Measuring autism in males and females with a differential item functioning approach: Results from a nation-wide population-based study

Caroline Mårland a,⁎, Thomas Nilsson a, Henrik Larsson b, c, Christopher Gillberg d, Gitta Lubke e, Sebastian Lundström a, d

a Centre for Ethics Law and Mental Health (CELAM), Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Rågården hus 1, SU-Ostra Sjukhuset, Gothenburg 416 85, Sweden
b Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
c School of Medical Sciences, Örebro University, Örebro, Sweden
d Gillberg Neuropsychiatry Centre, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Sweden
e Department of Psychology, University of Notre Dame, Notre Dame, USA

⁎ Corresponding author.
E-mail address: caroline.marlund@vgregion.se (C. Mårland).

https://doi.org/10.1016/j.psychres.2022.114674
Received 8 March 2022; Received in revised form 30 May 2022; Accepted 4 June 2022
Available online 8 June 2022

0165-1781/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

ARTICLE INFO

Keywords:
A-TAC
Screening
Sex
Autism spectrum disorder
Item response theory

ABSTRACT

Existing screening instruments for Autism Spectrum Disorder (ASD) might be prone to detect a male manifestation of ASD. Here, we examined the 17 items from the ASD domain in the Autism–Tics, ADHD and other Comorbidities inventory (A–TAC) for Differential Item Functioning (DIF). Data were obtained from the Child and Adolescent Twin Study in Sweden (CATSS) in which parents have responded to the A–TAC. Information regarding a registered diagnosis of ASD were retrieved from the National Patient Register. The cohort was divided into a developmental sample for evaluation of DIF, and a validation sample for examination of the diagnostic accuracy of the total ASD domain, and a novel male and female short form. Our main finding included the identification of DIF for six items, three favouring males and three favouring females. The full, 17 item, ASD domain and the male and female short form showed excellent ability to capture ASD diagnoses in both males and females up to the age of nine years. The full ASD domain in A–TAC is psychometrically largely equivalent across sex and the limited differences between males and females diminish the need for a sex-specific scoring when utilizing the 17 item total score.

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder and the core diagnostic criteria include persistent deficits in social communication and restricted, repetitive patterns of behaviour according to the Diagnostic and Statistical Manual of Mental Disorders – fifth edition (DSM-5) (American Psychiatric Association, 2013). A commonly replicated finding in literature is the male preponderance in ASD and the most recent meta-analytic study reported a male-female ratio of 3:1 (Loomes et al., 2017). It has been argued that the overrepresentation of males to some extent is related to under identification of females with ASD (Lai et al., 2015; Loomes et al., 2017) which could be attributed to the current diagnostic procedures that may be less prone to capture ASD traits in females (Halladay et al., 2015; Kreiser and White, 2014; Ratto et al., 2018).

The manifestation of ASD might be different, or less pronounced, in females. Several large-scale studies have reported lower levels of restricted and repetitive behaviours in females on clinical instruments such as the Autism Diagnostic Observational Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R) (Charman et al., 2017; Frazier et al., 2014; Mandy et al., 2012; Wilson et al., 2016). Previous studies regarding social communication have revealed inconsistent results; a meta-analysis reported no differences in social communication between males and females with ASD (Wijngaarden-Cremers et al., 2014) while other studies have reported higher levels of deficits in social communication for females with ASD (Evans et al., 2019; Frazier et al., 2014). The possible differences in symptomatology between males and females is likely influenced by several factors such as age, IQ, and language level. Kaat et al. (2021) took these factors into account and examined sex-differences in the ADOS, ADI-R, and Social Responsiveness Scale (SRS) in a sample of children with a clinical diagnose of ASD. The result showed higher raw scores in boys regarding restricted and
repetitive behaviours in the ADOS and ADI-R. In the SRS, a difference emerged in adolescence and girls received more severe scores than boys in both the domain of restricted and repetitive behaviours and social communication. However, the effect sizes were small and the clinical significance was considered to be minimal.

In a recent article by Lundström et al. (2019), raw and standardized scores were examined for males and females separately on the ASD domain in the Autism–Tics, ADHD and other Comorbidities inventory (A–TAC). The result showed higher raw scores for boys with a registered diagnosis of ASD. However, the opposite pattern emerged by the use of sex-standardized scores when girls with a registered diagnosis of ASD deviated further away from the sex-specific mean compared to boys on continuous scores of ASD. The reported differences between males and females might represent true differences in mean levels of ASD traits, but it may also reflect that different constructs are being measured across sex. Taken together, the suggested differences between males and females in the core symptomatology of ASD highlights the need to evaluate existing instruments for measurement equivalence (or invariance) which indicates that the same latent construct is measured across sex. This becomes particularly important for ASD since most instruments are based on the diagnostic criteria in DSM-IV in which the original field trials were based on a sample with a reported male-female ratio of 4.5:1 for individuals with a clinical diagnosis of ASD (Volkmar et al., 1994).

