Psychological Assessment
2016, Vol. 28, No. 5, 598–612

1040-3590/16/$12.00 http://dx.doi.org/10.1037/pas0000185

Identifying an Efficient Set of Items Sensitive to Clinical-Range Externalizing Problems in Children

Isaac T. Petersen and John E. Bates
Indiana University

Kenneth A. Dodge and Jennifer E. Lansford
Duke University

Gregory S. Pettit
Auburn University

The present study applied item response theory to identify an efficient set of items of the Achenbach Externalizing scale from the Child Behavior Checklist (CBCL; 33 items) and Teacher's Report Form (TRF; 35 items) that were sensitive to clinical-range scores. Mothers and teachers rated children's externalizing problems annually from ages 5 to 13 years in 2 independent samples (*N*s = 585 and 1,199). Item properties for each rater across ages 5–8 and 9–13 were examined with item response theory. We identified 10 mother- and teacher-reported items from both samples based on the items' measurement precision for subclinical and clinical levels of externalizing problems: externalizing problems that involve meanness to others, destroying others' things, fighting, lying and cheating, attacking people, screaming, swearing/obscene language, temper tantrums, threatening people, and being loud. Scores on the scales using these items had strong reliability and psychometric properties, capturing nearly as much information as the full Externalizing scale for classifying clinical levels of externalizing problems. Scores on the scale with the 10 CBCL items had moderate accuracy, equivalent to the full Externalizing scale, in classifying diagnoses of conduct disorder based on a research diagnostic interview. Of course, comprehensive clinical assessment would consider additional items, dimensions of behavior, and sources of information, too, but it appears that the behaviors tapped by this select set of items may be core to externalizing psychopathology in children.

Keywords: externalizing behavior problems, antisocial behavior, aggression and conduct problems, item response theory, oppositional and defiant problems

Supplemental materials: http://dx.doi.org/10.1037/pas0000185.supp

Externalizing behavior problems in children are common and burdensome. The median prevalence estimate of disruptive behavior disorders (conduct disorder, oppositional defiant disorder, or attention-deficit/hyperactivity disorder) among children and adolescents in the general population from studies between 1993 and 2005 was about 6% (Costello, Egger, & Angold, 2005). The present study applies a method for increasing the efficiency of assessment for clinical levels of externalizing problems by selecting a subset of useful items that is faster to administer than the full questionnaire while still retaining measurement precision for the targeted goal.

Efficient assessment is important for both research and clinical purposes. Efficient assessment tools permit more frequent assessments, which can be useful when monitoring treatment progress.

This article was published Online First August 31, 2015.

Isaac T. Petersen and John E. Bates, Department of Psychological and Brain Sciences, Indiana University; Kenneth A. Dodge and Jennifer E. Lansford, Center for Child and Family Policy, Duke University; Gregory S. Pettit, Department of Human Development and Family Studies, Auburn University.

The authors gratefully acknowledge Thomas Achenbach and his associates for their research and development of the Achenbach System of Empirically Based Assessment. The authors declare that they have no conflict of interest. The Child Development Project has been funded by Grants MH42498, MH56961, MH57024, and MH57095 from the National Institute of Mental Health; HD30572 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development; and DA016903 from the National Institute on Drug Abuse. The Fast Track project was supported by National Institute of Mental Health (NIMH) Grants R18 MH48043, R18 MH50951, R18 MH50952, R18 MH50953, and by National Institute on Drug Abuse (NIDA) Grant 1 RC1 DA028248-01. The Center for Substance Abuse Prevention and NIDA also provided support for Fast Track through a memorandum of agreement with the NIMH. Isaac T. Petersen was supported by Grant TL1 TR000162 (A. Shekhar, PI) from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award, and by a Ruth L. Kirschstein National Research Service Award grant from the National Institute of Mental Health (1 F31 MH100814-01A1). Kenneth A. Dodge was supported by the National Institute of Drug Abuse Grant 2K05 DA015226. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Institutes of Health or Indiana University.

