Post-print de: Ezpeleta, L., Penelo, E., de la Osa, N., Navarro, J.B., & Trepat, E. (2020). How the Affective Reactivity Index (ARI) works for teachers as informants. *Journal of Affective Disorders, 261*, 40-48. doi: 10.1016/j.jad.2019.09.080.
Versió de l'editor consultable a: https://www.sciencedirect.com/science/article/abs/pii/S0165032719302666

How the Affective Reactivity Index (ARI) Works for Teachers as Informants

Lourdes Ezpeleta[1,2]

Eva Penelo[1,3]

Núria de la Osa[1,2]

J. Blas Navarro[1,3]

Esther Trepat[1,2]

[1]Unitat d'Epidemiologia i de Diagnòstic en Psicopatologia del Desenvolupament (2017 SGR 33, Generalitat de Catalunya)

[2]Departament de Psicologia Clínica i de la Salut

[3]Departament de Psicobiologia i Metodologia de les Ciències de la Salut

Universitat Autònoma de Barcelona

Funding

This work was supported by the Spanish Ministry of Science, Innovation and Universities (MICINN/FEDER) [grant PGC2018-095239-B-I00].

Mailing address:

Lourdes Ezpeleta

Departament de Psicologia Clínica i de la Salut. Edifici B

Universitat Autònoma de Barcelona

08193 Bellaterra (Barcelona). SPAIN

Phone: (34) 935 812883

E-mail: lourdes.ezpeleta@uab.cat

Abstract

Background: The Affective Reactivity Index (ARI) is a brief instrument originally designed as a self- and parent report. However, the view of teachers, who can observe social situations that may give rise to irritability, is relevant. The goal is to provide the measurement qualities of the ARI score as reported by teachers. Method: Children formed part of a longitudinal study on behavior problems in Barcelona (Spain) and they were assessed when they were 7 (*N*=471) and 11 years old (*N*=454) with questionnaires about psychopathology, anger and aggressive behavior, and a diagnostic interview answered by the parents, youths and teachers. Confirmatory factor analysis, measurement invariance, reliability and validity were studied for the ARI answered by teachers. Results: The 6-item, 1-factor model fitted well. Almost full metric invariance and partial scalar invariance was obtained across sex and over age. The ARI scores largely converged with other teacher-reported measures of anger and irritability, and with other measures of psychopathology, aggressive behavior, and callous-unemotional traits at a medium level. The associations with parent's measures were medium to low, and very low for child self-reported measures. The ARI scores significantly differentiated children with and without psychopathology and functional impairment, both cross-sectionally and longitudinally. Limitations: only one child self-report measure of irritability included. Limited internal consistency of some scale scores. Findings are mostly generalizable to Spanish children. Conclusions: ARI could be a suitable instrument for measuring irritability as reported by teachers. The teacher's view can be useful when planning treatment by helping to identify treatment targets.

How the Affective Reactivity Index (ARI) Works for Teachers as Informants

DSM-5 (American Psychiatric Association, 2013a) has shifted from a purely categorical conceptualization of mental disorders to incorporate the notion that psychopathology occurs along dimensional continua. To this effect, the DSM-5 task force and some working groups have developed dimensional measures that evaluate the presence and severity of symptoms that cut across diagnostic boundaries and may inform clinical decision making (Clarke and Kuhl, 2014). One of these cross-cutting symptoms is irritability. Irritability is a shared symptom in different disorders, such as anxiety disorders, depressive disorders, oppositional defiant disorder (ODD), bipolar disorder, post-traumatic stress disorder, and disruptive mood dysregulation disorder, among others (Toohey and DiGiuseppe, 2017). Irritability can be defined as an elevated proneness to anger relative to peers at the same development level (Brotman et al., 2017; Stringaris et al., 2018; Vidal-Ribas et al., 2016) and is characterized by easy annoyance, a low or decreased threshold for frustration, touchiness, and anger/temper outbursts. Irritability is a frequent reason for mental health referral and predicts negative outcomes from childhood to adulthood (Copeland et al., 2014; Ezpeleta et al., 2016). Meta-analyses report that irritability is associated with future depression, anxiety problems, and ODD (Vidal-Ribas et al., 2016). Irritability also predicts ADHD (Shaw et al., 2014), comorbidity (internalizing and externalizing), difficulties with peers (Ezpeleta et al., 2016), functional impairment (Wiggins et al., 2018), suicide (Benarous et al., 2018), and, in adulthood, adverse health, educational and social outcomes (Copeland et al., 2014). There is a consequent need to measure irritability (see Toohey and DiGiuseppe, 2017, for a review).

The Affective Reactivity Index (ARI; Stringaris et al., 2012) is the instrument proposed by the American Psychiatric Association (2013b) to assess the cross-cutting

symptom of irritability. The ARI contains 6 irritability symptom items plus one impairment item that must be scored if present during the previous six months. According to Stringaris et al. (2012), the items cover (a) the threshold for an angry reaction, (b) the frequency, and (c) the duration of the feelings/behaviors. The questionnaire was designed to be answered by youth (6-17 years old) and their parents.

Supplementary material Table 1 synthesizes the current available research on the psychometric properties of the ARI. The questionnaire has been studied in several countries (UK, US, Brazil, Australia and China) with samples of children/youth aged 5 to 19 years. Most of the studies have used the child and parent version, but Mulraney et al. (2014b) also studied an adult version. Regarding reliability, internal consistency is in the high range (alpha between .84 and .90 for youth, between .80 and .92 for parents and .80 for the adult version score). Test-retest reliability is acceptable for youth (ICC: .66) and good for parents (.82) (Pan and Yeh, 2018) and adults (.80) (Mulraney et al., 2014b). Parent-child agreement ranges between .42 (Pan and Yeh, 2018) and .73 (Stringaris et al., 2012). All the developed versions show that there is consistency in the answers to the items and the scores are more stable for parents. Regarding validity, confirmatory factor analyses has obtained a one-factor good fit for parents (Mulraney et al., 2014b; Stringaris et al., 2012) and the fit was acceptable for youth in one study (DeSousa et al., 2013) but not in two others (Mulraney et al., 2014b; Stringaris et al., 2012); the adult version has also shown poor fit. Parent and youth ARI scores differentiate control groups from groups with severe mood disorders, bipolar disorders, ODD (only parent) (Pan and Yeh, 2018; Stringaris et al., 2012), and DSM-IV diagnoses (Mulraney et al., 2014a). ARI scores converge to a large extent with other irritability measures (Pan and Yeh, 2018) and within a small to medium range with measures of psychopathology, impairment, and adaptation (Mulraney et al., 2014b; Pan and Yeh, 2018). Peer problems for the parent version and prosocial problems for the youth version of the Strengths and

Difficulties Questionnaire (SDQ) are not associated (Mulraney et al., 2014b; Stringaris et al., 2012). Cross-sectionally, the youth and parent ARI scores have been associated with SDQ emotional problems, conduct problems, and hyperactivity (Mulraney et al., 2014a; Mulraney et al., 2014b; Stringaris et al., 2012). No study has made longitudinal predictions. The validity results indicate that the ARI is related to outcomes that have been shown to be related to irritability and that ARI scores can differentiate groups with different psychopathologies. In summary, these results indicate not only a broad use of the questionnaire internationally, but also an added interest in extending the initial use to adults. In all the versions the ARI presents good psychometric properties.