Measurement equivalence can be assessed by an item-level analysis utilizing an item response theory approach (IRT) (Osterlind and Sterba, 2009). Thus, an item is equivalent when males and females with the same level of the underlying latent trait (e.g. ASD traits) have the same probability to endorse an item. Differential Item Functioning (DIF) arises when one group, e.g. females, systematically do not endorse items that reflect behaviours that are commonly endorsed by males. Investigations of measurement equivalence in screening instruments for ASD have been scarce. To the authors’ knowledge, five previous studies have examined DIF across sex with an IRT approach in screening instruments for ASD. Measurement equivalence was established for the total scores in all of the investigated instruments; the Autism Spectrum Quotient – 10 items (AQ-10) (Murray et al., 2017; Murray et al., 2019), Social Communication Questionnaire – lifetime version (SCQ) (Wei et al., 2014) and the Social Responsiveness Scale (Frazer and Hardan, 2017; Sturm et al., 2017). However, several items in all of the scales revealed a DIF.

The ASD domain in the A–TAC includes 17 items which have been used for screening purposes in the Child and Adolescent Twin study in Sweden (CATSS) for almost two decades. The purpose of the present study is to further expand our understanding of possible differences in ASD traits across sex by evaluating the full, 17 item, ASD domain in A–TAC for potential DIF. More specific, the present study aims to (1) examine each item in the full ASD domain for DIF in a nation-wide population of twins, (2) investigate if a subset of items have a higher ability to detect ASD traits in males and females, (3) examine the diagnostic accuracy in the selected subset of items, i.e. a separate male and female short form, compared to the full, 17 item, ASD domain.

2. Methods

In this study, data was collected from two sources: CATSS and the National Patient Register (NPR) in Sweden.

2.1. The Child and Adolescent Twin Study (CATSS)

The CATSS is a population based longitudinal study which aims to assess somatic and mental health problems during childhood (for a detailed over-view see Anckarsäter et al., 2011). Since 2004, twins are identified through the Swedish Twin Registry and the parents are invited to participate in a telephone interview in connection with the twins’ 9th birthday (12-year-old twins were included during the first three years of CATSS). The response rate in CATSS is ≈ 70% and small differences in prevalence of neurodevelopmental disorders have been reported between responders and non-responders when compared in the NPR. The prevalence of ASD has been estimated to be 0.95% in non-responders compared to 0.84% in responders, 2.1% had ADHD versus 0.84% and 2.0% LDs versus 0.99% (Anckarsäter et al., 2011).

2.2. The National Patient Register (NPR)

The NPR in Sweden includes information about assigned diagnoses in psychiatric inpatient care since 1987 and from outpatient clinics since 2001. The ascribed diagnoses are assigned according to the International Classification of Diseases ninth (ICD-9) and tenth (ICD-10) revision (World Health Organization, 1980, 1992). Data from CATSS was merged with the NPR by using a personal identification number which all individuals in Sweden are assigned at birth or when receiving a Swedish citizenship. Diagnostic data, including the date of the first diagnosis, were retrieved by searching for ICD-9 and ICD-10 codes that correspond to an ASD diagnosis (299 (ICD-9) and F84.0, F84.1, F84.5, F84.8 and F94.9 (ICD-10)). The validity of the ASD diagnoses in NPR have previously been evaluated and an agreement of 96% between medical records and registered diagnoses of ASD have been reported when several medical registers were compared (Idring et al., 2012).

2.3. The Autism–Tics, ADHD and other Comorbidities inventory (A–TAC)

The A–TAC is a structured and comprehensive parental interview, utilized as a telephone interview by laymen in CATSS. It consists of 96 items which are divided into theoretically defined modules. The ASD domain includes 17 items from three modules; Language (6 items (module H)), Social interaction (6 items (module I)) and Flexibility (5 items (module J)). The items are based on the diagnostic criteria for pervasive developmental disorder in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) (American Psychiatric Association, 1994), clinical expertise, and relevant aspects that have been captured in other instruments: the Asperger Syndrome Screening Questionnaire (Ehlers and Gillberg, 1993; Ehlers et al., 1999), the Asperger Syndrome Diagnostic Instrument (Gillberg et al., 2001), and the Five to Fifteen questionnaire (Kadesjö et al., 2004). Items are assessed in a “whole-life” frame and each module begins with a reminder that “The essential aspect of each question is whether the problem/-peculiarity has been pronounced compared to peers during any period of life”. The items are coded on a dimensional scale with three response categories: “No” (0), “Yes, to some extent (0.5)” and “Yes” (1). Two cut-off values have previously been established, a low cut-off (>4.5) with high sensitivity for screening purposes and a high cut-off value (>8.5) with high specificity that can be used as a clinical proxy. Furthermore, a short form of the ASD domain have previously been established using IRT in the CATSS-sample. The short form includes four items (H35, I40, I44 and J47) with high discriminatory ability in the far end of the ASD trait continuum (Mårländ et al., 2019).

The full, 17 item, ASD domain have previously been validated in cross-sectional and longitudinal studies. The cross-sectional studies have reported an Area under the Receiver Operating Characteristics Curve (AUC) between 0.88 and 0.96 (Hansson et al., 2005; Larson et al., 2010). The corresponding AUC in the longitudinal studies ranged from 0.81 to 0.91 (Larson et al., 2013; Mårländ et al., 2017). In addition, the psychometric properties of the previously established short form have been reported to be in agreement with the validation studies of the full ASD domain (Mårländ et al., 2019). Furthermore, an independent research group in Spain has validated the full, 17 item, ASD domain in Spanish which showed excellent psychometric properties (Cubó et al., 2011). Regarding reliability, one study has examined the test-retest intraclass correlation and reported a value of 0.84. The x-value was 0.59 for the low cut-off and 1.0 for the high cut-off value in the full ASD domain (Larson et al., 2014).
2.4. Definition of the samples

The total sample from CATSS included 34 033 subjects, (27 541 were 9-year-old and 6492 were 12-year-old when the A–TAC interview was conducted). A registered diagnosis of ASD were reported for 2.4% males and 1.1% females, giving a male: female ratio of 2.26:1.