Correspondence concerning this article should be addressed to Isaac T. Petersen, Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th Street, Bloomington, IN 47405. E-mail: itpeters@indiana.edu

Efficient assessments also reduce participant burden. Questionnaires are often used as assessment tools, in part, because of their brevity. In order for questionnaires to be clinically practical for assessing externalizing disorders, questionnaire items should be: (a) few, (b) developmentally appropriate, (c) rater appropriate, and (d) maximally informative for deciding whether a child has clinical or subclinical levels of externalizing problems (Achenbach, 1991a; Chorpita et al., 2010; Studts & van Zyl, 2013).

Item response theory (IRT) provides an empirical way to evaluate the clinical utility of questionnaire items. In the present study, we fit IRT models that estimate two properties of each item: (a) discrimination, and (b) difficulty (severity). An item's discrimination parameter describes how well the item distinguishes between low and high levels of the trait being measured (i.e., how well the item relates to the trait). For example, if an item asking how often a child hits others is more relevant to externalizing problems than an item asking how often a child eats ice cream, then "hits others" will have a higher discrimination parameter than "eats ice cream" for externalizing problems. An item's difficulty parameter describes the trait level at which the probability of endorsing the item is 50%. For example, if a child is described as using alcohol or drugs, the child is likely to be higher in externalizing problems than children described as arguing. Thus, "uses alcohol or drugs" will have a higher difficulty parameter than "argues" for externalizing problems. In this context, a higher difficulty parameter reflects a higher, more severe level of externalizing problems, so henceforth we refer to the difficulty parameter as severity, consistent with prior studies (Krueger et al., 2004).

Based on items' discrimination and severity, one can determine how much information, or measurement precision, each item (and the assessment tool as a whole) provides at different trait levels. The goal of an assessment device is to maximize the amount of information that the items provide at the trait levels of interest. For a screening measure of children at risk for developing externalizing problems, the goal is to maximize information provided by the assessment tool at moderate trait levels (e.g., 0 to 1.5 *SD* above the full population mean) to identify at-risk children for further screening (Harford et al., 2013). On the other hand, to screen cases at a more extreme, clinical or diagnosable level, the goal is to maximize information at higher trait levels (e.g., 1.5 to 3 *SD* above the mean) in order to distinguish between clinical and subclinical symptomatology with greater confidence (Krueger et al., 2004).

IRT methods can evaluate the information coverage of individual items to inform the selection of clinically useful items, as defined by the trait levels of interest. Further, extending IRT to the study of items and how they relate to the construct of externalizing behavior problems at different ages and by different raters can advance understanding of the content space of externalizing problems and how it changes with context and development. Such an analysis improves our understanding of the construct and improves the reliability and validity of assessment. Ultimately, having a shorter questionnaire that removes poorly discriminating items or ensures item coverage across target trait levels may yield more measurement precision than a longer questionnaire, which is different from common assumptions based on classical test theory (Embretson & Reise, 2000).

Previous studies have examined questionnaire items' utility for measuring externalizing problems in children. Studts and van Zyl (2013) examined 18 externalizing problem items from parents' reports of 3- to 5-year-old children on the Behavior Problems Index and the Pediatric Symptom Checklist. The authors identified eight useful items for distinguishing clinical and subclinical externalizing problems in preschoolers (based on item information from 1.5 to 3 *SD* above the mean): "bullying/cruelty to others," "lack of remorse after misbehavior," "difficulty getting along with other children," "not being liked by other children," "deliberately breaking/destroying things," "fighting with other children," "blaming others," and "taking things that do not belong to him/her." Lambert et al. (2003) examined self-reports on the Externalizing scale of the Achenbach Child Behavior Checklist (CBCL) of Jamaican children ages 11–18 years. The authors identified three items that provided good information, six mediocre items, and 21 poor items, but they did not specify which were the "good" items.