However, none of the previous studies have tested how the ARI scores work when answered by teachers. Discrepancies between informants are the norm in child psychopathology assessment (De los Reyes and Kazdin, 2005), highlighting the need to obtain information from various contexts (De Los Reyes, 2011). Regarding ODD, non-equivalence across parents and teachers has been found for some items of its dimensions (one of which was irritability), as it has been observed that parents tend to rate ODD behaviors as more frequent than teachers (Ezpeleta and Penelo, 2015). Teachers are familiar with children's normative development and may be the best reporter of peer and social relations and behaviors at school (Konold and Pianta, 2007). Teachers can observe the child in social situations that may give rise to irritability and its expression. For some disorders, the diagnosis system may require the presence of the symptoms in more than one setting and teachers may inform whether the symptoms are present in the school context (Evans et al., 2016). In this line, the information provided by the teacher may help the clinician to identify pervasive and severe symptoms that occur in several contexts and require special attention, which could potentially facilitate treatment planning and boost treatment efficacy (De Los Reyes et al., 2015). The goal of the study was to test the psychometric properties of the ARI

answered by teachers. Since appropriate and proper comparison of a construct between groups and across times depends first on ensuring equivalence of meaning of the construct (e.g., Putnick and Bornstein, 2016), invariance of the measurement model across sex and over measurement occasions was analyzed. Based on the position of the teacher as observer in the school context, we expected high convergence between the ARI scores and other constructs related to irritability, such as aggressive behavior, psychopathology, functional impairment, and other measures of irritability. In line with the literature, we expected the same-informant values to be higher than the cross-informant ones. We also expected higher irritability scores to be associated with worse outcomes cross-sectionally and longitudinally, also differentiating between groups with and without problems. This is the first study to report on ARI predictive validity longitudinally.

Method

Participants

The sample comes from a longitudinal study of behavioral problems in a sample of children from Barcelona (Spain). A two-phase design was employed (see Ezpeleta et al. (2014) and supplementary material Figure 1). The children were evaluated yearly from age 3 to age 11. For the purpose of this study, we used the data recruited when the children were 7 and 11 years old (we used age 6 for just one questionnaire).

The ARI teacher-reported data were available for 471 children at age 7 (*M* = 7.7 years, *SD* = 0.36; 51.0% girls; 96.9% born in Spain; 92.1% Caucasian, 4.0% Latin American, and 3.9% other ethnicity; 32.1% high, 46.7% middle, and 21.2% low socioeconomic status [SES]; 64.1% from state schools and 35.9% from state-subsidized private schools) and 454 children at age 11 (*M* = 11.6 years, *SD* = 0.34; 51.5% girls; 97.4% born in Spain; 91.0% Caucasian, 4.8% Latin American, and 4.2% other ethnicity; 32.6% high, 48.5% middle and 18.9% low

SES; 66.0% from state and 34.0% from state-subsidized private schools); 396 were rated at both ages (74 only at age 7, and 58 only at age 11). At age 7, 135 teachers from 74 different schools (54 state and 20 state-subsidized) each rated between 1 and 11 children (*Md* = 3); and at age 11, 121 teachers from 72 different schools (48 state and 24 state-subsidized) each rated between 1 and 15 children (*Md* = 3). No statistically significant differences were observed by sex ($p \geq .362$) or SES ($p \geq .156$) between the children remaining at each follow-up and those not retained at ages 7 and 11. Regarding type of school, no differences were observed at age 7 ($p = .681$), but at age 11 there were more children from state than state-subsidized schools (78.6% vs. 67.9%, $p = .003$).

Measures

The *Affective Reactivity Index* (ARI; Stringaris et al., 2012) contains 6 items about feelings and behaviors related to irritability (0: *not true*; 1: *somewhat true*; 2: *certainly true*) plus one item assessing impairment due to irritability during the last 6 months. The children's teachers, who had known them for a mean of 10.2 months, answered the ARI questionnaires when the children were 7 and 11 years old.

Promis- Anger—Parent/Guardian of Child Age 6-17 (PROMIS-Patient-Reported Outcomes Measurement Information System, 2016) measures angry mood with 5 items (feels mad, yells, throws thing when angry, feels upset, stays mad; 0: *Never;* 5: *Always)* as reported by teachers at ages 7 and 11.

The *Strengths and Difficulties Questionnaire* (SDQ; Goodman, 2001) assesses children's mental health with 25 items on five scales: emotional symptoms, conduct problems, hyperactivity, peer relationship problems, and prosocial behavior. The items on the first four scales provide a total difficulties score. Two broader internalizing (emotional and peers) and externalizing (conduct and hyperactivity) scales (Goodman et al., 2010) were also

analyzed. The questionnaire has an impact supplement which is useful for considering the need for mental health services, where the informant judges whether a problem is present and the degree of distress, social impairment and burden it is causing others. The dimension *irritability* (Stringaris and Goodman, 2009) was made up of the items loses their temper, touchy and angry. The SDQ was completed by teachers when the children were ages 7 and 11.

The *Children's Aggression Scale* (CAS; Halperin and McKay, 2008) assesses aggressive behavior with 22 items (0: *never* to 4: *many days*) that make up the verbal aggression, aggression against objects/animals, physical aggression, and use of weapons scales. The total score plus two clusters (aggression towards peers and aggression towards adults) derived from teachers' responses at ages 7 and 11 were also analyzed.

The *Inventory of Callous-Unemotional Traits* (ICU; Frick, 2004) includes 24 items (0: *not at all true* to 3: *definitely true*) structured in three dimensions: Callousness, Uncaring and Unemotional. The total score is the sum of the raw scores as reported by the teachers when the children were ages 7 and 11.

The *Diagnostic Interview for Children and Adolescents for Parents of Preschool and Young Children* (DICA-PPYC; Ezpeleta et al., 2011) is a computerized, semi-structured diagnostic interview for assessing the most common psychological disorders in children as reported by parents, following the DSM-5 criteria (American Psychiatric Association, 2013a). Diagnoses of ADHD, ODD, any anxiety disorder (separation, generalized, specific phobia and social anxiety), and any comorbidity (includes the previously listed disorders plus conduct problems and major depression) at ages 7 and 11 were used for the study.

The *Children's Global Assessment Scale* (CGAS; Shaffer et al., 1983) is a global measure of functional impairment as rated by the interviewer/clinician based on information from diagnostic interviews with the parents when the children were ages 7 and 11.

The *Child Adolescent Functional Assessment Scale* (Hodges, 1995) assesses the degree of functional impairment secondary to the presence of psychological problems in several contexts as rated by the interviewer/clinician. Impairment at school, home and behavior toward others at ages 7 and 11 were used for the study.

The *Child Behavior Checklist* and the *Youth Self-Report* (CBCL/6-18, YSR; Achenbach and Rescorla, 2001) measure behavioral and emotional problems as reported by parents (ages 7 and 11) and youth (age 11), respectively, through 112 items with 3 response options (0: *not true*, 1: *somewhat/sometimes true*, 2: *very true/often true*). Empirical scales were used for the analyses. The sum of three items (*temper tantrum or hot temper, stubborn, sullen or irritable* and *sudden changes in mood or feelings)* measured irritability.

The *Emotion Regulation Checklist* (ERC; Shields & Cicchetti, 1997) assesses children's ability to regulate their emotions as reported by parents. It has 24 items (1: *never* to 4: *always true*). The negativity/lability and emotion regulation subscales were applied when the children were 7 years old.