The total sample was randomly divided in two samples; a developmental sample that was used in order to conduct the DIF analyses and a validation sample that was used to calculate the previous and predictive validity. The developmental sample included 16 631 individuals, 8359 males (80.5% 9-year-olds) and 8270 females (81.4% 9-year-olds). A registered diagnosis of ASD were reported for 220 males and 96 females.

The mean age at diagnosis were 11.07 years (SD = 4.8) for males compared to 12.84 years for females (SD = 4.7). For both males and females, a registered diagnosis of autistic disorder (F84.0, 44.81% of the males and 37.5% of the females) was registered in most cases together with Asperger’s syndrome (F84.5, 33.6% and 38.5%, respectively). A co-occurring diagnosis of intellectual disability was reported for 38 males and 13 females.

The validation sample included 16 719 individuals, 8517 males (80.8% 9-year-olds) and 8202 females (80.9% 9-year-olds). A registered diagnosis of ASD were reported for 186 males and 83 females. The mean age at diagnosis were 11.52 years (SD = 4.8) for males compared to 12.14 years (SD = 4.9) for females. For both males and females, a registered diagnosis of autistic disorder (F84.0, 47.8% of the males and 44.6% of the females) was registered in most cases together with Asperger’s syndrome (F84.5, 29.6% and 22.9%, respectively). A co-occurring diagnosis of intellectual disability was reported for 38 males and 12 females.

2.5. Statistical analyses

2.5.1. Preliminary analysis

The item response frequencies were examined for males and females in the total sample (see supplementary material: Additional file 1). As expected, the distribution of responses were negatively skewed and few subjects have endorsed the higher response categories (i.e. “Yes, to some extent” and “Yes”). Overall, more males than females had endorsed the higher response categories. Since some items had a response frequency of < 1% in the highest response category (males: 142; females: H35, H37, I40, I42, J47), we decided to collapse the higher response categories for parsimony. Subsequently, a binary version of the full, 17 item, ASD domain was used for in the calculation of DIF across sex.

2.5.2. Unidimensionality

IRT requires unidimensionality, i.e. that a single underlying trait accounts for the vast majority of the covariance among the items in a scale. The full, 17 item, ASD domain was therefore evaluated for unidimensionality by an exploratory factor analysis (EFA) with principal axis factoring and a promax rotation in order to take account for the correlation between items. The analyses were conducted for males and females separately with a randomly selected sample of approximately 2% of males (N = 345) and females (N = 338). The subjects included in the EFA were excluded in the rest of the statistical analyses in this paper.

The factor extraction was determined by an examination of the scree plot. For both males and females the scree plot indicated that one factor should be retained. The scree plot and the factor loadings are available as supplementary material (Supplementary material: Additional file 2). The result indicated that the full, 17 item, ASD domain were sufficiently unidimensional. Therefore, all the 17 items were used to fit the IRT models which also provides the possibility to investigate if the total score can be considered to be equivalent across sex.

2.5.3. Differential item functioning

The developmental sample was used to fit a two parameter logistic model (2PL) with two groups: males (reference group) and females (focal group). The 2PL model yields a trace line which is defined as a logistic function of the relationship between a subject’s response to a specific item and the subject’s level of the underlying trait (here: severity of ASD traits). The trace line establish two parameters: difficulty and discrimination. The difficulty parameter, denoted as b, represents the point on the latent trait continuum were the probability of giving an affirmative response is 50%. Thus, a higher b indicates that a subject must have a higher level of the underlying trait in order to give an affirmative response. The discrimination parameter, denoted a, identifies the slope of the trace line that corresponds to the difficulty parameter. A higher a indicates that an item has a better ability to discriminate between subjects within a small range of the underlying trait where b=0.

Three different 2PL models, a constrained model, a DIF-difficulty model, and a DIF-discrimination model, were fitted in order to separately examine if the difficulty (b) and discrimination (a) parameter were equal across sex (i.e. indicating measurement equivalence). Constraints were specified by utilizing the slope-intercept parameterization. In a first step, the constrained model was fitted with the slopes and intercepts constrained to be equal across groups. In the next step, a DIF-difficulty model was fitted for each item where the difficulty parameter b was allowed to vary while constraints were set to the slope parameter a. Secondly, the DIF-discrimination model were fitted for each item where the slope were allowed to vary while the intercept was constrained to be equal across groups. Both models were tested with a likelihood ratio test (LRT) to examine whether the difficulty or discrimination parameter of each item differed across sex. The result from the LRT is distributed as a chi-square variable where a significant result indicate the presence of DIF. The Akaike information Criterion (AIC) and Bayesian information criterion (BIC) were used to compare the models since minor differences between the models can be flagged as statistically significant in large samples when using LTR. A better fit is indicated for the model with the lowest AIC and BIC values. According to Raftery (1995) between-model differences in BIC of 0-2 can be categorized as "weak", 2-6 "Positive", 6-10 “Strong”, and >10 “Very strong” evidence in favour of the model with the better fit.

