The validity of the SNAP-IV in children displaying ADHD symptoms

Running head: the structure of the SNAP-IV for ADHD

Authors

Charlotte L Hall, Boliang Guo, Althea Z Valentine, Madeline J Groom, David Daley, Kapil Sayal, Chris Hollis

[1*]Charlotte L Hall, Division of Psychiatry and Applied Psychology, Institute of Mental Health, University of Nottingham Innovation Park, Triumph Road, Nottingham, NG7 2TU, UK. (ORCID: 0000-0002-5412-6165)

[2]Boliang Guo, Division of Psychiatry and Applied Psychology, Institute of Mental Health, University of Nottingham Innovation Park, Triumph Road, Nottingham, NG7 2TU, UK. (ORCID: 0000-0002-1844-705X)

[3]Althea Z Valentine, Division of Psychiatry and Applied Psychology, Institute of Mental Health, University of Nottingham Innovation Park, Triumph Road, Nottingham, NG7 2TU, UK. (ORCID: 0000-0003-2129-9278)

[4]Madeleine J Groom, Division of Psychiatry and Applied Psychology, Institute of Mental Health, University of Nottingham Innovation Park, Triumph Road, Nottingham, NG7 2TU, UK. (ORCID: 0000-0002-5182-518X)

[5]David Daley, Division of Psychiatry and Applied Psychology, Institute of Mental Health, University of Nottingham Innovation Park, Triumph Road, Nottingham, NG7 2TU, UK. (ORCID ID 0000-0002-3597-0408)

[6]Kapil Sayal, Division of Psychiatry and Applied Psychology, Institute of Mental Health, University of Nottingham Innovation Park, Triumph Road, Nottingham, NG7 2TU, UK. (ORCID: 0000-0002-2050-4316)

[7]Chris Hollis, University of Nottingham, CANDAL (Centre for ADHD and Neuro-developmental Disorders across the Lifespan) & MindTech, Developmental Psychiatry, University of Nottingham, Queen's Medical Centre, Nottingham, NG7 2UH, UK. (ORCID: 0000-0003-1083-6744).

*Corresponding Author

Dr Charlotte L Hall

Institute of Mental Health, Division of Psychiatry and Applied Psychology, University of Nottingham Innovation Park, Triumph Road, Nottingham, NG7 2TU, UK

Email: Charlotte.hall@nottingham.ac.uk. Phone: 0115 823 2438.

Abstract

The Swanson, Nolan, and Pelham Rating Scale (SNAP-IV) is a widely used scale that measures the core symptoms of attention deficit hyperactivity disorder (ADHD). However, there are contradictory findings regarding factor structure. Factor structure and measurement equivalence/invariance (ME/I) analysis on parent and teacher SNAP-IV for children referred for an ADHD assessment (n = 250; 6-17 years), revealed a 2-factor structure provided the best fit. SNAP-IV scores were also compared with clinician diagnosis of ADHD and research diagnoses of ADHD and hyperkinetic disorder (HKD). Parent ratings of inattention and hyperactivity/impulsivity were good predictors of research but not clinician diagnosis. For teacher ratings, only hyperactivity/impulsivity scores were associated with research and clinician diagnosis. SNAP-IV scores showed high sensitivity but low specificity to clinician diagnosis. The SNAP-IV is a valid outcome measure for use in RCTs and clinical settings, and is best used as a screening rather than a diagnostic tool for ADHD.

Keywords: Exploratory Structural Equation Modelling; Factor Analysis; Criterion validity; Measurement equivalence/invariance; SNAP-IV

Introduction

Attention deficit hyperactivity disorder (ADHD) is one of the most prevalent neuropsychiatric disorders of childhood, affecting 3-5% of children and characterised by symptoms of inattention, hyperactivity and impulsivity (NICE, 2008). The clinical assessment of ADHD is largely subjective, relying on clinical opinion which is typically informed by an observation of the young person as well as the opinions of parents and teachers. In an attempt to improve standardisation and comparability of parent and teacher reports, rating scales are often used as a tool to gather this information.

There are several rating scales which have been developed to measure ADHD as based on the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 1994). These rating scales are typically similar in using the DSM symptom descriptions but vary on the assessment of comorbid disorders (Bussing et al., 2008). For example, the Vanderbilt ADHD Diagnostic Rating Scales (Wolraich, Hannah, Baumgaertel, & Feurer, 1998; Wolraich et al., 2003) includes assessment of externalising and internalising disorders (such as mood and anxiety) as well as impairment.

One of the most extensively used questionnaires in treatment studies, the Swanson, Nolan and Pelham Rating Scale (SNAP-IV; Swanson et al., 2001) is a behavioural rating scale that measures the core symptoms of ADHD and oppositional defiant disorder (ODD) as defined by the DSM-IV (American Psychiatric Association, 2013), which can be completed by parents and teachers. The SNAP-IV is a well-used clinical and research tool, which has been used extensively to determine treatment outcome in research trials (Abikoff et al., 2005; Correia Filho et al., 2005; Wigal et al., 2004), including being used as the primary outcome for the Multimodal Treatment of ADHD (MTA) study (Jensen, 1999). However, despite its popularity, there is insufficient evidence with regards to: 1) the factor structure and criterion

validity of the SNAP-IV in a clinic-referred population, 2) the measurement invariance between parent and teacher ratings, 3) the predictive validity of the tool for ADHD and 4) the validity of the tool as a longitudinal outcome measure in research trials. It is important to address these issues to inform clinicians about the different dimensions of psychopathology being measured through the SNAP and to aid the interpretation of results from randomised controlled trials (RCTs) and epidemiological studies.

The original SNAP consisted of 43-items but was shortened to 26-items for the use in the MTA study (Swanson et al., 2001). This shortened version consists of DSM-IV symptoms for hyperactivity/impulsivity and inattention (totalling 18-items), and ODD (8-items). The 18-items are similar to that used by the Vanderbilt (Wolraich et al., 1998). The scoring for the SNAP-IV is available online (http://www.adhd.net/snap-iv-instructions.pdf), providing average ratings for inattentive, hyperactivity/impulsive, combined ADHD and ODD subscales. The limited research available indicates that the SNAP-IV has good internal consistency (Correia Filho et al., 2005; Stevens & Quittner, 1998), but Collett, Ohan, and Myers (2003) criticised the SNAP-IV for lack of published data on its psychometric properties.

There is uncertainty as to whether ADHD is best considered as two broad symptom domains of inattention and hyperactivity/impulsivity (as classified in both the DSM-IV and DSM-5; dsm5.org) or three broad symptoms of inattention, hyperactivity and impulsivity (as classified by the International Classification of Diseases (ICD)-10 criteria for hyperkinetic disorder). Ongoing attempts to clarify this issue have involved conducting factor analyses on SNAP-IV scores to elucidate the most appropriate dimensionality of ADHD. The majority of studies exploring the factor structure of the SNAP-IV have used general population samples and have supported a two-factor structure for the ADHD items. For example, Bussing et al. (2008) investigated the 26-items of the scale completed by both parents and teachers. Using

Confirmatory Factor Analysis (CFA) they found a three-factor structure, with two ADHD factors (inattention, hyperactivity/impulsivity) and an opposition factor best fit the data. In support of this, Swanson et al. (2012) used Exploratory Factor Analysis (EFA) on only the teacher rated 18-ADHD items of the SNAP-IV and found a two-factor structure (inattention, hyperactivity/impulsivity) in a community sample.