The *Anger Questionnaire* (AQ; Unitat d'Epidemiologia, 2012) was created for this research project. It contains 40 items (0: *not at all*; 1: *a little;* 2: *a lot*) organized in two sections, one asking about the situations that make children angry and the other about what they do when they are angry. Exploratory factor analysis gave a one-dimensional factor structure for the first section (entitled Tolerance to frustration) and a bi-dimensional factor structure for the second section (External expression and Anger control), all three factors with acceptable internal consistency. The scores correlate with other measures of psychopathology, aggressive behavior and functional impairment. The development of the questionnaire and the psychometric properties can be found in Unitat d'Epidemiologia (2012). The children answered the questionnaire when they were 6 years old.

Supplementary material Table 2 shows the internal consistency reliability of the measures in the sample.

Procedure

The study was approved by the Ethics Commission of Animal and Human Experimentation of the authors' institution. The school principals and the families were provided with a detailed description of the research project. The families were recruited at the schools and gave written consent. They completed the questionnaires and were interviewed each academic year. The primary teachers were asked to complete the questionnaires by the end of the school year. The participating teachers had known the 7-year-olds for a mean of 11.2 months (*SD* = 4.8) and the 11-year-olds for a mean of 9.4 months (*SD* = 4.7).

Statistical Analysis

The statistical analyses were carried out with Stata 15 and MPlus8.1. Given the multistage sample, the data were analyzed using the case weighting procedure with sampling weights inversely proportional to the probability of participant selection.

Confirmatory Factor Analysis (CFA) was conducted with MPlus using Weighted Least Squares Means and Variance (WLSMV) adjusted for the categorical data method of estimation, which addresses floor and ceiling effects (Muthén and Muthén, 1998-2017). First, fit for a one-dimensional 6-item model with a multigroup approach across sex at ages 7 and 11 was examined. Goodness of fit was assessed with the common fit indices (Jackson et al., 2009): χ^2, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA); satisfactory fit was considered as > .95 for CFI and TLI and < .06 for RMSEA, and acceptable fit was > .90 for CFI and TLI and < .08 for RMSEA (e.g., Brown, 2006). Second, the measurement invariance of factor loadings (metric

invariance) and item thresholds (scalar invariance) across sex were analyzed, progressively comparing more constrained nested models across groups of responses following the common sequence (e.g., Marsh et al., 2013; Putnick and Bornstein, 2016; Vandenberg and Lance, 2000). And third, longitudinal measurement invariance over age involved a single-sample approach, considering the responses at ages 7 and 11 as repeated measures; in this case, error covariances between analogous items over time were also estimated to account for the non-independence of ratings at each age (Ferrando, 2000). For both invariance analyses, model identification for each step was set as described in Ezpeleta and Penelo (2015), by using the factor-variance strategy and taking into account the hierarchical data structure of teachers' ratings by means of a design-based multilevel CFA strategy (i.e., the Type = COMPLEX routine in MPlus). To compare nested models, the α level for scaled difference chi-square tests was set at .01. For a given step, when full invariance was not attained partial invariance was examined by sequentially releasing constraints (backward method). Any items that had unequal loadings in the metric invariance model and were allowed to vary were also allowed to vary in the scalar invariance model (e.g., Putnick and Bornstein, 2016).

Internal consistency of the ARI total score was calculated using Cronbach's alpha coefficient. Next, relations of ARI scores and external measures were examined. The nesting of children within teachers and schools accounted for a significant variance percentage of the ARI scores at both age 7 (ICC = 15.4%, χ^2 = 16.9, $p < .001$) and 11 (ICC = 17.9%, χ^2 = 33.5, $p < .001$), so analyses were carried out using a multilevel approach through linear mixed models using full maximum likelihood estimation (restricted maximum likelihood estimation was not possible in conjunction with weighting) and the Huber/White/sandwich estimator of the variance-covariance matrix (White, 1980). Convergent and discriminant validity between the ARI scores and quantitative external measures at the same age were evaluated using Pearson's correlation coefficients (r) with the school and the teacher at that age as random

factors. The predictive validity of the ARI scores at age 7 using SDQ at age 11 as outcome was also calculated using Pearson's correlation with the school and the teacher at age 7 as the random factors. Criterion-related validity, considering the presence of DSM-5 diagnoses and functional impairment at the same age (concurrent) and from age 7 to 11 (predictive), was also examined using linear mixed models taking the binary diagnosis as predictor and the ARI score as outcome. When the estimation process presented convergence problems the nesting structure was simplified by deleting the school random factor.

Effect sizes were measured with Cohen's *d* and interpreted according to the following rules of thumb (in absolute values): small effect for values 0.20-0.50, medium effect for values 0.50-0.80 and large effect for values > 0.80 (Cohen, 1992).

Results

Confirmatory factor analysis and invariance by sex and age

Fit for the 1-factor configural baseline model across sex was satisfactory at age 7 (Table 1, model A0) but insufficient at age 11 (Table 1, model B0; RMSEA = .121). Fit was acceptable (Table 1, models A1 and B1) after item uniquenesses between Item 2 ("often loses their temper") and Item 6 ("loses their temper easily") were correlated based on their similar wording. These models were therefore examined for invariance analyses, considering that the inclusion of such correlated uniquenesses (CU) did not meaningfully alter other parameter estimates (Marsh et al., 2013).

Full metric and scalar invariance across sex were found at age 7 (Table 1, models A#), since all factor loadings and item thresholds were equivalent across girls and boys, respectively. The latent mean in boys (fixed at 0 in girls) was slightly higher at age 7 (d = 0.599, p = .008). At age 11 (Table 1, models B#) partial metric and scalar invariance was attained: all factor loadings except one and all item thresholds except four were equivalent

across girls and boys. The item with non-equivalent factor loading was Item 3 ("Stays angry for a long time"), being higher in girls (.97) than in boys (.85). In addition to both thresholds for Item 3, the non-equivalent item thresholds were the first thresholds for Item 1 ("Easily annoyed by others") and Item 2 ("Often loses temper"). The higher first threshold for Items 1, 2 and 3 in girls indicates that the option "*not true*" was more frequently endorsed in girls than in boys, and the higher second threshold for Item 3 also in girls indicates that the option "*certainly true*" was more frequently endorsed in boys than in girls. The latent mean in males (fixed at 0 in females) was slightly lower at age 11 ($d = -0.372$, $p = .039$).

Given that full or partial metric plus scalar invariance (strong invariance) was accomplished across sex at each age, repeated-measure invariance analysis over age was performed on the whole sample (Table 1, models C#). Partial metric and scalar invariance were observed since only one of the six factor loadings and five of the 12 item thresholds were found to be non-equivalent at ages 7 and 11 (Table 2, right). The item with a non-equivalent factor loading was Item 1 ("easily annoyed by others"), being lower in girls (.71) than in boys (.94), despite showing very high values for both. The other items with non-equivalent item thresholds were Item 3 (both thresholds) and Item 5 ("Gets angry frequently"; first threshold). For the three items, the higher first threshold at age 7 indicates that the option "not true" was more frequently endorsed at this age than at age 11. The latent mean (fixed at 0 at age 7) was lower at age 11 ($d = -1.038$, $p \leq .001$). The factor correlation between the ARI responses at ages 7 and 11 (which can be interpreted as an indicator for test-retest reliability over an average of about 4 years) was .57.