2.5.4. Selection of items: male and female short form

The item information functions (IIF) for each item in the DIF-difficulty model and DIF-discrimination model were examined in order to select a subset of items as candidates for a separate male and female short form. The IIF is dependent on the item parameters and illustrate how much information an item provide and the range of the underlying trait where the item provide the most precise measurement. The IIF is inversely proportional to the error of measurement, therefore IIF indicate the range where the error of measurement is low and the precision of the item is at its best (Carlson, 2020).

Based on the results from the DIF analysis in the developmental sample, a separate male and female short form were developed. A subset of the items with high and narrow IIFs were selected in combination with items flagged with DIF in favour of either males or females. Thus, only items with DIF in favour of males were included in the male short form and vice versa.

2.5.5. Validation

The validation sample was used to establish the diagnostic accuracy in the separate male and female short form compared to the full, 17 Item, ASD domain and previously established short form. Receiver Operating Characteristic curves (ROC) were calculated in order to examine the Area Under the Curve (AUC). The AUC is a measure of an instruments ability to discriminate between the presence or absence of a disorder that also provides sensitivity and specificity values for all possible steps on a continuous scale. Furthermore, the AUC can be used to illustrate the validity of an instrument. An AUC of 0.5 signals random prediction, 0.60–0.70 poor validity, 0.70–0.80 is fair, 0.80–0.90 is good, and AUC >0.90 represents excellent validity (Tape, 2004). The full, 17 item, ASD domain, previously established short form, and the separate male and
female short form were used as independent predictors while the ASD diagnoses from NPR was used as dependent variable. All analyses in the validation sample were stratified by age. The sample was divided into two groups: previous (a registered ASD diagnosis in the NPR before or the same year as the A-TAC interview) and predictive (a registered ASD diagnosis in the NPR after the A-TAC interview). The analyses were also conducted in a total group with no consideration taken to the age of diagnosis.

3. Results

3.1. Differential item functioning

The difficulty and discrimination parameter for each item in the DIF-difficulty model are presented in Table 1 together with results from the comparison with the constrained model utilizing the LRT, AIC, and BIC. In total, eight items were flagged with DIF after the LRT. Item H34 (Was his/her language development delayed or does s/he not speak at all?), H37 (‘Does he/she have difficulties with games of make-believe or does he/she imitate others considerably less than other children?’) and J46 (‘Does s/he get absorbed by his/her interests in such a way as being repetitive or too intense?’) was flagged with DIF in favour of males. Similarly, five items favoured females; H38 (‘Does s/he talk in too high a pitch or too quietly?’), H39 (‘Does s/he have difficulties keeping “on track” when telling other people something?’), I43 (‘Can s/he only be with other people on his/her terms?’), J47 (‘Does s/he get absorbed by routines in such a way as to produce problems for himself or for others?’), and J49 (Does s/he get absorbed by details?). However, the result from the AIC and BIC for items J47 and J49 indicated a better fit for the model with parameters constrained to be equal across groups Therefore, item J47 and J49 were not considered to show a significant amount of DIF. Taken together, three items, H34, H37, and J46, showed significant DIF in favour of males, i.e. males have a higher probability of an affirmative response compared to women with the same level of the underlying latent trait. Similarly, three items, H38, H39, and I43, showed significant DIF in favour of females.

The difficulty and discrimination parameter for each item in the DIF-discrimination model are presented in Table 2 together with comparison with the constrained model utilizing the LRT, AIC, and BIC. The results from the LRT flagged the same eight items listed above with DIF. However, the results from the AIC and BIC only indicated a significant amount of DIF in three items. A significantly better ability to discriminate in males were found in item J46 while item H38 and I43 discriminated better in females.

3.2. Item selection: male and female short form

Items with narrow and moderately high or high IIF in the upper end of the continuum (i.e. severity of autism traits) were found in all three modules: H35 in the language module, I40, I41, and I44 in the social interaction module and J47 and J49 in the flexibility module. All items with high IIF did not show significant DIF. Four of these items (H35, I40, I44, and J47) are included in the previously established short form (Mårland et al., 2019).

A separate male and a female short form were developed based on the results above, in order to evaluate the possibility of better detection of ASD across sex. The four items included in the previously established short form was used as a base for both the male and female short form since these items provided high discrimination in the far end of the ASD trait continuum. The male short form included the three items with DIF in favour of males (H34, H37, and J46) plus the four items without DIF from the previously established short form. Similarly, the female short form included the three items with DIF in favour of females (H38, H39, and I43) and the same four items without DIF.

3.3. Validation

The AUC are reported in Table 3, (see supplementary material: additional file 3 for the sensitivity and specificity values for each possible scale step in the total group for all versions of the full, 17 item, ASD domain). The previous validity was excellent for all versions while the predictive validity was fair with slightly lower AUC estimates for females. The full, 17 item, ASD domain had the highest AUC while the male and female short form yielded a slight increase in AUC compared to the previously established short form.

