford JA, Turner A. *Manual for the Integrated Visual and Auditory Continuous Performance Test*. Richmond, VA: Braintrain; 1995.

8. Greenberg LM, Waldmant ID. Developmental Normative Data on The Test of Variables of Attention (T.O.V.A.?). *J Child Psychol Psychiatry*. 1993;34(6):1019-1030. doi:10.1111/j.1469-7610.1993.tb01105.x

9. Huang-Pollock CL, Karalunas SL, Tam H, Moore AN. Evaluating vigilance deficits in ADHD: A meta-analysis of CPT performance. *J Abnorm Psychol*. 2012;121(2):360-371. doi:10.1037/a0027205

10. Goharinejad S, Goharinejad S, Hajesmaeel-Gohari S, Bahaadinbeigy K. The usefulness of virtual, augmented, and mixed reality technologies in the diagnosis and treatment of attention deficit hyperactivity disorder in children: an overview of relevant studies. *BMC Psychiatry*. 2022;22(1):4. doi:10.1186/s12888-021-03632-1

11. Parsons TD, Duffield T, Asbee J. A Comparison of Virtual Reality Classroom Continuous Performance Tests to Traditional Continuous Performance Tests in Delineating ADHD: a Meta-Analysis. *Neuropsychol Rev*. 2019;29(3):338-356. doi:10.1007/s11065-019-09407-6

12. Coghill DR, Seth S, Matthews K. A comprehensive assessment of memory, delay aversion, timing, inhibition, decision making and variability in attention deficit hyperactivity disorder: advancing beyond the three-pathway models. *Psychol Med*. 2014;44(9):1989-2001. doi:10.1017/S0033291713002547

13. Mulraney M, Arrondo G, Musullulu H, et al. Systematic Review and Meta-analysis: Screening Tools for Attention-Deficit/Hyperactivity Disorder in Children and Adolescents. *J Am Acad Child Adolesc Psychiatry*. 2022:1-15. doi:10.1016/j.jaac.2021.11.031

14. Hall CL, Valentine AZ, Groom MJ, 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):677-699. doi:10.1007/s00787-015-0798-x

15. Berger I, Slobodin O, Cassuto H. Usefulness and validity of continuous performance tests in the diagnosis of attention-deficit hyperactivity disorder children. *Arch Clin Neuropsychol*. 2017;32(1):81-93. doi:10.1093/arclin/acw101

16. Breaux RP, Griffith SF, Harvey EA. Preschool Neuropsychological Measures as Predictors of Later Attention Deficit Hyperactivity Disorder. *J Abnorm Child Psychol*. 2016;44(8):1455-1471. doi:10.1007/s10802-016-0140-1

17. Tallberg P, Råstam M, Wenhov L, Eliasson G, Gustafsson P. Incremental clinical utility of continuous performance tests in childhood ADHD – an evidence-based assessment approach. *Scand J Psychol*. 2019;60(1):26-35. doi:10.1111/sjop.12499

18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. March 2021:n71. doi:10.1136/bmj.n71

19. Whiting PF. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529. doi:10.7326/0003-4819-155-8-201110180-00009

20. Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *J Thorac Oncol*. 2010;5(9):1315-1316. doi:10.1097/JTO.0b013e3181ec173d

21. Glascoe FP. Screening for developmental and behavioral problems. *Ment Retard Dev Disabil Res Rev*. 2005;11(3):173-179. doi:10.1002/mrdd.20068

22. Schwarzer G. Meta: General Package for Meta-Analysis. R package version 4.15-1. https://www.rdocumentation.org/packages/meta/versions/4.9-6/topics/metagen.

23. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36. doi:10.1148/radiology.143.1.7063747

24. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558. doi:10.1002/sim.1186

25. Higgins J, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane; 2021. training.cochrane.org/handbook.

26. Doebler P. MADA: Meta-analysis of diagnostic accuracy. R package version 0.5.10. https://cran.r-project.org/web/packages/mada/mada.pdf.

27. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*. 2005;58(10):982-990. doi:10.1016/j.jclinepi.2005.02.022

28. Jarrett MA, Meter A Van, Youngstrom EA, Hilton DC, Ollendick TH. Evidence-Based Assessment of ADHD in Youth Using a Receiver Operating Characteristic Approach. *J Clin Child Adolesc Psychol*. 2018;47(5):808-820. doi:10.1080/15374416.2016.1225502

29. Kim JW, Lee YS, Han DH, Min KJ, Kim DH, Lee CW. The utility of quantitative electroencephalography and Integrated Visual and Auditory Continuous Performance Test as auxiliary tools for the Attention Deficit Hyperactivity Disorder diagnosis. *Clin Neurophysiol*. 2015;126(3):532-540. doi:10.1016/j.clinph.2014.06.034