These approaches to factor analysis cannot account for any commonality between factors. However, the three core symptom domains of ADHD (inattention, hyperactivity and impulsivity) demonstrate high inter-correlations (Adams, Derefinko, Milich, & Fillmore, 2008). Using bi-factor or second order factor models can account for correlations between factors, allowing for a general, broader factor to emerge (Gustafsson & Åberg-Bengtsson, 2010). In a bi-factor model, the manifest variables are explained by both sub-factors and a single general factor. In second order models, the manifest variables are explained by the first-order factors, and the first-order factors are explained by a general factor. This approach has been used to investigate the factor structure of the SNAP-IV. Ulleb et al. (2012) examined the factor structure of the 18-ADHD SNAP-IV items in a large general child population in Norway. Using CFA they found a bi-factor model with a general ADHD factor and two specific factors for impulsivity and inattention best fit the data. Supporting the previous studies, Ulleb et al. (2012) found the sub-factors for 'hyperactivity' alone accounted for very little unique variance and was completely absorbed by the general ADHD factor. Their findings were consistent across both parent and teacher completed SNAP-IV scores. Using a clinical sample, Pillow et al. (1998) utilised a second-order approach and found the best fitting model contained a general 'ADHD' factor and two sub-factors (inattention and hyperactivity/impulsivity), with the impulsivity sub-factor being almost exactly determined by a general ADHD factor. However, their study only used ratings from parents to conduct the factor analysis. Given that the SNAP-IV is completed by both parent and teachers it is

important to consider the factor structure for both informants. The scores from parents and teachers are often used both clinically and in research settings under the assumption they have the same operational meaning. However, there is often poor inter-rater reliability between parent- and teacher- rated SNAP-IV's (Swanson, Lerner, March, & Gresham, 1999; Swanson et al., 2001), which may be because the raters are observing the child in different environments and completing the scale at different time points. However, to date there is no published evidence of measurement equivalence/invariance (ME/I) between raters and little evidence on the rating scales from teachers in a clinic referred sample.

Furthermore, there is a lack of research on the predictive validity of the SNAP-IV. Scores above the 95th percentile are generally considered to be clinically relevant, although criticism has been levelled at the generalizability of the sample used to define these cut-offs (Bussing et al., 2008), which were a group of low-income Hispanic secondary-school students (Gaub & Carlson, 1997). Based on their findings from a community sample, Bussing et al. (2008) demonstrated that parent scores above 1.8 on inattention and 2.4 for hyperactivity/impulsivity were predictive of an ADHD diagnosis, but there was no relationship between teacher scores and an ADHD diagnosis. Alda and Serrano-Troncoso (2013) investigated the predictive validity of the SNAP-IV in a sample of clinic-referred Spanish children. The parent-rated SNAP-IV demonstrated 82.3% sensitivity and 82.4% specificity with the clinicians' impression of ADHD, indicating that the SNAP was a useful screening tool, however, they did not investigate teacher scores.

The SNAP-IV is also a popular measurement for ADHD symptoms in randomised controlled trials (RCTs) (Jensen, 1999) to compare outcomes across groups (i.e. control and intervention arm) and time points (i.e. pre and post an intervention). Recently, the SNAP-IV was used in a randomised controlled trial 'Assessing QbTest Utility in ADHD' (AQUA-Trial (Hall et al., 2014; Hollis et al., 2018), which compared ADHD diagnostic rates between two

study arms for children and young people who had been referred to child and adolescent mental health services for an ADHD assessment. Currently, the measurement equivalence/invariance (ME/I) of the SNAP-IV between treatment groups and informants has not been investigated, however, this is needed to ensure it measures the same latent construct in the same way across groups/informants and different follow-up time points (Guo et al., 2017; Vandenberg & Lance, 2000).

Given the sparsity of research on the SNAP-IV factor structure and the inconsistencies in existing evidence, there is need for more rigorous factor analysis modelling to be employed to further understand the psychometric properties of the questionnaire, particularly in clinical settings if the tool is to be useful to aid diagnostic decision making. The existing evidence base has used either EFA or CFA, both of which are methodologically limited. The methodology of EFA means it is unable to incorporate latent EFA factors into subsequent analysis, and it does not lend itself to measure invariance across groups and/or times (Marsh, Morin, Parker, & Kaur, 2014). In CFA modelling, each item is allowed to load on only one factor and all non-target loadings are constrained to zero. In applied research, it is generally justifiable by theory and/or item contents that item(s) can cross load on different latent factors (Dickey & Blumberg, 2004; Goodman, 2001; Niclasen et al., 2012). Thus, restrictive zero loading typically results in an inflated CFA factor correlation and leads to biased estimates in CFA modelling when other variables are included in the model (Marsh et al., 2014). More recently, there have been methodological advances which integrate the best features of both EFA and CFA together as Exploratory Structural Equation Modelling (ESEM). This method applies a more rigorous test of the underlying factor structure together with the advanced statistical methods typically associated with CFA, including testing measurement invariance between groups. To date this technique has not been applied to investigate the factor structure of the SNAP-IV.

As the SNAP-IV is a frequently used measure of ADHD symptoms in both clinical and research settings, there is a need to further understand the factor structure and the accuracy of the SNAP-IV in detecting ADHD in a clinic-referred sample. Using data from children who were referred for a clinic assessment of ADHD and who participated in the AQUA-Trial, this study used ESEM to explore: 1) the factor structure of the SNAP-IV, 2) the measurement invariance between the two treatment groups and two informants (parent and teacher), and 3) the measurement invariance across follow-up time points (baseline, three and six months). The study also used logistic regression analyses to investigate the diagnostic accuracy of the SNAP-IV. These data are required to aid the interpretation of SNAP-IV results in clinical settings and in randomised controlled trials (RCTs) and epidemiological studies.

Method

Participants

Parent and teacher SNAP-IV rating scales were obtained from participants in the two-arm, multisite RCT 'AQUA-Trial' (Hall et al., 2014; Hollis et al., 2018). The trial evaluated the impact of providing a computerised test of attention and activity (QbTest) report on the speed and accuracy of diagnostic decision-making in children with suspected ADHD. Participants and their assessing clinician were randomised to either immediately receiving the QbTest report (QbOpen group) or having the report withheld until the study end (QbBlind group). Participants were followed from first appointment (baseline) until six-months later.

Eligible participants were 250 children aged 6-17 years referred for their first ADHD assessment to a child and adolescent mental health service (CAMHS) or community paediatric clinic in England. Exclusion criteria were previous or current ADHD diagnosis or assessment for ADHD, being non-fluent in English, and suspected moderate/severe

intellectual disability. When the child was under 16-years-old, parents provided written consent for their child's participation and assent (verbal or written) was gained from the young person. Ethical approval was granted by Coventry and Warwick Research Ethics Committee (Ref: 14/WM/0166) and research and development (R&D) permissions were obtained from each Trust. Outcome assessors were blind to group allocation throughout the study. Further details on the trial protocol and primary outcome have been previously reported (Hall et al., 2014; Hollis et al., 2018).