Chorpita et al. (2010) identified six externalizing problem items that were useful for monitoring children's response to treatment, based on ratings of 8- to 12-year-old children on the Achenbach Youth Self-Report and by their parents on the CBCL (based on item information from 0 to 2 *SD* above the mean): disobedient at home or school, temper tantrums, argues, stubborn, threatens others, and destroys others' things. Wakschlag et al. (2014) examined parent-reported questionnaire items of 3- to 5-year-old children on the Multidimensional Assessment of Preschool Disruptive Behavior, which includes subdimensions of externalizing problems. However, the authors did not provide estimates of items' discrimination or diagnostic information, and did not examine the items in relation to the general externalizing problems factor. In addition, other studies have examined Attention-Deficit/Hyperactivity Disorder (ADHD) symptoms with IRT (Gomez, 2008; Gumpel, Wilson, & Shalev, 1998), but they did not examine general externalizing problems. Limitations of prior studies include that they were cross-sectional and involved only a single rater, with one exception that examined two raters (Chorpita et al., 2010). IRT has been used to develop briefer assessments of externalizing problems in adults (e.g., Patrick, Kramer, Krueger, & Markon, 2013) and children's response to treatment (Chorpita et al., 2010), and factor analysis has also been used to develop brief scales (Peterson & Zill, 1986; Zill, 1990), but to our knowledge, no studies have used IRT to develop brief assessments with the intended purpose of screening children with clinical-range externalizing problems.

The present article reports two longitudinal studies of item properties from annual mother and teacher reports of externalizing problems on the Achenbach scales from ages 5 to 13. The Achenbach scales are well-normed scales for children's behavior problems (Lambert et al., 2003). The Achenbach scales were developed using factor analysis, and scores on the Achenbach scales have good internal consistency, test–retest reliability, and interrater reliability. Interpretations of scores on the Achenbach scales have satisfactory content, criterion, and construct validity (Sattler & Hoge, 2006). As of 2014, the Achenbach scales had been used by over 15,000 authors in 9,000 studies from 80 countries, with 400–500 new publications every year (Bérubé & Achenbach, 2014). The Achenbach scales are also widely used in clinical contexts including mental health, school, and medical settings. Despite the widespread usage of the Achenbach CBCL and Teacher's Report Form (TRF) and the appearance of a few IRT studies of Achenbach questionnaires, this is the first study, to our knowledge, to examine items' sensitivity to clinical-range scores from both the CBCL and TRF longitudinally with IRT. Longitudinal data permit examining the stability or change in item properties and usefulness across time. The present report describes two studies using two independent samples, which is also unprecedented in prior IRT studies of child behavior problems, to our knowledge, although the Cole et al. (2011) study of depression in children used IRT with a dataset from multiple samples. We present the item properties for ages 5–8 and 9–13 separately for mothers and teachers to compare the useful items by rater and at different developmental eras because externalizing behaviors may appear different and have different meanings in middle childhood compared to the transition to adolescence (i.e., heterotypic continuity; Petersen, Bates, Dodge, Lansford, & Pettit, 2015). Based on the information provided by each item at subclinical to clinical trait levels, we identify an optimal subset of useful items that are sensitive to clinical-range scores on the Achenbach Externalizing scale across the two independent samples.

Study 1

Method

Participants. Children (N = 585) were recruited for the Child Development Project (Dodge, Bates, & Pettit, 1990) from two cohorts in 1987 and 1988 from three sites: Nashville, TN; Knoxville, TN; and Bloomington, IN. Children's parents were approached at random during kindergarten preregistration, on the first day of class, and by phone or mail. About 75% of parents approached agreed to participate. The schools and the sample represented families with a broad range of socioeconomic status, representative of the populations at the respective sites. The Hollingshead index of SES (M = 39.53, SD = 14.01, range: 8 to 66) reflected a broad range for the original sample, which was 52% male, 81% European American, 17% African American, and 2% of "other" ethnicity. Children were followed up annually with mothers' and teachers' ratings of the children's externalizing problems. The present study focuses on ratings of children's externalizing problems from 5 to 13 years of age.

Measures. Externalizing problems were measured by the Achenbach scales annually from ages 5–13: Mothers' scores came from the relevant factor of the CBCL (33 items; Achenbach, 1991a), and teachers' scores came from the TRF (35 items; Achenbach, 1991b).[1] Reporters rated whether a given behavior was "not true," "somewhat or sometimes true," or "very or often true" (scored 0, 1, and 2, respectively). Correlations, means, and standard deviations of mothers' and teachers' summed ratings of externalizing problems are in Table 1. Cronbach's alpha of externalizing problems for mothers' and teachers' ratings at each age are in Table 2. Rates of missingness ranged from 3%–32% (M = 21%) for mothers' reports, depending on the year, and from 2%–31% for teachers' reports (M = 18%). We focused on mothers' and teachers' reports because (a) parent and teacher report are common in the context of child clinical assessment, and (b) they were the only raters for whom we had annual ratings from ages 5–13 (the time frame of the present study for assessing externalizing problems with the Achenbach norms). For more information about missingness including rates of missingness of mothers' and teachers' reports of externalizing problems at each age and an attrition analysis (see Petersen et al., 2015).