### Table 1

**Difficulty and discrimination estimates from the DIF-difficulty model together with the DIF analysis utilizing the likelihood ratio test, Akaike information criterion, and Bayesian information criterion.**

| Item number and item content | Male | SE | Male | SE | Female | SE | Female | SE | LRT $\chi^2$ | $p$ | AIC | BIC |
|------------------------------|------|----|------|----|--------|----|--------|----|-------------|----|------|------|
| H34 delayed language development? | 1.10 | 0.04 | 2.19 | 0.07 | 1.10 | 0.04 | 2.72 | 0.10 | 89.29 | <0.001 | 87.3 | 79.6 |
| H35 difficulties sustaining a conversation? | 2.64 | 0.10 | 2.06 | 0.04 | 2.64 | 0.10 | 2.08 | 0.05 | 0.17 | 0.68 | -1.9 | -9.5 |
| H36 repeat words and expressions / use words in a strange way? | 2.00 | 0.07 | 2.04 | 0.05 | 2.00 | 0.07 | 2.04 | 0.05 | 0.01 | 0.94 | -2 | -9.7 |
| H37 difficulties with games of make-believe / less imitation | 1.80 | 0.07 | 2.21 | 0.06 | 1.80 | 0.07 | 2.42 | 0.07 | 15.69 | <0.001 | 13.7 | 6 |
| H38 talk in too high a pitch or too quietly? | 1.21 | 0.04 | 2.17 | 0.06 | 1.21 | 0.04 | 1.85 | 0.06 | 44.44 | <0.001 | 42.4 | 34.7 |
| H39 difficulties keeping “on track” when telling other people something? | 1.72 | 0.05 | 1.68 | 0.04 | 1.72 | 0.05 | 1.53 | 0.04 | 18.70 | <0.001 | 16.7 | 9 |
| I40 difficulties expressing emotions and reactions? | 2.92 | 0.13 | 2.25 | 0.05 | 2.92 | 0.13 | 2.25 | 0.06 | 0.00 | 0.97 | -2 | -9.7 |
| I41 difficulties interacting with peers? | 3.56 | 0.13 | 1.57 | 0.03 | 3.56 | 0.13 | 1.54 | 0.03 | 1.78 | 0.18 | -0.3 | -7.9 |
| I42 uninterested in sharing joy, interests, and activities? | 2.67 | 0.11 | 2.21 | 0.05 | 2.67 | 0.11 | 2.21 | 0.06 | 0.01 | 0.93 | -2 | -9.7 |
| I43 be with other people on his/her terms? | 2.66 | 0.09 | 1.68 | 0.03 | 2.66 | 0.09 | 1.47 | 0.03 | 47.66 | <0.001 | 45.6 | 37.9 |
| I44 difficulties behaving as expected? | 3.54 | 0.13 | 1.47 | 0.03 | 3.54 | 0.13 | 1.44 | 0.03 | 1.09 | 0.30 | -0.9 | 8.6 |
| I45 Easily influenced by others? | 1.09 | 0.03 | 1.09 | 0.04 | 1.09 | 0.03 | 1.02 | 0.04 | 3.52 | 0.06 | 1.5 | -6.2 |
| J46 Absorbed by interests, repetitive or too intense? | 1.89 | 0.06 | 1.32 | 0.03 | 1.89 | 0.06 | 1.79 | 0.04 | 196.73 | <0.001 | 194.7 | 187 |
| J47 Absorbed by routines? | 2.62 | 0.10 | 1.98 | 0.04 | 2.62 | 0.10 | 1.90 | 0.05 | 4.30 | 0.04 | 2.3 | -5.4 |
| J48 strange hand movements/toe-walking? | 1.47 | 0.06 | 2.59 | 0.08 | 1.47 | 0.06 | 2.64 | 0.09 | 0.81 | 0.37 | -1.2 | 8.9 |
| J49 obsessed with details? | 2.29 | 0.07 | 1.38 | 0.03 | 2.29 | 0.07 | 1.30 | 0.03 | 7.79 | 0.01 | 5.8 | -1.9 |
| J50 dislike changes in daily routines? | 2.00 | 0.06 | 1.26 | 0.03 | 2.00 | 0.06 | 1.30 | 0.03 | 1.75 | 0.19 | 0.3 | -8.8 |

Note. N = 8359 males & 8270 females. a = item discrimination, b = item difficulty, SE = standard error, LRT $\chi^2$ = Likelihood-ratio test, AIC = Akaike information criterion, BIC = Bayesian information criterion. Item number and item content in italics denotes inclusion of item in the previously established short form.
Table 2

Difficulty and discrimination estimates from the DIF-discrimination model together with the DIF analysis utilizing the likelihood ratio test, Akaike information criterion, and Bayesian information criterion.

| Item | Male | Female | SE Male | Female | LRT $\chi^2$ | p | AIC* | BIC* |
|------|------|--------|---------|--------|-------------|---|------|------|
| H34 delayed language development? | 1.23 | 0.05 | 2.18 | 0.07 | 1.03 | 0.06 | 2.61 | 0.14 | 9.93 | 0.001 | 7.9 | 0.2 |
| H35 difficulties sustaining a conversation? | 2.65 | 0.10 | 2.06 | 0.04 | 2.62 | 0.12 | 2.08 | 0.06 | 0.14 | 0.71 | -1.9 | -0.6 |
| H36 repeat words and expressions / use words in a strange way? | 2.02 | 0.07 | 2.01 | 0.05 | 1.94 | 0.09 | 2.10 | 0.07 | 1.22 | 0.27 | -0.8 | -0.8 |
| H37 difficulties with games of make-believe / less imitation | 1.91 | 0.07 | 2.18 | 0.06 | 1.66 | 0.09 | 2.50 | 0.11 | 10.21 | 0.001 | 8.2 | 0.5 |
| H38 talk in too high a pitch or too quietly? | 1.07 | 0.05 | 2.27 | 0.08 | 1.38 | 0.06 | 1.76 | 0.07 | 21.51 | <0.001 | 19.5 | 11.8 |
| H39 difficulties keeping “on track” when telling other people something? | 1.66 | 0.06 | 1.67 | 0.04 | 1.77 | 0.07 | 1.57 | 0.05 | 2.48 | 0.12 | 0.4 | -7.2 |