Measures

Swanson Nolan and Pelham – 4th version (SNAP-IV)

The SNAP-IV consists of 26-items that are rated on a 4-point scale ('not at all', 'just a little', 'quite a bit', 'very much'). The items are divided between three sub-scales: inattention (9-items), hyperactivity/impulsivity (9-items), and oppositional (8-items). Sub-scale scores are calculated by creating an average. Items for inattention and hyperactivity/impulsivity can be combined to also create a 'combined ADHD' score (Bussing et al., 2008). Higher scores represent more problem symptoms. The SNAP-IV was a secondary measure in the AQUA-Trial, used to assess ADHD symptoms at baseline (first appointment for an ADHD assessment), three months, and six months. The SNAP-IV was completed by parents and teachers online or on paper, and took approximately 15-minutes to complete.

Psychiatric diagnosis

Development and Well Being Assessment (DAWBA)

Children were assigned psychiatric diagnosis based on the Development and Well Being Assessment (DAWBA; Goodman, Ford, Richards, Gatward, & Meltzer, 2000). The

DAWBA is a package of interviews, questionnaires and rating techniques completed by parents and teachers and designed to generate ICD-10 and DSM-IV / DSM-5 output regarding childhood psychiatric diagnoses.

The DAWBA computer algorithm estimates the probability of having a psychiatric disorder in bands of <.1%, .5%, 3%, 15%, 50% and > 70% based on large community-based populations (Goodman et al., 2000), the top two levels have been shown to reliably indicate presence of a clinician-rated diagnosis and can be used as an alternative to clinician-rated diagnoses in research studies (Goodman, Heiervang, Collishaw, & Goodman, 2011). The parent DAWBA can take between 20 minutes to 2 hours to complete depending on the complexity of symptoms, and the teacher version takes less than 30 minutes. The DAWBA's were completed at baseline, either online or on the telephone with a researcher.

Consultation pro forma

As part of the AQUA trial, clinicians completed a short clinical record pro forma after each consultation. This pro forma documented whether the clinician had reached a confirmed diagnostic decision about the presence of ADHD. Clinicians could make a confirmed positive ADHD diagnosis, a confirmed excluded ADHD diagnosis or not reach a diagnostic decision about ADHD, within the six-month follow-up period. The clinician's diagnosis was made in accordance to DSM-IV/V criteria. For the purpose of this study, analysis was only conducted on confirmed positive or confirmed excluded ADHD diagnoses.

Data analysis

Stage 1 of the analysis was to explore the factor structure of the 18 ADHD items of the SNAP-IV using ESEM (Marsh et al., 2014). With reference to existing studies on the factor structure of SNAP-IV, first-order factors ranging from two to five and corresponding

bi-factor models were tested using baseline, three-month and six-month follow-up data combined using ESEM. Although both second order and bi-factor models have been used in previous studies, we chose to use bi-factor models as it is preferable to a second order CFA when exploring the general latent construct (Ullebø et al., 2012) of a measure. The factor structure was tested separately for both parent and teacher data. For each of the factor structures (two to five), factor loading patterns were additionally checked to finalise the most clinically meaningful factor structure as recommended by Kaplan (2008).

For the second stage of the analysis, Measurement invariance between baseline and follow up time (longitudinal ME/I), between treatment arms at/across measurement time, and between informant at/across measurement time were tested sequentially with configural invariance and scalar invariance model testing for ordinal items (Guo et al., 2017; Muthen & Muthen, 2017; Vandenberg & Lance, 2000). Configural invariance means the same pattern of fixed and free factors loadings is specified for each group, scalar invariance means alike items have the same factor loading and threshold estimates between each groups. ME/I is a prerequisite for meaningful group score comparison (Vandenberg & Lance, 2000). To explore between arm and informant measurement invariance at/across measurement times, the model with the relevant parameters set to be equal between groups at each follow-up time point was tested first, followed by the relevant parameters set to be equal between groups across follow-up time points.

Each ESEM model fitting was evaluated using the chi-square (χ^2) test together with comparative fit index (CFI, > 0.9), non-normed fit index (NNFI,> 0.9), and root mean square error of approximation (RMSEA, < 0.1) (Guo et al., 2017; Vandenberg & Lance, 2000). However, the χ^2 test is sensitive to a large sample size and non-normal data (Cheung & Rensvold, 2002). Thus, although χ^2 change ($\Delta\chi^2$) test was originally recommended to compare ME/I test model improvement, due to this sensitivity to non-normal data and large

sample size the CFI drop (ΔCFI <0.01) is generally recommended as the best indicator that two nested models are equivalent (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000), because ΔCFI is independent of both model complexity and sample size and not correlated with the overall fit measurements.

All ESEM models were conducted using software Mplus 8.1 (Muthen & Muthen, 2017). Ordinal item score was analysed with the WLSMV estimator and missing values were automatically accounted for using the full-information maximum likelihood approach built into Mplus (Enders & Bandalos, 2001; Graham, 2003). Logistic regression with STATA 14 was conducted to investigate whether the SNAP-IV can predict ADHD diagnosis made by independent research criteria rated via DAWBA using DSM-IV and clinician rated diagnosis.

Results

Table 1 shows the characteristics of the participants, and indicates that participants in the two trial arms were of similar composition. A total of 250 participants were consented, randomised and received the intervention (QbTest with the report either disclosed or withheld). Of these 123 were in the intervention arm (QbOpen) and 127 in the control arm (QbBlind).

<<Insert Table 1>>

SNAP-IV factor structure

In order to test for the proposed factor structures (including bi-factor models) in the existing literature and fully demonstrate the incremental value of the proposed models, we explored factor structures ranging from 2-5 factors. Based on previous research we explored both first order factors and bi-factor models in terms of item loading. A bi-factor model and first order model have identical modelling fitting estimates if they both have the same number of latent factors (i.e., the fit of a bi-factor with 1 specific sub-factor model (bi-1factor) is

identical to the fit of a 2 first order factor model fit; a bi-factor with 2 specific sub-factor model (bi-2factor) is identical fit to a 3 first order factor model fitting, and so on. Thus, for the purpose of presentation, the model fit for each factor (2 to 5) is labelled against its bi-factor equivocal in Table 2 and 3.

All ESEM modelling were conducted on combined baseline and follow-up data. The results show a different model fit for parents and teachers. The parent data showed that the CFI for the 3-factor model was slightly improved compared to the 2-factor model, but no substantial improvements were made with models of 4 and 5 factors, indicating a 3-factor model best fit the parent data (RMSEA =0.28, CFI = 0.977, NNFI = 0.975, ΔCFI = 0.015) (Table 2). For the teacher data, the CFIs showed no substantial improvement with the addition of more factors (3, 4, and 5) gains of less than 0.01, indicating that a 2-factor model best fit the teacher data (RMSEA = 0.32, CFI = 0.984, NNFI = 0.983) (Table 2).