As a validation of the set of items selected via IRT, we tested whether the same items from mothers' reports on the CBCL at ages 16 and 17 predicted later research diagnosis of conduct disorder at age 18. In Study 1, we only considered mothers' reports at ages 16 and 17 because we did not collect teachers' reports at age 16 or 17 and we did not collect the Achenbach scales at age 18. Conduct disorder was measured by an in-person assessment on the National Institute of Mental Health Diagnostic Interview Schedule (DIS; Robins et al., 1999) as administered by a specially trained interviewer to participants at age 18. The DIS was not administered at other ages. Interviews were conducted privately in the child's home or in the lab, depending on the adolescent's preference. Interviewers recorded participants' responses in a computer program designed to handle skip patterns depending on participants' responses (e.g., follow-up questions about specific aspects of a disorder were skipped if the participant did not meet the diagnostic criteria for having the disorder). Scores on the DIS have good convergent validity with clinical scales (Fantoni-Salvador & Rogers, 1997) and reliability (Compton & Cottler, 2004; Robins, Helzer, Croughan, & Ratcliff, 1981; Robins, Helzer, Ratcliff, & Seyfried, 1982). Data on conduct disorder diagnoses were missing for 25% of the sample. Of the 75% with data, 6% met criteria for conduct disorder diagnosis, consistent with epidemiological studies (Costello et al., 2005).

Statistical Analysis

IRT assumptions. First, we evaluated three IRT assumptions: (a) unidimensionality—the items have one predominant dimension reflecting the underlying (latent) trait (i.e., externalizing problems); (b) local independence—the items are uncorrelated when controlling for the latent dimension; and (c) monotonicity—the probability of endorsing a higher level on an item increases as the child's severity on externalizing problems increases. We evaluated the unidimensionality of externalizing problem ratings by each rater at each age by examining the percentage of variance accounted for by the first factor in maximum-likelihood exploratory factor analysis (EFA) using varimax rotation when restricting the extraction to one factor. It has been suggested that the first factor should account for at least 20% of the variance to meet the assumption of unidimensionality (Reckase, 1979). We present the proportion of variance explained by the first factor in EFA rather than CFA estimates of unidimensionality for simplicity and for easier comparison across years and raters. We evaluated the local independence of items by each rater at each age by evaluating the item correlations after partialing out the latent externalizing factor (the first EFA factor). It has been suggested that, when controlling for the latent factor, the absolute value of items' residual correlations should not exceed .20 (Reckase, 1979). We evaluated the monotonicity of items by examining the order of score categories in item response category characteristic curves from generalized partial credit models. In other words, we examined whether higher score categories (e.g., 2) corresponded to higher levels of externalizing problems compared to lower score categories (e.g., 0) for

[1] More recent editions of the CBCL and TRF have been published with some item changes (Achenbach & Rescorla, 2001). For instance, in the 2001 editions, there are separate items for alcohol use and drug use (whereas these are combined in the 1991 editions).

Table 1
Correlations, Means, and Standard Deviations of Mothers' and Teachers' Ratings of Externalizing Problems Across Time (Study 1)

Age (Years)	5	6	7	8	9	10	11	12	13
Mean	5.75	6.61	7.02	6.63	6.59	7.21	7.11	6.68	7.80
SD	8.67	9.66	10.42	10.09	10.08	10.11	10.56	9.81	11.60
5	.24	.56	.55	.52	.51	.47	.38	.41	.29
6	.66	.29	.57	.57	.56	.49	.39	.36	.36
7	.58	.69	.33	.61	.58	.57	.45	.40	.49
8	.62	.70	.72	.36	.63	.55	.49	.53	.50
9	.55	.66	.69	.74	.34	.62	.50	.51	.55
10	.58	.67	.70	.70	.76	.40	.53	.59	.44
11	.47	.58	.57	.68	.67	.70	.24	.52	.46
12	.50	.58	.61	.69	.69	.74	.75	.36	.53
13	.44	.55	.54	.64	.65	.65	.67	.77	.42
Mean	11.51	10.34	9.60	9.33	9.27	8.73	8.74	9.24	9.19
SD	7.02	7.02	6.80	7.46	7.41	7.28	7.12	7.14	7.18