Table 2 (left) shows the descriptives at the item level, in addition to the standardized factor loadings and item thresholds for the final measurement model analyzed using CFA (right). The mean item at age 7 ranged from 0.10 (4. Angry most of the time) to 0.36 (2. Often loses their temper); at age 11 the range was from 0.07 (4. Angry most of the time) to 0.37 (1.

Easily annoyed). Internal consistency was satisfactory, with Cronbach's alpha values of .85 at age 7 and .88 at age 11. All standardized factor loading values were above .80 and statistically significant ($p < .001$).

The means (and *SD*) for the ARI direct scores at age 7 were 1.28 (2.14) in girls and 1.78 (2.34) in boys, with the boys scoring slightly higher than the girls ($z = 3.16$, $p = .002$, $d = 0.22$). At age 11, the means (and *SD*) were 1.02 (1.96) for the girls and 1.34 (2.30) for the boys, with no statistically significant differences ($z = 1.45$, $p = .148$, $d = 0.15$). In addition, the mean scores at age 7 were slightly higher than at age 11 ($z = 2.20$, $p = .028$, $d = 0.15$) (Table 2). All these differences in the derived direct scores show small or almost null effect sizes and are aligned with the results found for latent means using measurement invariance analysis. The ARI direct score correlated highly and significantly with the Impairment item both at age 7 ($r = .77$, $p < .001$) and age 11 ($r = .78$, $p < .001$).

The rounded average total score was also calculated by dividing the raw total score by 6, according to the scoring and interpretation section of the APA report (American Psychiatric Association, 2013b). At age 7, 80.3% of girls and 70.5% of boys showed no irritability (0 points average total score), whereas 18.8% of girls and 28.7% of boys showed mild-moderate irritability (1 point average total score) and 0.9% of girls and boys showed moderate-severe irritability (2 points average total score), with statistically significant differences between sexes ($\chi^2 = 6.41$, $df = 2$, $p = .040$). At age 11, no differences were found between girls and boys for irritability levels: 84.1% none, 14.6% mild-moderate, and 1.3% moderate-severe ($\chi^2 = 3.40$, $df = 2$, $p = .182$).

Convergence with external measures

Same informant (teacher) measures

Table 3 shows the associations of the ARI scores with other measures reported concurrently by teachers at ages 7 and 11. The highest associations were for anger and irritability measures (Promis-anger, SDQ-irritability), SDQ conduct, externalizing and total scores, and CAS total, verbal aggression and aggression towards peers. Medium correlation values were obtained for the remaining SDQ and CAS scale scores (except for use of weapons, which was very poor) and the ICU scale scores (except unemotional). In addition, higher ARI scores at age 7 were significantly associated with higher SDQ scores at age 11 (Table 4).

Other informants

Table 3 also shows the convergence of the ARI teacher scores with measures answered by other informants. The ARI score correlated at a low-medium level with functional impairment as rated by the interviewer (*r* from –.33 to .45). The associations with the CBCL parents' measures were between low and medium and were highest for the attention, rule-breaking, aggressive behavior and attention problems scale scores (between .21 and .40), and very low for the withdrawn/depressed and anxious/depressed scale scores (–.03 to .11). Correlations for CBCL-irritability were low at age 7 (.17) and somewhat higher at age 11 (.32). Correlation with the ERC-Lability score was low (.26) and very low for ERC-emotion regulation (–.06). All the associations with the child self-reported measures were very low [the highest were YSR aggressive behavior (.24), rule-breaking (.19) and AQ-external expression of anger (.22) scores].

Discriminative ability

Table 5 shows the mean of the total ARI teacher score by diagnostic groups. Cross-sectionally, the ARI scores were significantly higher in the groups with functional impairment

or with a diagnosis other than anxiety, with large effect sizes for ADHD and ODD (at age 11), medium effect sizes for comorbidity and functional impairment, and low effect sizes for ODD (at age 7), any anxiety, and any disorder.

Longitudinally, the ARI scores at age 7 were significantly higher in the groups with diagnoses and functional impairment at age 11, with large effect sizes for ADHD and small effect sizes for ODD, any disorder, and functional impairment. The ARI scores were not associated with either comorbidity or any anxiety disorders (Table 5).

Supplementary material Figures 2 and 3 show the mean of each of the ARI items by DICA-PPYC diagnoses of ADHD, ODD, and any anxiety, and the mean of the children with no diagnoses at ages 7 and 11. With few exceptions, the ARI items discriminated ADHD and ODD from the control groups, but they did not differentiate between any anxiety and the control groups.

Discussion

This is the first study that reports on ARI measurement qualities when answered by the teacher. The results indicate that the ARI could be a suitable instrument for measuring irritability as reported by teachers and that these scores could be used to predict psychopathology and functional impairment in children from the general population. As expected, the ARI scores have a high convergence with other teacher-reported measures of anger and irritability, and a medium convergence with other measures of psychopathology, aggressive behavior, and callous-unemotional traits. The convergence with measures of psychopathology and functional impairment as reported by parents was medium to low and for child self-reported measures it was very low. The ARI scores significantly differentiated groups with and without psychopathology and functional impairment both cross-sectionally and longitudinally.

The six ARI items make up one factor for the teacher reports, similar to the results obtained with the parent (Mulraney et al., 2014b; Stringaris et al., 2012) and child versions (DeSousa et al., 2013) (findings on measurement invariance will be commented on below). The teacher's ARI scores converge significantly with measures related to irritability such as anger, other measures of irritability, and psychopathology questionnaires such as the SDQ and the CBCL, aggressive behavior, callous unemotional traits, functional impairment, and emotional regulation scales. Other versions of the questionnaire (parent, child and adult) have shown convergence with similar constructs such as irritability, aggressive behavior, social problems, and impairment (DeSousa et al., 2013; Mulraney et al., 2014a; Stringaris et al., 2012; Pan and Yeh, 2018). As expected, the associations were higher when the informant was the same (different measures reported by the teacher) and when the construct was closely related to irritability (anger and other measures of irritability) (Konold and Pianta, 2007; Rowe et al., 2019).

The teachers' ARI scores showed good discriminative ability both cross-sectionally and longitudinally as they were able to differentiate children with DSM-5 diagnoses, comorbidity (only cross-sectionally), and functional impairment from those without. In terms of dimensional measures, higher teacher ARI scores at age 7 predicted higher SDQ scales at age 11. These results are in line with the other versions of the instrument, the scores of which also differentiate groups with and without disorders and with higher scores on the SDQ scales (Mulraney et al., 2014a, 2014b; Pan and Yeh, 2018; Stringaris et al., 2012). In our study effect sizes were large for ADHD both cross-sectionally and longitudinally and moderate for ODD cross-sectionally. Previous work has reported the association between irritability and ADHD (Aebi et al., 2013; Eyre et al., 2019; Kolko and Pardini, 2010), suggesting that irritability may be an early marker of mood problems in these children, which may imply that children with ADHD need to be routinely assessed for irritability to identify those most at risk

for current impairment and future depression (Eyre et al., 2017). In the same line, the dimensional model of ODD (one of the dimensions being irritability) has been tested in several samples (Burke et al., 2010; Lavigne et al., 2015; Whelan et al., 2013), so the association between irritability and ODD was expected.