<<Insert Table 2>>

This finding was confirmed by the model fitting results of a longitudinal configural invariance model fit results, which showed that for parent data, factors 2, 4, and 5 did not fit the data as well as a 3-factor model (RMSEA = 0.034, CFI = 0.970, NNFI = 0.965, ΔCFI = 0.017) (Table 3), and for teachers, factors 3-5 did not fit the data any better than the 2-factor model, with CFI gains of less than 0.01 (RMSEA = 0.34, CFI = 0.982, NNFI = 0.981) (Table 3).

<<Insert Table 3>>

Next, we investigated the factor loading pattern of the different structures, including for the bi-factor models. Meaningful loadings were assessed using the criteria of 0.32

("poor"), 0.45 ("fair"), 0.55 ("good"), 0.63 ("very good"), and 0.71 ("excellent") (Tabachnick & Fidell, 2007).

For parent data, the item loadings for the proposed 3-factor model did not show good face validity: items 1-9 (inattention items) mapped with poor-to-excellent loadings on to a inattention factor, and items 10-18 (hyperactivity/impulsivity items) mapped with poor-to-excellent loadings on to a hyperactivity/impulsivity factor. However, the third factor did not show loadings that were meaningful (see Electronic Supplement Table S1). Conversely, although the results of model fittings demonstrated an improved fit from 2 to a 3-factor model, the item loadings were more theoretically consistent for the 2-factor model, with items 1-9 mapping on to the inattention factor and items 10-18 mapping on to the hyperactivity/impulsivity factor with loadings ranging from good to excellent (see Electronic Supplement Table S2). The lowest rating was for the impulsivity item-10 "fidgets" with a loading of 0.59 (good), all other items loaded at 0.63 and over (very-good to excellent).

Findings from the bi-factor models did not support a general factor across any of the models, with multiple items loading under 0.32 (see Electronic Supplement Table S3 for parent bi-factor 2).Thus, a 2-factor solution was considered to be theoretically consistent where items loaded strongly on to one factor and weakly on to other factors.

For the teacher data, the item loadings were consistent with the model results and showed a 2-factor model best fit the data, with items 9-17 loading with excellent fit on to the inattention factor and items 18-26 loading poor/fair-to excellent on the hyperactivity/impulsivity factor (see Electronic Supplement Table S4). One item, item 10 "fidgety", loaded onto both factors greater than 0.30, with a difference of <0.20 between factor loadings (inattention = 0.42, hyperactivity/impulsivity = 0.44) at baseline. Again, this item demonstrated the lowest loading on both factors across all three time points (baseline, follow-up 1 and follow-up 2).

As with parent data, there was no evidence of a 3-factor model for teacher data (see Electronic Supplement Table S5). The bi-factor 2 models for teachers did show support for a general factor, with all items loading onto a general ADHD factor (Electronic Supplement Table S6). Although all inattention factors loaded, hyperactivity/impulsivity items 18-20 did not load onto any sub-factor and loadings onto items 21 (difficulties playing or engaging with leisure activities) and 22 (is "on the go" or often acts as if "driven by a motor") were poor (0.32 and 0.34 respectively), indicating that these hyperactivity/impulsivity factors were better explained by a general ADHD factor. Although both the 2 factor model and bi-2 factor models showed theoretically reasonable loadings, the 2 factor model showed best model fitting results. As such, after examining both the model results and item loadings, a 2-factor solution best fit both the parent and teacher data (Kaplan, 2008).

Correlations between the two factors are shown in Figure 1 (parent) and Figure 2 (teacher) Correlations ranged from poor-to-excellent for parent data and poor-to-good for teacher across the three time points.

<<<Insert Figure 1 & Figure 2 here>>>>

Measurement invariance test of a 2 factor structure across time points

The model fitting indices for longitudinal ME/I test are presented in Table 4. The threshold invariance model fitting results showed that the 2-factor structure model evidenced a strong factorial invariance across measurement time points (baseline and follow-up) (parent: RMSEA = 0.041, CFI = 0.948, NNFI = 0.948, ΔCFI = -0.005; teacher: RMSEA = 0.40, CFI = 0.974, NNFI = 0.974, ΔCFI = -0.008).

<<Insert Table 4>>

Table 5 presents the results from the ME/I model between arms at and across measurement time. The results indicated the 2-factor structure remained stable between

16

treatments across baseline and follow-ups, only with a small amount of item threshold estimates freely estimated between baseline and follow-up time for both parent (RMSEA = 0.38, CFI = 0.957, NNFI = 0.958, ΔCFI = -0.008) and teacher data (RMSEA = 0.31, CFI = 0.985, NNFI = 0.985, ΔCFI = -0.004).

<<Insert Table 5>>

Finally, we compared the longitudinal measurement invariance between parent and teacher data across time points. The results are presented in Table 6 and indicate limited evidence for strong measurement invariance between teacher and parents' ratings. Specifically, although the scalar invariance model fitted the data well, the invariant threshold model fitting dropped too much when moved from the invariant configural model. This is demonstrated by the CFI drop of larger than 0.01 for all the threshold equivalence models.

<<Insert Table 6>>

Association between SNAP-IV score and ADHD diagnosis

To test the criterion validity of the SNAP-IV, we investigated the association between SNAP scores and the child's diagnosis assigned by an independent research diagnosis (DAWBA DSM-IV/V and ICD) and the clinician. This was conducted on the two factors (inattention and hyperactivity/impulsivity). The findings are presented in Table 7 and show that children with higher parent or teacher ratings on the SNAP-IV hyperactivity/impulsivity scale were more likely to receive a DAWBA diagnosis of ADHD (DSM-IV/V) (parent: OR = 4.31, p = 0.000; teacher: OR = 2.30, p = 0.001) and hyperkinetic disorder (ICD-10) (parent: OR = 3.75, P = 0.000; teacher: OR = 2.32, p = 0.001) , and also a clinician diagnosis of ADHD (parent: OR = 1.92, p = 0.052; teacher: OR = 2.51, p = 0.011) (although for parent scores this only approached statistical significance when assessed against clinician diagnosis). Scores on inattention showed less association with diagnostic predictions. Teacher

ratings on SNAP-IV inattention scale were not associated with either DAWBA or clinician

diagnosis. Parent ratings on SNAP-IV inattention were associated with DAWBA predictions

of ADHD (DSM-IV/V) (OR = 3.37, p = 0.001) and hyperkinetic disorder (ICD-10) (OR =

2.81, p = 0.003), but not with clinician diagnosis. Combining parent and teacher scores did

not improve the association (see Electronic Supplement Table S7).

<<Insert Table 7>>

Table 8 presents the sensitivity/specificity and positive/negative predictive value of the

SNAP-IV scores against clinician and independent research diagnoses (DAWBA) and shows

largely similar results for parent and teacher data: SNAP-IV scores are sensitive to picking up

ADHD but less accurate in determining those without ADHD, particularly when compared to

clinician diagnosis. The specificity of the scale is however better for hyperkinetic disorder,

which has a more strict diagnostic criteria.