Note. Means, *SD*s, and correlations above the diagonal refer to teacher-reported externalizing problems; mother-reported externalizing problems are below the diagonal. The diagonal represents the concurrent correlations between mother- and teacher-reported externalizing problems. All correlations are significant at $p < .001$ level.

all items. Generalized partial credit models (Muraki, 1992) were used for testing monotonicity because they do not restrict the order of score categories. Because of multiple testing stemming from the large number of items (612) examined across two raters and 9 years in 18 models and 9,819 item correlations, it was unlikely for all items for both raters at all years to strictly meet all IRT assumptions. Thus, we examined the degree to which the assumptions were supported for each rater at each age.

IRT models and parameters (item discrimination and severity). Questionnaire items of externalizing problems were analyzed with graded response models in IRT (Samejima, 1969) using the ltm package (Rizopoulos, 2006) in R 3.0 (R Development Core Team, 2009). The ltm package uses marginal maximum likelihood estimation, which uses all available data and provides valid inferences when data are missing at random or completely at random. Graded response models allow polytomous variables with more than two response categories (e.g., 0–2 Likert scale in the present study). The models estimated three parameters for each item: (a) discrimination (a_1); (b) severity for the threshold from 0–1 (b_1); and (c) severity for the threshold from 1–2 (b_2). We examined model fit for the graded response models using likelihood ratio tests (LRT) that compared the model fit (log-likelihood) of nested models that (a) constrained the discrimination parameter across items, and (b) allowed the discrimination parameter to vary across items. RMSEA and CFI estimates of model fit were obtained using the mirt package (Chalmers, 2012) in R. After settling on a best fitting model, we fit separate models for each rater and year.[2] The item properties from each individual model are in Tables S1–S4 of the supplementary appendix. Because of outliers in the distributions of item properties, we calculated robust averages across raters and years using the Hodges-Lehmann estimator (Hodges & Lehmann, 1963), which is the median of all pairwise means, and is a smooth version of the median (Rousseeuw & Croux, 1993) that approximates the population median.

Table 2
Evaluating IRT Assumptions: Unidimensonality and Local Independence (Study 1)

Rater	Age	Cronbach's alpha	Unidimensionality	Local independence
Mother	5	.86	19.4	.010
Mother	6	.88	19.9	.042
Mother	7	.88	19.4	.044
Mother	8	.90	24.1	.034
Mother	9	.90	24.5	.050
Mother	10	.90	23.6	.038
Mother	11	.89	22.6	.042
Mother	12	.89	21.8	.055
Mother	13	.90	24.1	.050
Teacher	5	.94	35.7	.116
Teacher	6	.95	37.4	.153
Teacher	7	.96	42.6	.144
Teacher	8	.95	41.4	.148
Teacher	9	.95	42.0	.176
Teacher	10	.95	40.6	.137
Teacher	11	.96	39.6	.163
Teacher	12	.95	39.2	.187
Teacher	13	.96	45.3	.200

Note. "Unidimensionality" column represents the percentage of variance accounted for by Factor 1 in EFA when restricting the extraction to one factor. "Local Independence" column represents the proportion of correlations among items whose absolute value was greater than .20 after partialling out the latent externalizing factor (the first EFA factor).

Item properties are depicted visually with item operation characteristic curves (see Figure 1), which represent the likelihood of endorsing a particular item threshold (1 or 2) as a function of one's level on the latent trait of externalizing problems (theta). Items' severity parameters are represented as the location of each curve's

[2] Although conducting longitudinal IRT was outside the scope of the current article, we fit abbreviated longitudinal IRT models in Mplus version 6.12 (Muthén & Muthén, 2010) to compare IRT parameter estimates from cross-sectional models with parameter estimates from longitudinal models. Parameter estimates from cross-sectional models were highly correlated with parameter estimates from longitudinal models. For instance, our parameters from a cross-sectional IRT model of mothers' ratings at age 5 and an abbreviated longitudinal IRT model from ages 5 to 8 were highly correlated for both discrimination (r = .981) and severity (r = .861) parameters. Thus, evidence suggests that our estimates were quite similar to longitudinal IRT estimates. Parameter estimates from the longitudinal IRT models are available by request.