Note. N = 8359 males & 8270 females. a = item discrimination, b = item difficulty, SE = standard error, LRT $\chi^2$ = Likelihood-ratio test, AIC = Akaike information criterion, BIC = Bayesian information criterion. Item number and item content in italics denotes inclusion of item in the previously established short form.

The reported AIC and BIC values represents the difference between the constrained model and the DIF-difficulty model. Thus, negative values indicate a better fit of the constrained model.

Table 3

Area under the Curve (AUC) for males and females in the previous, predictive and total group for the full ASD domain, previously established short form, and the male and female short form.

| Item | Previous - AUC | Predictive - AUC | Total - AUC |
|------|----------------|-----------------|-------------|
|      | $d$            | $\hat{\varphi}$ | $d$         | $\hat{\varphi}$ | $d$         | $\hat{\varphi}$ |
| ASD domain | 0.96 | 0.98 | 0.79 | 0.76 | 0.86 | 0.83 |
| Short form | 0.94 | 0.93 | 0.75 | 0.71 | 0.83 | 0.77 |
| Male short form | 0.95 | 0.97 | 0.78 | 0.75 | 0.85 | 0.81 |
| Female short form | 0.94 | 0.97 | 0.77 | 0.74 | 0.84 | 0.81 |

Note. N $d$ = 8517 (ASD = 186), N $\hat{\varphi}$ = 8202 (ASD = 83). Previous group: a registered ASD diagnosis in NPR before, or the same year, as the A-TAC interview. Predictive group: a registered ASD diagnosis after the A-TAC interview. Total group: previous and predictive group collapsed. Please see the supplementary material: additional file 3 for the sensitivity and specificity values for each scale step in the ASD domain for the total group.

4. Discussion

The primary aim of this study was to evaluate the full, 17 item, ASD domain in A-TAC for item-level equivalence across sex in a large population-based sample. Our main finding included the identification of DIF in six items, three items favouring males and three favouring females. The four items included in the previously established short form showed item-level measurement equivalence and the items with the best ability to capture ASD in both males and females. An excellent ability to capture ASD diagnoses in males and females were reported for the full, 17 item, ASD domain as well as for the previously established short form and the male and female, seven item, short form.

The highest AUC estimates were reported for the full ASD domain, where item-level measurement equivalence was found in 11 of the 17 items. The total score of the full ASD domain is expected to be largely equivalent across sex since the majority of items were invariant and the six items with DIF were equally distributed across sex. This is an important finding speaking against any systematic measurement error, and rather indicating that the observed differences in mean scores across sex reflect a meaningful variation in the severity of ASD traits. This is in line with previous research that have found a limited clinical significance for the difference in total scores between males and females with ASD (Kaat et al., 2021; Evans et al., 2019). The suggested sex-specific short forms did not contribute to an improved diagnostic accuracy in males and females, compared to the full, 17 item, ASD domain or the previously validated short form. However, the previous validity was excellent for all the included forms, which could indicate that the ASD phenotype captured by the A-TAC may differ depending on the age at diagnosis, i.e. individuals with a more severe symptomatology are identified at an earlier age while a subler manifestation of ASD could delay treatment starting.

The identification of items with DIF highlights the need for caution in single item-level interpretation as well as the importance of psychometric evaluation of existing instruments during both revisions and development of new instruments. For example, item J46 “Does s/he get absorbed by his/her interests in such a way as being repetitive or too intense?” favoured males ($d = 1.32$, $\hat{\varphi} = 1.79$). The opposite pattern emerged for item H38 “Does s/he talk in too high a pitch or too quietly?” which showed significant DIF in favour of females ($d = 2.17$, $\hat{\varphi} = 1.85$). The explanations for DIF may vary substantially. For J46, the observed difference may depend on biological underpinnings since a higher liability threshold have been reported for females regarding restrictive and repetitive behaviour in particular, but also for ASD symptoms in general (Sztamari et al., 2012). However, several studies have reported lower levels of restricted and repetitive behaviour in females with ASD and it has been suggested that the For H38, different voice or speech have previously been highlighted by Kopp and Gillberg as a possible female-specific item (i.e. only endorsed by girls) in the Autism Spectrum Screening Questionnaire, Revised Extended Version (ASSQ-REV). Furthermore, a robot-like language were considered to reflect a more male manifestation of ASD which in turn could be more easily recognized than unusual voice or speech (Kopp and Gillberg 2011). In contrast, Frazier and Hardan, reported a higher difficulty threshold for females on an item in SRS concerning a child whose voice volume on others (Frazier and Hardan, 2017). Thus, it is possible that, in this instance, the wording could have affected the given response to H38.