According to teachers, the ARI scores had an impact on the children's daily lives. The association between the total score and the impact it caused was in the high range. The two items most associated with Item 7 (impairment) at ages 7 and 11 were related to frequency (5. Gets angry frequently and 2. Often loses their temper), indicating that frequent irritability (more than the threshold or the duration) is the item most associated with impairment according to teachers. Regarding the impact score of the SDQ, ARI irritability is associated with difficulties with peers and learning at school. Last, higher scores in the ARI are associated with worse functioning as rated by the interviewer/clinician and the highest association (at a medium-low level) is with difficulties in functioning at school. The parent and child versions of the ARI have also been associated with functional impairment (DeSousa et al., 2013; Pan and Yeh, 2018; Stringaris et al., 2012).

The associations between the teacher's ARI scores and the SDQ scores were higher for externalizing scale scores and close to medium for peer problems, whereas for the parent and youth versions the associations found by Stringaris et al. (2012) were higher for emotional problems and non-significant or low for peer problems. The associations were also higher (in the medium to low range) for externalizing than for internalizing problems for the parents' CBCL scales. However, the teachers' ARI scores were not able to differentiate the groups with and without anxiety (convergence with anxiety scales was also in the low range). Cross-context differences in children's behavior may explain these differing levels of associations and highlight the need to also obtain information from the teacher (Konold and Pianta, 2007). To this effect, Rowe et al. (2019), who investigated the proportion of variance

attributable to teachers, children's characteristics and different occasions of measurement of adjustment in elementary school children, found that teachers' externalizing behavior ratings were based primarily on the child's behavior whereas internalizing problem ratings may be more situational. Different contexts elicit different problems and emotional problems may be better captured by the parents at home than by the teachers at school (for instance, separation anxiety, fears, complaints of headaches, expressing worries). The meta-analysis by De Los Reyes et al. (2015) confirmed a consistent finding over the last 50 years that informants have higher levels of agreement when they report on observable characteristics (externalizing vs. internalizing) and when they observe the same context. However, irritability correlates have shown to differ by age. Savage et al. (2015) reported that irritability predicts anxious/depressed symptoms at older ages, which could be another possible explanation for the non-emergence of the associations between irritability and emotional symptoms.

Anyhow, in general, the agreement between informants reporting on child mental health problems is only moderate (van der Ende, Verhults & Tiemeier, 2012), and this was also the case of the relation between teacher-irritability (ARI) and parent externalizing problems (between .35 to .40). In the seminal work of Achenbach, McConaughy, and Howell (1987) on cross-informant agreement, the average correlations between informant observing the child in different settings (e.g., parent and teacher) were only .28. We obtained higher values, of moderate level, which indicate that observed teacher's irritability concerns observed at school share about 16% of variability with externalizing problems concerns observed by parents in non-school context.

The associations between the ARI and the youth self-reports, while expected (de los Reyes et al. 2005), must also be highlighted. There was only a small correlation between YSR-aggressive behavior and AQ-external expression of anger. On the one hand, these associations are consistent with the most reliable self-reported scales, which could provide a

higher association. On the other hand, as observed in other disorders, the child is the one who best knows their feelings (AQ-tolerance to frustration and control obtained acceptable reliability but a low association with teacher information). Therefore, if children as young as 6 years old report reliably in a questionnaire about anger, again the influence of the method (informant) is more important than the trait (Konold and Pianta, 2007). Additionally, there was one year between both measures. It seems, then, that the ARI teacher report captures irritability that is mostly associated with externalizing observable behaviors, whether these behaviors are observed by the teacher or by the parents. Alternatively, this could indicate that the validity of the internalizing measures is poorer (Ferdinand, 2008).

This is the first study to analyze the invariance of scores. Since all the factor loadings except one and more than half the item thresholds were found to be equivalent across sex and age, full or partial metric plus partial scalar invariance were obtained, respectively. Overall, since partial strong invariance (metric plus scalar) was obtained, it can be assumed that when reported by the teacher through ARI scores the construct (irritability) has the same meaning across sex and essentially does not change over time. This means that teacher ARI scores can be related to other constructs for different groups and that patterns of relations with other variables in the same group can be studied and compared across the sexes and over age. However, for the latter, three items showed one or both item thresholds as being non-equivalent over time. More specifically, the option "*not true*" was less endorsed at age 11 than at age 7 for 1. Easily annoyed by others, 3. Stays angry for a long time, and 5. Gets angry frequently. Given these results and the fact that measurement invariance constitutes a prerequisite for subsequent mean comparisons, caution may be required when comparing teacher-reported ARI scores over a long period of time. Internal consistency was good and the medium factor correlation obtained with an interval of 4 years (ages 7 to 11) may be indicative of temporal stability.

The unique contribution of this work is the evidence it provides of the measurement qualities of the ARI scores reported by teachers. We report positively on reliability and validity in a wide sample of the general population, both cross-sectionally and longitudinally, using different informants (teachers, parents and children) reporting about several constructs related to irritability. However, some limitations should be considered when interpreting the results. First, irritability may not always be associated with overt or aggressive behavior (Toohey and DiGiuseppe, 2017) and teachers (and parents) might not capture this subjective feeling. It is therefore advisable to include some self-reported measure of irritability. We could only include a child anger questionnaire answered when they were 6 years old, which could limit the association with later teacher measures. Second, the high association between the PROMIS (defined to measure anger) and the ARI (defined to measure irritability) could be explained by both questionnaires being answered by the same informant (teacher), but it may also suggests that they are both measuring the same construct of anger. Irritability and anger used to be used interchangeably in the literature (Stringaris et al., 2018). Anger is "the emotion that characterizes irritability" (Stringaris et al., 2018, p. 722) and irritability is "an increased proneness to anger compared with peers at the same development level" (Stringaris et al., 2018, p. 722). ARI could be measuring anger. Third, the availability of both parent and child ARI scores would have strengthened the study considerably, providing the opportunity to study incremental validity and agreement. Fourth, the predictive analyses were not adjusted by baseline levels, first because it was a different teacher or informant that assessed the child on the two occasions, and second because as is usual in psychometric validation studies (as opposed to clinical studies) we wanted to determine if ARI scores at age 7 are useful to predict later problems and diagnoses at age 11, bearing in mind that these available external measures reflect the association at earlier ages. Fifth, the internal consistency of some of the scale scores of the instruments used did not reach optimal levels and this could limit the

findings based on relations to external variables. Last, we report on a sample of children from Barcelona and this could limit the generalizability of the results.

To conclude, the use of the ARI answered by the teacher is supported by good psychometric properties. We know that the reports of multiple informants share little variance and that multi-informant assessments may reflect the specific contexts where subjects display their mental health difficulties (De Los Reyes et al., 2015). Parents may not be fully aware of the difficulties their children are having at school in expressing or controlling irritability and teachers may contribute with their observations in this context. Evidence-based assessment proposes a personalized assessment that considers the within- and between-context variations of the children's difficulties to optimally fit their unique needs (De Los Reyes et al., 2015). Hence, we need psychometrically sound instruments that capture information for each context where the subject may need care. If irritability increases risk of externalizing disorders via poor anger and sadness coping, internalizing problems via poor anger coping and anxiety symptoms via intolerance to uncertainty (Evans et al., 2019), teacher information about how irritability is presenting in the school context may be very valuable. Among other advantages, teacher information could: a) assist with early assessment to prevent externalizing and internalizing disorders; b) help to identify what the targets for the treatment of irritability are; c) help establish the severity of the irritability if present cross-context; and d) help CBT treatment by giving clues as to how to improve emotion coping, intolerance to uncertainty, and rumination at school.