<<Insert Table 8>>

Discussion

The SNAP-IV is an internationally recognised tool to aid the diagnostic assessment

and symptom management of ADHD. This study investigated the factor structure and

measurement invariance of the SNAP-IV in a clinic-referred sample using novel and robust

ESEM analysis to determine its validity as an outcome measure. Tests of association and

diagnostic accuracy were also conducted to investigate its validity as a diagnostic aide. The

findings showed that a 2-factor structure consisting of one inattention factor and one

hyperactivity/impulsivity factor best fit the data. The 2-factor structure was invariant across

baseline and follow-ups, indicating it is a valid measure of outcome. Although the two-factor

structure was found using parent and teacher data, there was a difference in model-fitting

results and strong measurement invariance was not demonstrated between these two

informants, indicating that the SNAP-IV measures the same construct across parents and

teachers but in a slightly different way. To the best of the authors' knowledge, this is the first investigation into measurement invariance for SNAP-IV across time and informants.

Parent and teacher ratings on the hyperactivity/impulsivity scale were associated with a research (DAWBA) and clinician diagnosis of ADHD. However, for the inattention scale, only parent scores on inattention were associated with research (DAWBA) diagnoses only. Further analysis to establish sensitivity/specificity indicated that the SNAP-IV scores were sensitive to picking up ADHD but less accurate in determining those without ADHD. The specificity of the scale was particularly poor when compared to clinician diagnosis but was substantially increased for hyperkinetic disorder – which is defined more restrictively than ADHD.

Model fitting data comparing 2-5 factor models indicated that a 2-factor model best fit teacher data. However, for parents, a 3-factor model produced slightly superior model fitting results to a 2-factor model. Further analysis of item loadings for the 3-factor model showed poor face validity, whereas the 2-factor model showed meaningful loading patterns onto a factor of inattention and a factor of hyperactivity/impulsivity, thus the 2-factor model was selected as the best fit. As the SNAP-IV is based on the DSM items, which categorises ADHD into a 2-factor structure, it is not surprising that our results reveal the same structure. Furthermore, although the paper aimed to investigate the structure of the SNAP-IV, given the close overlap between the items of SNAP-IV and DSM-IV, the findings also support the structure of the DSM-IV. The measurement invariance of this 2-factor model across treatment groups (QbOpen and QbBlind) indicate the validity of the scale to compare outcomes in RCTs. Furthermore, the measurement invariance across time points (baseline and follow-ups) validate the SNAP-IV as a measure of treatment outcome. Measurement invariance across treatment groups and time points is necessary to be able to meaningfully compare outcomes between two treatment groups. As the SNAP-IV is often used as a

measure of treatment outcome in trials, this is a particularly noteworthy finding. Given the similarity of the SNAP-IV to the Vanderbilt, it is likely that this measurement invariance may also be demonstrated in the 18-ADHD items of this scale.

Although the same model fit was selected for both parent and teacher data, strong measurement invariance was not demonstrated between parent and teacher ratings indicating that there are systematic response differences towards the same child's behaviour between teachers and parents. Measurement invariance indicates that the same construct is being measured. To elaborate, a parent and teacher would rate the same child's behaviour in the same way (as the same model best fit the data) but to a different degree (as there was no strong evidence for measurement invariance). This suggests there are some differences in the way parents and teachers are rating the SNAP-IV and thus a direct comparison between parent and teacher scores is not advisable. These results may partially reflect the different environments in which the two informants observe the child and indicate that the two scores are not directly comparable. For example, Pappas (2006) reports that the behavioural characteristics of ADHD are different across school and home. Given that school is a more structured environment, it is possible that issues of attention can be perceived as noncompliance (DuPaul, Weyandt, & Janusis, 2011) and issues of academic performance deficits, organisational skills and disruptive behaviours are of key importance (Fabiano et al., 2009), which may be less relevant in the home environment.

The confirmation of the two-factor structure indicates that in a clinic-referred sample ADHD is best considered in terms of two broad symptom domains: inattention and hyperactivity/impulsivity, as classified by the DSM-IV/V and supported by the findings of Bussing et al. (2008) and Swanson et al. (2012). In doing so, we also support findings from Wolraich et al. (2003) who demonstrated that a two-factor structure best-fit the 18-items of the Vanderbilt, which are also based on the same DSM items used by the SNAP-IV. Our

findings showed little support for a 'general' ADHD factor: for parent SNAP-IV there was no evidence of this, while for teacher rated SNAP-IV item loading patterns showed some evidence for a general factor. Given that model fitting results clearly indicated a 2-factor solution best fitted the data, this model was selected over the bi-factor model. Although the item loading patterns were not supportive of a 'general' ADHD factor, correlations between the two factors ranged from poor-to-excellent, indicating a degree of commonality between the core symptom domains of ADHD (Adams et al., 2008). To our knowledge, only one study, conducted over 20 years ago, has looked at parent SNAP-IV bi-factor models in an American clinic referred-sample, whereby support for a general ADHD factor was shown (Pillow et al., 1998). Our findings supporting a first-order factor model over a bi-factor model may be a result of a cultural difference in the way ADHD is perceived in the UK/America. For example, a review of the literature speculated that there may be greater emphasis on hyperactivity in non-north American samples (Buitelaar et al., 2006) which may explain why the variance in the hyperactivity items were subsumed by a general factor. Alternatively, the support of a first-order model may reflect a change over time in the way ADHD is reviewed as our knowledge advances. Or, the difference in findings may be a result of more advanced and robust statistical procedures.

For the two-factor model, all items, bar one, showed good-to-excellent loadings for the inattention and hyperactivity/impulsivity sub-scales. The exception was item 10 "fidgety". This item showed the worst loading for both parent and teacher SNAP-IV, although for both informants the item mapping reached statistical significance. For teacher data, this item also loaded onto the inattention scale, indicating that for teachers, fidgeting is viewed equally as inattentive as it is impulsive. Items were also examined for cross-loading onto the other factor, with cross loadings being deemed problematic if the item loading

reached statistical significance. There was no evidence of significant cross-loadings onto the other factor, indicating the validity of two distinct factors of ADHD.

In support of previous findings, we found a greater association between parent scores on the SNAP-IV and ADHD diagnosis than teacher scores (Bussing et al., 2008). Interestingly, for teachers, inattention scores were not associated with any ADHD diagnosis, and for parents, inattention scores were not associated with a clinician diagnosis (although they were associated with DAWBA diagnoses). The findings suggest that the hyperactivity/impulsivity scores may be more clinically useful when determining presence of ADHD. Our findings show that SNAP-IV scores showed the best combination of sensitivity and specificity to hyperkinetic disorder, which is considered a narrower sub-type reflecting the more severe ADHD. The very high sensitivity but poor specificity when compared to clinician diagnosis indicate that the SNAP-IV may be useful as a screening tool to identify ADHD, but less clinically useful as a diagnostic tool. This is consistent with current best-practice guidelines which recommend that ADHD is too complex to be diagnosed based on the findings of one instrument (NICE, 2008).

A notable strength to the research was the choice of analytical approach, drawing upon current 'best practice' principles endorsed by Vandenberg and Lance (2000), validating the use of chi-square, RMSEA, CFI to determine model fit. Furthermore, in line with their recommendations and previous research (Guo et al., 2017) we gave preference to the CFI score over the chi-square given that chi-square results are sensitive to large sample sizes and non-normal data, rendering it an imperfect methodology for this sample .