Noticeably, five of the items flagged with DIF asks about impairment in language or social interaction. Item H74 concerning delayed language development favoured males and showed the highest DIF across sex ($d = 2.19$, $\hat{\varphi} = 2.72$). This difference might be of clinical importance since
it could be one of the first concerns raised by the parents (Chawarska et al., 2007) and may also be a more noticeable marker for referral to diagnostic assessment. The four remaining items (H37, H38, H39, and I43) showed a smaller amount of DIF across sex. Given our large sample size, this limited amount of DIF may not be practically relevant but highlight the need for careful item-writing.

It has been suggested that females with ASD could present with a different phenotype not entirely captured by existing screening instruments. For example, females with ASD may show restricted interests that are more socially acceptable (i.e. books or animals compared to trains and rocks). These interests are less likely to be perceived as a fixation which increase the risk for under identification of females with ASD (Hiller et al., 2014, 2016; Mandy et al., 2012; Sutherland et al., 2017). A related aspect is camouflaging (i.e. the use of coping strategies to fit in) which is reported to be more common in females without intellectual disability and has been proposed to contribute to the under identification of females with ASD (Hull et al., 2020; Hull et al., 2017; Lai et al., 2017). Furthermore, the choice of responders may influence the way the differences are captured by a screening instrument. It has been reported that subtle social difficulties in females may go unnoticed by the parents during childhood while these deficits become more evident during adolescence with the increase of social demands (Mandy et al., 2018). Additionally, it has been suggested that both parents and teachers may perceive the symptomatology manifested by females as less impairing (Posserud et al., 2018). During childhood, females with ASD have been reported to show an ability to maintain a reciprocal conversation, initiate friendship, integrate verbal and non-verbal behaviour, and engage in imaginative play. Therefore, females with ASD may be considered to be less impaired on a surface level (Hiller et al., 2014). The possible sex differences may also vary with age, and the differences in the manifestation of ASD might be more pronounced during adolescence (Jamison et al., 2017). On that account, it is possible that the female manifestation of ASD is not adequately captured by a parent rated screening instrument, and other items or follow-up questions, might be required in order to enhance the identification of females with ASD. Furthermore, the change in ASD prevalence over the last decades have been associated with a decrease in the number of symptomatic individuals diagnosed with ASD (Aridsson et al., 2018) and an increase in individuals being assigned a diagnosis in adolescence (Kosidou et al., 2017). It is plausible that screening instruments, such as A–TAC, are not calibrated for a diagnosis that today encompass roughly 5% of the teenage population in Sweden.

To conclude, the majority of the variance in symptom presentation in ASD is shared between males and females. The full, 17 item, ASD domain in A–TAC is largely equivalent and the observed differences in mean scores across sex reflect a meaningful variation in the severity of autistic traits. The limited differences between males and females diminish the need for a sex-specific scoring procedure when utilizing the total score. However, the possible subtle differences between males and females may not be adequately captured during the screening process.

The present study has several strengths including a large nation-wide sample, a best estimate clinical diagnosis of ASD from the NPR and a high response rate. However, the result must be considered in the light of some limitations. First, the items in the full ASD domain were based on the diagnostic criteria of pervasive developmental disorders in the DSM-IV together with clinical expertise and well known aspects that were included in other ASD assessment instruments. It cannot be ruled out that the inclusion of items in the full, 17 item, ASD domain have been influenced by the male preponderance in ASD. Second, the A–TAC interview was conducted by the parents in connection with the twins 9th birthday, or in a minority of the sample, 12th birthday and it is possible that sex differences are more noticeable later in life. Third, register-based ASD diagnoses were used in the validation process. Concerns have been raised over ascertainment bias since females with ASD may be under-recognized and therefore not included in clinically recruited samples. The male-female ratio of 2.26:1 in the present study is similar to the reported ratio in other population-based studies (Kim et al., 2011; Idsinga et al., 2012). Furthermore, the latest meta-analytic study reported a male-to-female ratio of 3:1 (Loomes et al., 2017), which indicate only minor, if any, underrepresentation of females diagnosed with ASD in the present sample from CATSS. Fortth, possible parental bias should be considered since parents may have different social expectations on males and females. Finally, the sample is based on twins. Some studies have suggested an increased risk for ASD in twins (Betancur et al., 2002; Greenberg et al., 2001). However, this has not been established in large-scale epidemiological studies (Croen et al., 2002; Hallmayer et al., 2002; Hultman et al., 2002) or within the CATSS (Lundström et al., 2015).

Ethical considerations

The CATSS and the linkage to the NPR have ethical approval from the Regional Ethical Review Board in Stockholm (DNR 02-289, 2010/597–31/1 and 2016/2135–31).

CRediT authorship contribution statement

Caroline Mårland: Conceptualization, Methodology, Formal analysis, Writing – original draft. Thomas Nilsson: Writing – review & editing. Henrik Larsson: Writing – review & editing. Christopher Gillberg: Writing – review & editing. Gitta Lubke: Methodology, Writing – review & editing. Sebastian Lundstrom: Conceptualization, Methodology, Writing – original draft, Supervision.

Declaration of Competing Interest

The authors declare that they have no competing or potential conflicts of interests.

Acknowledgements

The Child and Adolescent Twin Study in Sweden is supported by Swedish Research Council (Humanities and Social Science) Grant number: 2017-02552 and by funds under the ALF agreement, grant number ALFGBG-776031. This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The funding sources had no involvement in the study design, collection, analysis, and interpretation of the data, preparation of the manuscript, or the decision to submit this manuscript for publication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2022.114674.