Acknowledgements

We would like to thank the participating schools and families.

References

Achenbach, T. M., McConaughy, S. H., Howell, C. T., 1987. Child/adolescent behavioral and emotional problems: Implications of cross informant correlations for situational specificity. Psychol. Bull. 101, 213–232.

Achenbach, T.M., Rescorla, L.A., 2001. Manual for the ASEBA school-age forms & Profiles. University of Vermont, Research Center for Children, Youth & Families, Burlington, VT.

Aebi, M., Plattner, B., Metzke, C.W., Bessler, C., Steinhausen, H.-C., 2013. Parent- and self-reported dimensions of oppositionality in youth: construct validity, concurrent validity, and the prediction of criminal outcomes in adulthood. J Child Psychol and Psychiatry 54, 941-949.

American Psychiatric Association, 2013a. Diagnostic and statistical manual of mental disorders, 5th ed. American Psychiatric Association, Arlington, VA.

American Psychiatric Association, 2013b. DSM-5: Online Assessment measures.

Benarous, X., Consoli, A., Cohen, D., Renaud, J., Lahaye, H., Guile, J.-M., 2018. Suicidal behaviors and irritability in children and adolescents: a systematic review of the nature and mechanisms of the association. Eur. Child Adolesc. Psychiatry.

Brotman, M.A., Kircanski, K., Stringaris, A., Pine, D.S., Leibenluft, E., 2017. Irritability in Youths: A Translational Model. Am. J. Psychiatry 174, 520-532.

Brown, T.A., 2006. Confirmatory factor analysis for applied research. Guilford, New York.

Burke, J.D., Hipwell, A.E., Loeber, R., 2010. Dimensions of oppositional defiant disorder as predictors of depression and conduct disorder in preadolescent girls. J Am Acad Child Adolesc Psychiatry 49, 484-492.

Clarke, D.E., Kuhl, E.A., 2014. DSM-5 cross-cutting symptom measures: A step towards the future of psychiatric care? World Psychiatry 13, 314-316.

Cohen, J., 1992. A power primer. Psychol. Bull. 112 155-159.

Copeland, W.E., Shanahan, L., Egger, H., Angold, A., Costello, E.J., 2014. Adult Diagnostic and Functional Outcomes of DSM-5 Disruptive Mood Dysregulation Disorder. Am. J. Psychiatry 171, 668-674.

De Los Reyes, A. 2011. More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. J Clin Child Adolesc Psychol 40, 1-9.

De Los Reyes, A., Augenstein, T.M., Wang, M., Thomas, S.A., Drabick, D.A.G., Burgers, D.E., Rabinowitz, J., 2015. The Validity of the Multi-Informant Approach to Assessing Child and Adolescent Mental Health. Psychol. Bull. 141, 858-900.

De los Reyes, A., Kazdin, A.E., 2005. Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. Psychol. Bull. 131, 483-509.

DeSousa, D.A., Stringaris, A., Leibenluft, E., Koller, S.H., Manfro, G.G., Salum, G.A., 2013. Cross-cultural adaptation and preliminary psychometric properties of the Affective Reactivity Index in Brazilian Youth: implications for DSM-5 measured irritability. Trends in psychiatry and psychotherapy 35, 171-180.

Evans, S.C., Blossom, J.B., Fite, P.J., 2019. Exploring longitudinal mechanisms of irritability in children: Implications for cognitive-behavioral intervention. Behavior Therapy.

Evans, S.C., Pederson, C.A., Fite, P.J., Blossom, J.B., Cooley, J.L., 2016. Teacher-Reported Irritable and Defiant Dimensions of Oppositional Defiant Disorder: Social, Behavioral, and Academic Correlates. School Mental Health 8, 292-304.

Eyre, O., Langley, K., Stringaris, A., Leibenluft, E., Collishaw, S., Thapar, A., 2017. Irritability in ADHD: Associations with depression liability. J. Affect. Disord. 215, 281-287.

Eyre, O., Riglin, L., Leibenluft, E., Stringaris, A., Collishaw, S., Thapar, A., 2019. Irritability in ADHD: association with later depression symptoms. Eur. Child Adolesc. Psychiatry.

Ezpeleta, L., de la Osa, N., Doménech, J.M., 2014. Prevalence of DSM-IV disorders, comorbidity and impairment in 3-year-old Spanish preschoolers. Soc. Psychiatry Psychiatr. Epidemiol. 49, 145-155.

Ezpeleta, L., Granero, R., de la Osa, N., Trepat, E., Domènech, J.M., 2016. Trajectories of oppositional defiant disorder irritability symptoms in preschool children. J. Abnorm. Child Psychol. 44, 115-128.

Ezpeleta, L., Osa, N.d.l., Granero, R., Doménech, J.M., Reich, W., 2011. The Diagnostic Interview for Children and Adolescents for Parents of Preschool and Young Children: Psychometric Properties in the general Population. Psychiatry Res. 190, 137-144.

Ezpeleta, L., Penelo, E., 2015. Measurement invariance of oppositional defiant disorder dimensions in 3-year-old preschoolers European Journal of Psychological Assessment 31, 45-53.

Ferdinand, R.F., 2008. Validity of the CBCL/YSR DSM-IV scales Anxiety Problems and Affective Problems. J. Anxiety Disord. 22, 126-134.

Ferrando, P.J., 2000. Testing the equivalence among different item response formats in personality measurement: A structural equation modeling approach. Structural Equation Modeling 7, 271-286.

Frick, P.J., 2004. The Inventory of Callous-Unemotional traits. Unpublished rating scale.

Goodman, A., Lamping, D.L., Ploubidis, G.B., 2010. When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the strengths and difficulties questionnaire (SDQ): Data from british parents, teachers and children. J. Abnorm. Child Psychol. 38, 1179-1191.

Goodman, R., 2001. Psychometric properties of the Strenghts and Difficulties Questionnaire. J. Am. Acad. Child Adolesc. Psychiatry 40, 1337-1345.

Halperin, J.M., McKay, K.E., 2008. Children's Aggression Scale. Psychological Assessment Resources, Lutz, FL.

Hodges, K., 1995. Child and Adolescent Functional Assessment Scale. Eastern Michigan University, Department of Psychology, Ypsilanti, MI.

Jackson, D.L., Gillaspy, J.A., Purc-Stephenson, R., 2009. Reporting practices in confirmatory factor analysis: an overview and some recommendations. Psychological Methods 14, 6-23.

Kolko, D.J., Pardini, D.A., 2010. ODD dimensions, ADHD, and callous-unemotional traits as predictors of treatment response in children with disruptive behavior disorders. J. Abnorm. Psychol. 119, 713-725.

Konold, T.R., Pianta, R.C., 2007. The influence of informants on ratings of children's behavioral functioning. J. Psychoeduc. Assess. 25, 222-236.

Lavigne, J.V., Bryant, F.B., Hopkins, J., Gouze, K.R., 2015. Dimensions of oppositional defiant disorder in young children: Model comparisons, gender and longitudinal invariance. J. Abnorm. Child Psychol. 43, 423-439.