30. Pan X, Ma H, Dai X. Value of integrated visual and auditory continuous performance test in the diagnosis of childhood attention deficit hyperactivity disorder. *Chinese J Contemp Pediatr*. 2007;9(3):210–212. doi:http://www.ncbi.nlm.nih.gov/pubmed/17582257

31. Perugini EM, Harvey EA, Lovejoy DW, Sandstrom K, Webb AH. The predictive power of combined neuropsychological measures for attention-deficit/hyperactivity disorder in children. *Child Neuropsychol*. 2000;6(2):101-114. doi:10.1076/chin.6.2.101.7059

32. Schatz AM, Ballantyne AO, Trauner DA. JAMA. *Assessment*. 2001;8(4):357-365. doi:10.1177/107319110100800401

33. Tang Y, Li Y-L, Yang H. A Comparison of IVA-CPT and EEG Biofeedback on Diagnosis for Children with ADHD. *Chinese J Clin Psychol*. 2005;13:94–95. doi:https://doi.org/10.16.16128

34. Won GH, Choi TY, Kim JW. Application of attention-deficit/hyperactivity disorder diagnostic tools: Strengths and weaknesses of the korean adhd rating scale and continuous performance test. *Neuropsychiatr Dis Treat*. 2020;16:2397-2406. doi:10.2147/NDT.S275796

35. Johansson V, Norén Selinus E, Kuja-Halkola R, et al. The Quantified Behavioral Test Failed to Differentiate ADHD in Adolescents With Neurodevelopmental Problems. *J Atten Disord*. 2021;25(3):312-321. doi:10.1177/1087054718787034

36. Berger I, Cassuto H. The effect of environmental distractors incorporation into a CPT on sustained attention and ADHD diagnosis among adolescents. *J Neurosci Methods*. 2014;222:62-68. doi:10.1016/j.jneumeth.2013.10.012

37. Bloch Y, Fixman M, Maoz H, et al. Can computerized cognitive tests assist in the clinical diagnosis of attention-deficit hyperactivity disorder? *J Neuropsychiatry Clin Neurosci*. 2012;24(1):111-114. doi:10.1176/appi.neuropsych.11010014

38. Chu KC, Huang YS, Tseng CF, Huang HJ, Wang CH, Tai HY. Reliability and validity of DS-ADHD: A decision support system on attention deficit hyperactivity disorders. *Comput Methods Programs Biomed*. 2017;140:241-248. doi:10.1016/j.cmpb.2016.12.003

39. Doyle AE, Biederman J, Seidman LJ, Weber W, Faraone S V. Diagnostic efficiency of neuropsychological test scores for discriminating boys with and without attention deficit-hyperactivity disorder. *J Consult Clin Psychol*. 2000;68(3):477-488. doi:10.1037//0022-006x.68.3.477

40. Edwards MC, Gardner ES, Chelonis JJ, Schulz EG, Flake RA, Diaz PF. Estimates of the validity and utility of the Conners’ Continuous Performance Test in the assessment of inattentive and/or hyperactive-impulsive behaviors in children. *J Abnorm Child Psychol*. 2007;35(3):393-404. doi:10.1007/s10802-007-9098-3

41. Hult N, Kadesjö J, Kadesjö B, Gillberg C, Billstedt E. ADHD and the QbTest: Diagnostic Validity of QbTest. *J Atten Disord*. 2018;22(11):1074-1080. doi:10.1177/1087054715595697

42. Inoue K, Nadaoka T, Oiji A, et al. Clinical evaluation of attention-deficit hyperactivity disorder by objective quantitative measures. *Child Psychiatry Hum Dev*. 1998;28(3):179-188. doi:10.1023/A:1022885827086

43. Karalunas SL. Editorial: Can We Accurately Screen for Attention-Deficit/Hyperactivity Disorder? Moving to a Dimensional, Multistep Process to Support Youth Development. *J Am Acad Child Adolesc Psychiatry*. 2022. doi:10.1016/j.jaac.2022.02.011

44. de la Fuente J, González-Torres MC, Aznárez-Sanado M, Martínez-Vicente JM, Peralta-Sánchez FJ, Vera MM. Implications of Unconnected Micro, Molecular, and Molar Level Research in Psychology: The Case of Executive Functions, Self-Regulation, and External Regulation. *Front Psychol*. 2019;10. doi:10.3389/fpsyg.2019.01919

45. Sonuga-Barke EJS, Coghill D. The foundations of next generation attention-deficit/hyperactivity disorder neuropsychology: building on progress during the last 30 years. *J Child Psychol Psychiatry*. 2014;55(12):e1-e5. doi:10.1111/jcpp.12360

46. Coghill D. Editorial: Comprehensive Child and Adolescent Mental Health Assessment Should Include Objective Assessments of Neurocognition. *J Am Acad Child Adolesc Psychiatry*. 2021;60(12):1461-1463. doi:10.1016/j.jaac.2021.01.003