Although our findings are strengthened by the use of a clinical sample, it could be considered a limitation that our sample had all been referred for an ADHD assessment. Utilising this sample provides the opportunity to understand the validity of the SNAP-IV in

children referred for queried ADHD, in which it is arguably most used in clinic, however, our criterion validity results may not apply to differentiating ADHD from a normal population or samples presenting with another primary difficulty.

Given our sample was predominantly male, we did not make comparisons across gender. We also did not compare measurement invariance across age, as SNAP-IV is designed to assess symptoms relative to other children of the same age. Additionally, our sample was predominantly white, British and lacked ethnic diversity. As such, our results may not be generalisable to other ethnic groups. Furthermore, there may be differences in reporting styles between parents and teachers who agree to participate in an RCT. Finally, there were missing data for both parent and teacher SNAP-IV's as well as DAWBAs and clinician-rated diagnoses and this should be reflected upon when interpreting the results. Despite this, the findings are strengthened by the use of the novel ESEM approach, which combines both the benefits of EFA and CFA (Marsh et al., 2014) and is unique in demonstrating the measurement invariance across time and group for SNAP-IV.

In conclusion, utilising ESEM a 2-factor structure was shown to best fit both parent and teacher rated SNAP-IV's for children and young people who had been referred to a clinic for an ADHD assessment and participating in an RCT. This 2-factor structure was invariant across time points and RCT treatment groups, demonstrating that the SNAP-IV is a robust and valid measure of outcome for research studies and as an aide to clinical interpretation of symptom improvement over time. Measurement invariance was not found between parent and teacher scores indicating parent and teacher ratings of SNAP-IV should not be directly compared. In general, parent scores showed more association with ADHD diagnoses than teacher scores and the findings indicated hyperactivity/impulsivity scores may be more clinically useful than inattention scores. The very high sensitivity and low specificity of

SNAP-IV scores to clinician diagnosis suggest the tool may be more useful as a screening rather than diagnostic tool.

References

Abikoff, H., McGough, J., Vitiello, B., McCracken, J., Davies, M., Walkup, J., Riddle, M., Oastis, M., Greenhill, L., Skrobala, A., March, J., Gammon, P., Robinson, J., Lazell, R., McMadon, D.J., Ritz, L. (2005). Sequential pharmacotherapy for children with comorbid attention-deficit/hyperactivity and anxiety disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(5), 418-427.

Adams, Z. W., Derefinko, K. J., Milich, R., & Fillmore, M. T. (2008). Inhibitory functioning across ADHD subtypes: recent findings, clinical implications, and future directions. *Developmental Disabilities Research Reviews, 14*(4), 268-275.

Alda, J. A., & Serrano-Troncoso, E. (2013). Attention-deficit hyperactivity disorder: agreement between clinical impression and the SNAP-IV screening tool. *Actas Espanolas de Psiquiatria, 41*(2), 76-83.

American Psychiatric Association (1994). Diagnostic and statistical manual of mental disorders. (4th edition ed.) Washington, DC.

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5®): American Psychiatric Pub, Washington, DC.

Buitelaar, J. K., Barton, J., Danckaerts, M., Friedrichs, E., Gillberg, C., Hazell, P. L., Hellemans, H., Johnson, M., Kalverdijk, L.J., Masi, G. and Michelson, D. (2006). A comparison of North American versus non-North American ADHD study populations. *European Child & Adolescent Psychiatry, 15*(3), 177-181.

Bussing, R., Fernandez, M., Harwood, M., Hou, W., Garvan, C. W., Eyberg, S. M., & Swanson, J. M. (2008). Parent and teacher SNAP-IV ratings of attention deficit hyperactivity disorder

symptoms: psychometric properties and normative ratings from a school district sample. *Assessment, 15*(3), 317-328.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Collett, B. R., Ohan, J. L., & Myers, K. M. (2003). Ten-year review of rating scales. V: scales assessing attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(9), 1015-1037.

Correia Filho, A. G., Bodanese, R., Silva, T. L., Alvares, J. P., Aman, M., & Rohde, L. A. (2005). Comparison of risperidone and methylphenidate for reducing ADHD symptoms in children and adolescents with moderate mental retardation. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(8), 748-755.

Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the strengths and difficulties questionnaire: United States, 2001. *Journal of American Academy Child and Adolescent Psychiatry, 43*. doi:10.1097/01.chi.0000132808.36708.a9

DuPaul, G. J., Weyandt, L. L., & Janusis, G. M. (2011). ADHD in the classroom: Effective intervention strategies. *Theory into Practice, 50*(1), 35-42.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of Full Information Maximum Likelihood Estimation for missing data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(3), 430-457.

Fabiano, G. A., Pelham Jr, W. E., Coles, E. K., Gnagy, E. M., Chronis-Tuscano, A., & O'Connor, B. C. (2009). A meta-analysis of behavioral treatments for attention-deficit/hyperactivity disorder. *Clinical Psychology Review, 29*(2), 129-140.

Gaub, M., & Carlson, C. L. (1997). Behavioral characteristics of DSM-IV ADHD subtypes in a school-based population. *Journal of Abnormal Child Psychology, 25*(2), 103-111.

Goodman, A., Heiervang, E., Collishaw, S., & Goodman, R. (2011). The 'DAWBA bands' as an ordered-categorical measure of child mental health: description and validation in British and Norwegian samples. *Social Psychiatry and Psychiatric Epidemiology, 46*(6), 521-532.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. . *Journal of American Academy Child and Adolescent Psychiatry,40*. doi:10.1097/00004583-200111000-00015

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Well-Being Assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry, 41*(05), 645-655.

Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 10*, 80-100. doi:10.1207/S15328007SEM1001_4

Guo, B., Aveyard, P., Fielding, A., & Sutton, S. (2009). The factor structure and factorial invariance for the decisional balance scale for adolescent smoking. *International Journal of Behavioral Medicine, 16*(2), 158-163.

Guo, B., Kaylor-Hughes, C., Garland, A., Nixon, N., Sweeney, T., Simpson, S., Dalgeleish, Ramana, R., Tang, M., Morriss, R. (2017). Factor structure and longitudinal measurement invariance of PHQ-9 for specialist mental health care patients with persistent major depressive disorder: Exploratory Structural Equation Modelling. *Journal of Affective Disorders, 219*, 1-8.

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97-121). Washington, DC, US: American Psychological Association.

Hall, C. L., Walker, G. M., Valentine, A. Z., Guo, B., Kaylor-Hughes, C., James, M., Daley, D., Sayal, K., & Hollis, C. (2014a). Protocol investigating the clinical utility of an objective measure of activity and attention (QbTest) on diagnostic and treatment decision-making in children and young people with ADHD—'Assessing QbTest Utility in ADHD' (AQUA): a randomised controlled trial. *BMJ Open, 4*(12). doi:10.1136/bmjopen-2014-006838

Hollis, C., Hall, C. L., Guo, B., James, M., Boadu, J., Groom, M. J., Boadu, J., Groom, M.J., Brown, N., Kaylor-Hughes, C., Moldavsky, M., Valentine., A.Z., Walker, G. M., Daley, D., Sayal, K., Morris, R. (2018). The impact of a computerised test of attention and activity (QbTest) on diagnostic decision-making in children and young people with suspected attention deficit hyperactivity disorder: single-blind randomised controlled trial. *Journal of Child Psychology & Psychiatry*, 59 (12), 1298-1308.