Marsh, H.W., Nagengast, B., Morin, A.J.S., 2013. Measurement invariance of Big-Five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and La Dolce Vita effects. Dev. Psychol. 49 1194-1218.

Mulraney, M., Melvin, G., Tonge, B., 2014a. Brief report: Can irritability act as a marker of psychopathology? J. Adolesc. 37, 419-423.

Mulraney, M.A., Melvin, G.A., Tonge, B.J., 2014b. Psychometric Properties of the Affective Reactivity Index in Australian Adults and Adolescents. Psychological Assessment 26, 148-155.

Muthén, L.K., Muthén, B.O., 1998-2017. Mplus User's Guide, 7th ed. Muthén & Muthén., Los Angeles, CA.

Pan, P.-Y., Yeh, C.-B., 2018. Irritability and Maladaptation Among Children: The Utility of Chinese Versions of the Affective Reactivity Index and Aberrant Behavior Checklist-Irritability Subscale. J. Child Adolesc. Psychopharmacol.

PROMIS-Patient-Reported Outcomes Measurement Information System, 2016. PROMIS Anger. U. S. Department of Health and Human Services.

Putnick, D.L., Bornstein, M.H., 2016. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Dev. Rev. 41, 71-90.

Rowe, E.W., Curby, T.W., Kim, H., 2019. Variance in Teacher Ratings of Children's Adjustment. J. Psychoeduc. Assess. 37, 26-39.

Savage, J., Verhulst, B., Copeland, W., Althoff, R.R., Lichtenstein, P., Roberson-Nay, R., 2015. A genetically informed study of the longitudinal relation between irritability and anxious/depressed symptoms. J. Am. Acad. Child Adolesc. Psychiatry 54, 377–384.

Shaffer, D., Gould, M.S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., Aluwahlia, S., 1983. A Children's Global Assessment Scale (CGAS). Arch. Gen. Psychiatry 40, 1228-1231.

Shaw, P., Stringaris, A., Nigg, J., Leibenluft, E., 2014. Emotion Dysregulation in Attention Deficit Hyperactivity Disorder. Am. J. Psychiatry 171, 276-293.

Shields, A., & Cicchetti, D. 1997. Emotion regulation among school-age children: The development and validation of a new criterion Q-sort scale. Dev Psycho 33, 906-916.

Stringaris, A., Goodman, R., 2009. Three dimensions of oppositionality in youth. Journal of Child Psychology and Psychiatry 50, 216-223.

Stringaris, A., Goodman, R., Ferdinando, S., Razdan, V., Muhrer, E., Leibenluft, E., Brotman, M.A., 2012. The Affective Reactivity Index: a concise irritability scale for clinical and research settings. Journal of Child Psychology and Psychiatry 53, 1109-1117.

Stringaris, A., Vidal-Ribas, P., Brotman, M.A., Leibenluft, E., 2018. Practitioner Review: Definition, recognition, and treatment challenges of irritability in young people. Journal of child psychology and psychiatry, and allied disciplines 59, 721-739.

Toohey, M.J., DiGiuseppe, R., 2017. Defining and measuring irritability: Construct clarification and differentiation. Clin. Psychol. Rev. 53, 93-108.

Unitat d'Epidemiologia I de Diagnòstic en Psicopatologia del Desenvolupament. (2012). *AQ - Anger Questionnaire*. Retrieved from http://www.ued.uab.cat/docs/AQ_Development_and_validation.pdf

van der Ende, J., Verhulst, F. C., Tiemeier, H., 2012. Agreement of informants on emotional and behavioral problems from childhood to adulthood. Psychological Assessment, 24, 293-300.

Vandenberg, R., Lance, C., 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods 3, 4-69.

Vidal-Ribas, P., Brotman, M.A., Valdivieso, I., Leibenluft, E., Stringaris, A., 2016. The status of irritability in psychiatry: A conceptual and quantitative review. J. Am. Acad. Child Adolesc. Psychiatry 55, 556-570.

Whelan, Y.M., Stringaris, A., Maughan, B., Barker, E.D., 2013. Developmental continuity of oppositional defiant disorder subdimensions at ages 8, 10, and 13 years and their distinct psychiatric outcomes at age 16 years. Journal of the American Academy of Child Adolescent Psychiatry 52, 961-969.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.

Wiggins, J.L., Briggs-Gowan, M.J., Estabrook, R., Brotman, M.A., Pine, D.S., Leibenluft, E., Wakschlag, L.S., 2018. Identifying Clinically Significant Irritability in Early Childhood. J. Am. Acad. Child Adolesc. Psychiatry 57, 191-199.

Table 1. Fit indices for measurement invariance analyses

	Goodness-of-fit indices				Comparison		
Model	χ^2 (*df*)	CFI	TLI	RMSEA (CI 90%)	Models	$\Delta\chi^2$ (Δdf)	*p*
Across sex at age 7 (*N* = 469)							
A0: configural (equal form)	35.6 (18)	.993	.988	.065 (.032; .096)			
A1: A0 modified (CU between it2-it6)	24.4 (16)	.997	.994	.047 (.000; .083)			
A2: A1 plus equal factor loadings	32.1 (21)	.995	.994	.047 (.000; .079)	A2 vs. A1	8.6 (5)	.125
A3: A2 plus equal item thresholds	50.8 (32)	.992	.993	.050 (.021; .075)	A3 vs. A2	20.9 (11)	.034
Across sex at age 11 (*N* = 451)							
B0: configural (equal form)	77.3 (18)	.994	.991	.121 (.094; .149)			
B1: B0 modified (CU between it2-it6)	43.1 (16)	.997	.995	.087 (.056; .118)			
B2: B1 plus equal factor loadings	66.4 (21)	.996	.994	.098 (.072; .125)	B2 vs. B1	23.1 (5)	<.001
B2+: B2 except 1 factor loading (it3)	48.1 (20)	.997	.996	.079 (.050; .108)	B2+ vs. B1	9.0 (4)	.061
B3: B2+ plus equal item thresholds (except 2 it3)	70.2 (29)	.996	.996	.079 (.056; .103)	B3 vs. B2+	26.7 (9)	.002
B3+: B3 except 4 item thresholds	60.8 (27)	.997	.996	.075 (.050; .100)	B3+ vs. B2+	15.9 (7)	.026
Across age, repeated measures (*N* = 527)							
C1: C0 modified (CU between it2-it6)	91.7 (45)	.993	.989	.044 (.031; .057)			
C2: C1 plus equal factor loadings	110.4 (50)	.991	.988	.048 (.036; .060)	C2 vs. C1	20.0 (5)	.001
C2+: C2 except 1 factor loading (it1)	97.8 (49)	.992	.990	.043 (.031; .056)	C2+ vs. C1	9.3 (4)	.053
C3: C2+ plus equal item thresholds (except 2 it1)	142.6 (58)	.997	.985	.053 (.042; .064)	C3 vs. C2+	58.1 (9)	<.001
C3+: C3 except 5 item thresholds	111.9 (55)	.991	.989	.044 (.032; .056)	C3+ vs. C2+	16.9 (6)	.010

CU: correlated uniquenesses

Table 2. Item descriptives, association with impairment item, and standardized factor loadings and item thresholds for CFA