References

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV, 4th ed. the Task Force on DSM-IV. prepared byed. American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders: DSM-5, 5th ed. American Psychiatric Association, Arlington, Va. Anczarske, H., Lundström, S., Kollberg, L., Kerekes, N., Palm, C., Carlstrom, E., Langström, N., Magnusson, P.K., Haldåker, L., Bolte, S., Gillberg, C., Gumpert, C., Råström, M., Lichtenstein, P., 2011. The child and adolescent twin study in Sweden (CATSS). Twin Res. Hum. Genet. 14 (6), 495–508.

Aridsson, O., Gillberg, C., Lichtenstein, P., Lundström, S., 2018. Secular changes in the symptom level of clinically diagnosed autism. J. Child Psychol. Psychiatry 59 (7), 744–751. https://doi.org/10.1111/jcpp.12864.

Betancur, C., Leboyer, M., Gillberg, C., 2002. Increased rate of twins among affected siblings with autism. Am. J. Hum. Genet. 70 (5), 1381–1383. https://doi.org/10.1086/340364.

Carlson, J.E., 2020. Introduction to Item Response Theory Models and Applications. Routledge, London.

Charman, T., Loth, E., Tillmann, J., Crawley, D., Woolidge, C., Goyard, D., Ahmad, J., Auyeung, B., Ambrosino, S., Banaschewski, T., Baron-Cohen, S., Baumeister, S., Beckmann, C.F., Bolte, S., Bourgeron, T., Bours, C.C.A.H., Brammer, M., Brandeis, D.,
Ratto, A.B., Kenworthy, L., Yerys, B.E., Bascom, J., Wieckowski, A.T., White, S.W., Wallace, G.L., Pugliese, C., Schultz, R.T., Ollendick, T.H., Scarpas, A., Seese, S., Register-Brown, K., Martin, A., Anthony, L.G., 2018. What about the girls? Sex-based differences in autistic traits and adaptive skills. J. Autism Dev. Disord. 48 (5), 1698–1711. https://doi.org/10.1007/s10803-017-3413-9.

Raftery, A.E., 1995. Bayesian model selection in social research. Sociol. Methodol. 25, 111–163. https://doi.org/10.2307/271065.

Sturm, A., Kuhfeld, M., Kasari, C., McCracken, J.T., 2017. Development and validation of an item response theory-based social responsiveness scale short form. J. Child Psychol. Psychiatry 58 (9), 1053–1061. https://doi.org/10.1111/jcpp.12731.

Sutherland, R., Hodge, A., Bruck, S., Costley, D., Klieve, H., 2017. Parent-reported differences between school-aged girls and boys on the autism spectrum. Aut. Int. J. Res. Pract. 21 (6), 785–794. https://doi.org/10.1177/1362361316668653.

Szatmari, P., Liu, X.Q., Goldberg, J., Zwaigenbaum, L., Paterson, A.D., Woodbury-Smith, M., Georgiades, S., Duku, E., Thompson, A., 2012. Sex differences in repetitive stereotyped behaviors in autism: Implications for genetic liability. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 159B (1), 5–12. https://doi.org/10.1002/ajmg.b.31238.

Tape, T.G. (2004). Interpreting diagnostics tests. http://darwin.unmc.edu/dxtests. Accessed 15 May 2020.

Volkmar, F.R., Klin, A., Siegel, B., Szatmari, P., Lord, C., Campbell, M., Freeman, B.J., Cicchetti, D.V., Rutter, M., Kline, W., et al., 1994. Field trial for autistic disorder in DSM-IV. Am. J. Psychiatry 151 (9), 1361–1367. https://doi.org/10.1176/ajp.151.9.1361.

Wei, T., Chesnut, S.R., Barnard-Brak, L., Richman, D., 2014. Psychometric analysis of the social communication questionnaire using an item-response theory framework: implications for the use of the lifetime and current forms. J. Psychopathol. Behav. Assess. 37 (3), 469–480. https://doi.org/10.1007/s10862-014-9468-4.

van Wijngaarden-Cremers, P.J.M., van Eeten, E., Groen, W.B., van Deurzen, P.A.M., Oosterling, I.J., van der Gaag, R.J., 2014. Gender and age differences in the core triad of impairments in autism spectrum disorders: a systematic review and meta-analysis. J. Autism Dev. Disord. 44 (3), 627–635. https://doi.org/10.1007/s10803-013-1913-9.

Wilson, C.E., Murphy, C.M., McAlonan, G., Robertson, D.M., Spain, D., Hayward, H., Woodhouse, E., Deereley, P.Q., Gillan, N., Olsdon, J.C., Zinkstok, J., Stoengcheva, V., Faulkner, J., Yildiran, H., Bell, V., Hammond, N., Craig, M.C., Murphy, D.G.M., 2016. Does sex influence the diagnostic evaluation of autism spectrum disorder in adults? Autism Int. J. Res. Pract. 20 (7), 808–819. https://doi.org/10.1177/1362361315611381.

World Health Organization, 1992. International Statistical Classification of Diseases and Related Health Problems: ICD-10. World Health Organization, Geneva, 10. rev. ed.

World Health Organization, 1980. The International Classification of Diseases: 9th Revision, Clinical Modification: ICD-9-CM, 2nd ed. World Health Organization, Washington, D.C. ed.