Jensen, P. S (1999). A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry, 56*(12), 1073.

Kaplan, D. (2008). *Structural equation modeling: foundations and extensions* (Vol. 10): Sage Publications.

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85-110.

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*(1), 85-110.

Muthen, B. O., & Muthen, L. K. (2017). *Mplus user's guide* Retrieved from www.StatModel.com

NICE. (2008). *Attention deficit hyperactivity disorder:diagnosis and managment of ADHD in children, young people and adults. Clinical Guideline 72*. London: National Institute for Health and Clinical Excellence.

Niclasen, J., Teasdale, T. W., Andersen, A.-M. N., Skovgaard, A. M., Elberling, H., & Obel, C. (2012). Psychometric properties of the Danish Strength and Difficulties Questionnaire: the SDQ assessed for more than 70,000 raters in four different cohorts. *PloS One, 7*(2), e32025.

Pappas, D. (2006). ADHD Rating Scale-IV: Checklists, norms, and clinical interpretation. *Journal of Psychoeducational Assessment, 24*(2), 172-178.

Pillow, D. R., Pelham, W. E., Hoza, B., Molina, B. S., & Stultz, C. H. (1998). Confirmatory factor analyses examining attention deficit hyperactivity disorder symptoms and other childhood disruptive behaviors. *Journal of Abnormal Child Psychology, 26*(4), 293-309.

Stevens, J., & Quittner, A. L. (1998). Factors influencing elementary school teachers' ratings of ADHD and ODD behaviors. *Journal of Clinical Child Psychology, 27*(4), 406-414.

Swanson, J., Lerner, M., March, J., & Gresham, F. M. (1999). Assessment and intervention for attention-deficit/hyperactivity disorder in the schools: lessons from the MTA study. *Pediatric Clinics of North America, 46*(5), 993-1009.

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P., Arnold, L. E., Conners, C. K., Abikoff, H. B., Clevenger, B.A., Davies, M.P.H., Elliott, G.R., Greenhill, L.L., Hetchtman, L., Hoza, B., Jensen, P. S., March., J. S., Newcorn, J. H., Owens, E.B., Pelham, W.E., Schiller, E., Severe, M.S., Simpson, S., Vitiellos, B., Wells, K., Wigal, T., Wu, M. (2001). Clinical relevance of the primary findings of the MTA: success rates based on severity of ADHD and ODD symptoms at the end of treatment. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(2), 168-179.

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., Lakes, K. (2012). Categorical and Dimensional Definitions and Evaluations of Symptoms of ADHD: History of the SNAP and the SWAN Rating Scales. *The International journal of educational and psychological assessment, 10*(1), 51.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*: Allyn & Bacon/Pearson Education.

Ullebø, A. K., Breivik, K., Gillberg, C., Lundervold, A. J., & Posserud, M. B. (2012). The factor structure of ADHD in a general population of primary school children. *Journal of Child Psychology and Psychiatry, 53*(9), 927-936.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4 - 70. doi:10.1177/109442810031002.

Wigal, S., Swanson, J. M., Feifel, D., Sangal, R. B., Elia, J., Casat, C. D., Zeldis, J.B., Conners, C. K. (2004). A double-blind, placebo-controlled trial of dexmethylphenidate hydrochloride and d, l-threo-methylphenidate hydrochloride in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(11), 1406-1414

Wolraich, M. L., Hannah, J. N., Baumgaertel, A., & Feurer, I. D. (1998). Examination of DSM-IV criteria for attention deficit hyperactivity disorder in a county-wide sample. *Journal of Developmental and Behavioral Pediatrics*, *19*(3), 162-168.

Wolraich, M. L., Lambert, W., Doffing, M. A., Bickman, L., Simmons, T., & Worley, K. (2003). Psychometric properties of the Vanderbilt ADHD diagnostic parent rating scale in a referred population. *Journal of Pediatric Psychology, 28(8),* 559-568.

Figure titles and captions

Figure 1: item factor loadings and correlations for baseline/follow-up parent

Note. Data is shown as baseline/follow-up one/follow-up two

Figure 2: item factor loadings and correlations for baseline/follow-up teacher

Note. Data is shown as baseline/follow-up one/follow-up two

Tables

Table 1

Sociodemographic and clinical characteristics of participants at baseline with QbTest report withheld (QbBlind group) or QbTest report disclosed (QbOpen group).

	QbBlind (control; report withheld) (*n* = 127)	QbOpen (intervention; report disclosed) (*n* = 123)	Total sample (*n* = 250)
Sex (%)			
Male	102 (80)	95 (77)	197 (79)
Female	25 (20)	28 (23)	53 (21)
Age (years)			
Mean age (SD)	9.4 (2.8)	9.5 (2.8)	9.5 (2.8)
Min-max	(5.9, 16.2)	(6.0, 17.4)	(5.9 – 17.4)
Ethnicity %*	*n*=89	*n*=83	*n*= 172
White	80 (90)	73 (88)	153 (89)
Mixed	5 (6)	6 (7)	11 (6)
Other	4 (4)	4 (5)	8 (5)
SNAP-IV – Parent *:mean(SD)	*n* = 92	*n* = 81	*n* = 173
Inattention score	2.22 (.55)	2.20 (.60)	2.22 (.58)
Hyperactivity/impulsivity score	2.07 (.71)	2.10 (.72)	2.08 (.72)
Combined ADHD score	2.15 (.50)	2.16 (.55)	2.15 (.52)
SNAP-IV – Teacher *:mean(SD)	*n* = 81	*n* = 74	*n*= 155
Inattention score	1.73 (.80)	1.74 (.85)	1.74 (.82)
Hyperactivity/impulsivity score	1.53 (.88)	1.47 (.99)	1.50 (.933)
Combined ADHD score	1.63 (.74)	1.61 (.85)	1.62 (.79)

Type of clinical service (%)	$n = 127$	$n = 123$	$n = 250$
CAMHS	60 (47)	59 (48)	119 (48)
Community Paediatrics	67 (53)	64 (52)	131 (52)

Data are n (%) or mean (SD/range). 'Other' ethnicity includes Pakistani, Indian and Other Asian. *Data not available for all randomised participants. CAMHS = child and adolescent mental health services. Higher scores indicate more problems.

Table 2

ESEM Model Fitting Indices for Parent and Teacher Data with equal item loadings across time points (baseline and follow-ups)

Factor	χ^2(df),p=	RMSEA	CFI	NNFI	ΔCFI	Δχ^2(df),p=
Parent data (*n* = 228)						
Bi factor 1/ 2-factor	1779.849(1378),0.000	.036	.962	.960		
Bi factor 2/ 3-factor	1583.543(1341),0.000	.028	.977	.975	.015	159.950(37),0.000
Bi factor 3/ 4-factor	1470.863(1297),0.000	.024	.984	.982	.007	118.936(44),0.000
Bi factor 4/ 5-factor	1364.792(1246),0.010	.020	.989	.987	.005	111.689(51),0.000
Teacher data (*n* = 205)						
Bi factor 1/ 2-factor	1665.613(1378),0.000	.032	.984	.983		
Bi factor 2/ 3-factor	1523.194(1341),0.000	.026	.990	.989	.006	112.986(37),0.000
Bi factor 3/ 4-factor	1431.659(1297),0.005	.023	.993	.992	.003	87.683(44),0.000

Bi factor 4/   1340.272(1246),0.000   .019   .995   .994   .002   95.157(51),0.000

5-factor

Note. RMSEA=root mean square error of approximation; CFI= comparative fit index; NNFI=non-normal fit index. Models with same number of latent factors have identical fitting results. e.g. Bi-factor with 1 sub-factor and 2-factor model have identical fitting results. The bi-factor equivocal to the first order factor number is presented to aid clarity.