	Age 7								Age 11							
	Response option (%) [a]			*M* (*SD*)	Association	CFA (model C3+) [c]			Response option (%) [a]			*M* (*SD*)	Association	CFA (model C3+) [c]		
	0	1	2		with impairment [b]	λ	τ_1	τ_2	0	1	2		with impairment [b]	λ	τ_1	τ_2
1. Easily annoyed by others	78.4	18.4	3.2	0.25 (0.50)	.42	**.71**	**1.11**	**2.62**	68.2	26.7	5.1	0.37 (0.58)	.61	**.94**	**−0.71**	**2.69**
2. Often lose temper	68.6	26.7	4.7	0.36 (0.57)	.67	.83	0.91	3.09	87.3	11.5	1.2	0.14 (0.38)	.65	.90	0.91	3.09
3. Stay angry for a long time	74.8	23.2	2.0	0.27 (0.49)	.54	.83	**1.18**	**3.63**	76.2	19.5	4.3	0.28 (0.54)	.57	.89	**0.07**	**2.31**
4. Angry most of the time	90.4	9.1	0.5	0.10 (0.32)	.47	.84	2.26	4.17	93.7	5.1	1.2	0.07 (0.30)	.53	.91	2.26	4.17
5. Get angry frequently	77.7	18.6	3.7	0.26 (0.52)	.70	.94	**2.22**	4.95	83.6	13.1	3.3	0.20 (0.47)	.65	.97	**0.94**	4.95
6. Lose temper easily	76.8	17.2	5.9	0.29 (0.57)	.72	.81	1.26	2.85	89.5	9.4	1.1	0.12 (0.35)	.68	.89	1.26	2.85
Impairment item	81.2	14.5	4.3	0.23 (0.51)					68.2	26.7	5.1	0.17 (0.45)				
Total score				1.53 (2.25)	.79							1.18 (2.14)	.76			

[a] 0: *Not true*; 1: *Somewhat true*; 2: *Certainly true*

[b] All association with impairment were statistically significant ($p < .001$)

[c] CFA: Confirmatory factor analysis (see Table 1); λ: standardized factor loading; τ_1 and τ_2: first and second threshold; in bold: non-equivalent parameters over age

Table 3: Cross-sectional Pearson's correlations between Teacher's ARI scores and other measures

Informant	Measure		Age 7	Age 11
Teacher	Promis-Anger		.86**	.83**
	Strengths Difficulties Questionnaire	Conduct	.73**	.75**
		Emotional	.30**	.32**
		Hyperactivity	.43**	.50**
		Peer	.40**	.42**
		Prosocial[a]	.46**	.52**
		Total	.62**	.69**
		Impact	.50**	.65**
		Externalizing (conduct+hyper.)	.59**	.67**
		Internalizing (emotional+peer)	.41**	.43**
		Irritability dimension	.79**	.84**
	Children Aggression Scale	Verbal aggression	.76**	.72**
		Aggression against objects/animals	.61**	.57**
		Physical aggression	.60**	.48**
		Use of weapons	.18*	na
		Aggression towards peers	.72**	.70**
		Aggression towards adults	.46**	.56**
		Total score	.72**	.64**
	Inventory Callous-Unemotional	Callousness	.50**	.61**
		Uncaring	.49**	.60**
		Unemotional	−.04	.06
		Total	.43**	.56**
Interviewer/	Children Global Assessment Scale (CGAS)		−.33**	−.36**
Clinician	Child Adolescent Functional Assessment Scale	School	.39**	.45**
		Home	.26**	.27**
		Behavior towards others	.28**	.37**
Parents	Child Behavior Checklist	Withdrawn/Depressed	−.03	.09
		Anxious/depressed	.11	.08
		Attention problems	.31**	.21**

		Aggressive behavior	.35**	.40**
		Rule-Breaking	.40**	.37**
		CBCL-Irritability	.17*	.32**
	Emotion Regulation Checklist	Lability-Negativity	.26**	
		Emotion regulation	−.06	
Self-report	Youth Self-Report	Withdrawn/Depressed		.03
		Anxious/depressed		.00
		Attention problems		.17**
		Aggressive behavior		.24**
		Rule-Breaking		.19**
		YSR Irritability		.11*
	Anger questionnaire (age 6)	Tolerance to frustration	*.12*	
		External expression	*.22***	
		Anger control	*−.06*	

* $p < .05$; ** $p < .01$

In italics: at age 6

[a] Scores were reversed: higher scores indicate higher prosocial problems.

na: non-applicable because there was no variability in the scores of use of weapons at age 11

Table 4. Pearson's correlations between Teacher's ARI scores at age 7 and Teacher SDQ scores at age 11

	r
Conduct	.34**
Emotional	.22**
Hyperactivity	.34**
Peer	.26**
Prosocial[a]	.21**
Total	.41**
Impact	.32**
Externalizing (conduct+hyperativity)	.39**
Internalizing (emotional+peer)	.29**
Irritability dimension	.42**

* $p < .05$; ** $p < .01$

[a] Scores were reversed: higher scores indicate higher prosocial problems.

Table 5. ARI scores and cross-sectional and longitudinal comparison between diagnostic groups (DICA-PPYC) and functional impairment (CGAS)

						Comparison between groups		
		No/Absent		Yes/Present				
ARI	Groups	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*z*	*p*	*d*
Age 7	Age 7							
	ADHD	424	1.34 (2.04)	45	3.52 (3.22)	3.87	**<.001**	**1.00**
	ODD	437	1.50 (2.25)	31	2.30 (2.33)	2.28	**.023**	0.36
	Any anxiety	424	1.49 (2.17)	45	2.09 (3.00)	1.33	.184	0.26
	Any disorder	312	1.26 (2.03)	157	2.12 (2.58)	3.69	**<.001**	0.38
	Comorbidity	421	1.42 (2.13)	48	2.74 (3.00)	2.78	**.005**	**0.59**
	CGAS < 70	372	1.24 (1.98)	97	2.73 (2.85)	4.73	**<.001**	**0.68**
Age 11	Age 11							
	ADHD	403	1.03 (1.94)	31	3.23 (3.43)	2.84	**.004**	**1.06**
	ODD	398	1.01 (1.88)	36	3.07 (3.66)	2.79	**.005**	**0.99**
	Any anxiety	390	1.17 (2.15)	43	1.27 (2.25)	0.52	.603	0.05
	Any disorder	288	0.88 (1.71)	146	1.79 (2.74)	2.90	**.004**	0.43
	Comorbidity	396	1.10 (2.07)	38	2.08 (2.74)	2.19	**.028**	0.46
	CGAS < 70	283	0.68 (1.34)	150	2.15 (2.94)	4.80	**<.001**	**0.72**
Age 7	Age 11							
	ADHD	374	1.30 (2.01)	33	3.01 (3.44)	2.71	**.013**	**0.79**
	ODD	368	1.36 (2.18)	39	2.15 (2.36)	2.23	**.015**	0.34
	Any anxiety	364	1.40 (2.16)	43	1.79 (2.57)	0.88	.213	0.18
	Any disorder	260	1.16 (1.88)	148	1.93 (2.63)	2.84	**.008**	0.35
	Comorbidity	363	1.35 (2.15)	45	2.18 (2.52)	2.03	.058	0.38
	CGAS < 70	259	1.10 (1.89)	148	2.00 (2.55)	3.40	**.002**	0.42

In bold: $p < .05$ and/or $d > 0.50$