Table 3

Fittings Indices of Configural Invariance ESEM Models with Different Factors (Parent and Teacher data)

Factor	χ^2(df),p=	RMSEA	CFI	NNFI	ΔCFI	$\Delta\chi^2$(df),p=
Parent data (n = 228)						
Bi factor 1/ 2-factor	1810.496(1314),0.000	.041	.953	.949		
Bi factor 2/ 3-factor	1572.846(1251),0.000	.034	.970	.965	.017	213.682(63),0.000
Bi factor 3/ 4-factor	1440.295(1185),0.000	.031	.976	.971	.006	149.031(66),0.000
Bi factor 4/ 5-factor	1296.052(1116),0.000	.027	.983	.978	.007	153.579(69),0.000
Teacher data (n = 205)						
Bi factor 1/ 2-factor	1631.708(1314),0.000	.034	.982	.981		
Bi factor 2/ 3-factor	1477.601(1251),0.000	.030	.987	.986	.005	144.265(63),0.000
Bi factor 3/ 4-factor	1348.997(1185),0.000	.026	.991	.989	.004	127.236(66),0.000

Bi factor 4/ 5-factor	1224.828(1116),0.000	.022	.994	.992	.003	123.715(69),0.000

Note. RMSEA=root mean square error of approximation; CFI= comparative fit index; NNFI=non-normal fit index. .Models with same number of latent factors have identical fitting results. e.g. Bi-factor with 1 sub-factor and 2-factor model have identical fitting results. The bi-factor equivocal to the first order factor number is presented to aid clarity.

Table 4

Model Fit Indices of Longitudinal ME/I Modelling

Model	χ^2(df),p=	RMSEA	CFI	NNFI	ΔCFI	Δχ^2(df),p=
Parent data (n = 228)						
Configural	1810.496(1314),0.000	.041	.953	.949		
Threshold	2003.181(1450),0.000	.041	.948	.948	-.005	251.801(136),0.000
Teacher data (*n* = 205)						
Configural	1631.708(1314),0.000	.034	.982	.981		
Threshold	1913.446(1445),0.000	.040	.974	.974	-.008	313.333(131),0.000

Note. RMSEA=root mean square error of approximation; CFI= comparative fit index; NNFI=non-normal fit index.

Table 5

*Model Fit Indices of ME/I Modelling between **Arms** at/cross Follow-up Time*

Model		χ^2(df),p=	RMSEA	CFI	NNFI	ΔCFI	Δχ^2(df),p=
Parent data (*n* = 228)							
Configural		3090.206(2628),0.000	.039	.959	.955		
Threshold	A*	3270.624(2887),0.000	.034	.966	.966	.007	277.583(259),0.204
	B*#	3418.395(2934),0.000	.038	.957	.958	-.008	371.107(47),0.000
Teacher data (*n* = 205)							
Configural		2870.454(2628),0.001	.030	.987	.986		
Threshold	A*	3098.046(2890),0.004	.027	.989	.989	.002	279.089(262),0.224
	B*#	3227.577(2940,0.000)	.031	.985	.985	-.004	211.093(50), 0.000

A*: equal parameters between arms at each time.. B*# threshold equal between arm and across time

Table 6

Model Fit Indices of ME/I Modelling between informants at/cross Follow-up Time (n = 244)

Model		χ^2(df),p=	RMSEA	CFI	NNFI	ΔCFI	Δχ^2(df),p=
Configural		6014.139(5508),0.000	.019	.977	.976		
Threshold	A^1	7104.244(5830),0.000	.030	.943	.943	-.033	1022.319(322),0.000
	A^2	6869.555(5780),0.000	.028	.951	.951	-.026	889.799(272),0.000
	A^3	6922.264(5812),0.000	.028	.950	.950	-.027	861.603(304),0.000

A1 threshold equal between informants at each time

A2 freed 20 threshold parameters for teacher and parents rating, equal loading within informants

A3 freed 20 threshold parameters freed for teacher and parents rating, equal loading between informants and across time

Table 7

Association between SNAP-IV scores and diagnosis from DAWBA and clinician for

ADHD/Hyperkinetic disorder

SNAP-IV scales	DAWBA ICD-10 HKD present	DAWBA DSM-5 ADHD present	Clinical diagnosis ADHD present
	OR(95%CI),p=	OR(95%CI),p=	OR(95%CI),p=
Parent	*n = 165*	*n = 166*	*n = 120*
Inatt - p	2.81(1.42-5.56), p=0.003	3.37(1.67-6.79),p=0.001	1.65(0.75-3.64),p=0.208
Hyp / imp - p	3.75(2.07-6.79),p=0.000	4.31(2.36-7.91),p=0.000	1.92(0.99-3.69),p=0.052
Teacher	*n = 153*	*n = 154*	*n = 109*
Inatt - t	0.95(0.56-1.59),p= 0.840	0.99(0.57-1.70),p=0.981	1.38(0.65-2.92),p=0.395
Hyp/ imp – t	2.32(1.44-3.76),p=0.001	2.30(1.38-3.82),p=0.001	2.51(1.23-5.15),p=0.011

Note: Clinical diagnosis represents the diagnosis recorded on the clinical pro forma. Only definitive 'ADHD present' or 'ADHD not present' diagnoses were included in the analysis. CI = confidence interval. Inatt = inattention. Hyp/imp = hyperactivity/impulsivity. HKD = hyperkinetic disorder. OR= Odds Ratio. P = parent. T = teachers.

Table 8

Diagnostic accuracy of SNAP-IV

	DAWBA ICD-10 HKD present	DAWBA DSM-5 ADHD present	Clinical diagnosis ADHD present
Parent data	**N = 165**	**N = 166**	**N = 120**
ADHD confirmed (%)	**75 (45.5)**	**108 (65.1)**	**97(80.8)**
Sensitivity	66.7%	87.0%	100.0%
Specificity	73.3%	56.9%	4.4%
Positive Predictive Value	67.6%	79.0%	81.5%
Negative Predictive Value	72.5%	70.2%	100%
Teacher data	**N = 153**	**N = 154**	**N = 109**
ADHD confirmed (%)	**78 (51.0)**	**106 (68.8)**	**86 (78.9)**
Sensitivity	66.7%	90.6%	96.5%
Specificity	61.3%	31.3%	26.1%
Positive Predictive Value	64.2%	74.4%	83.0%
Negative Predictive Value	63.9%	60.0%	66.7%

Note. CI = confidence interval. DAWBA = Development and Well Being Assessment. HKD = hyperkinetic disorder.