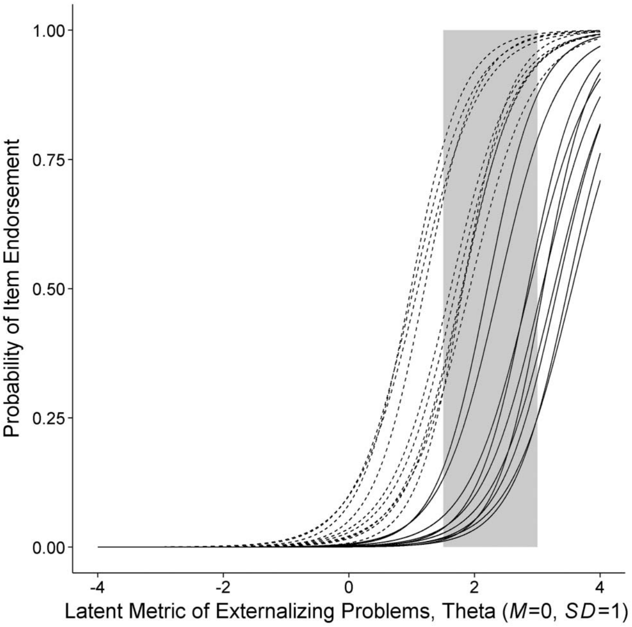

Figure 1. Item operation characteristic curves for the 10 selected useful items based on the items' information across the target trait levels (see gray box; 1.5 to 3 *SD* above the mean) as measured by mothers' and teachers' ratings from ages 5–13 (Study 1). Each item has two curves: one for each threshold (0 to 1 and 1 to 2). Curves represent the likelihood of endorsing an item with a score of 1 (dashed lines) or 2 (solid lines) as a function of one's level of externalizing problems.

midpoint on the *x*-axis, theta, with more severe items having midpoints further to the right (meaning that children of parents or teachers who endorse the item tend to have higher ratings of externalizing problems). Items' discrimination parameters are represented by the steepness of each curve's slope, with steeper slopes corresponding to greater item discrimination (meaning that as externalizing problems increase, one is more confident that the likelihood of endorsing the item increases).

Item information. Based on items' discrimination and severity, we calculated how much information (measurement precision; Dodd, De Ayala, & Koch, 1995) each item provided within the target clinical range of interest (1.5 to 3 *SD* above the mean). We chose this target range a priori because it distinguishes between subclinical (approximately < 2 *SD*) and clinical (approximately $>$ 2 *SD*) levels, consistent with prior studies (Studts & van Zyl, 2013) and the Achenbach norms (Achenbach, 1991a, 1991b). This target range approximately corresponds to the range from the 93rd to the 99th percentile on a nonclinical (i.e., standard normal z-score) distribution. Based on average item information across raters and years, items with low information were classified into several, nonmutually exclusive clusters describing possible reasons why the items did not provide much clinical information in this target range: (a) low discrimination ($a < 2$); (b) low severity ($b_1 < 1$ or $b_2 < 2$); or (c) high severity ($b_1 > 3$ or $b_2 > 5$). If the item provided high information (due to adequate discrimination and severity), it was classified as (d) useful (information > 1.4), where "useful" is defined as sensitive to clinical-range scores on the Achenbach Externalizing scale according to the Achenbach norms.

Selecting a subset of useful items. Our goal in selecting a subset of useful items was to identify items that provided high levels of measurement precision within the clinical-range scores of externalizing problems. Our preference was to select items that were filled out by both mothers on the CBCL and by teachers on the TRF for simplicity, purposes of comparison, and clinical utility. Based on these preferences, we derived a decision rule that allowed us to be fairly selective in choosing useful items. We selected an item if the item was administered to both raters and its average information was greater than 1.4. This decision rule was used because (a) it retained items that were administered to both mothers and teachers, (b) it included most of the best performing items, and (c) it minimized the number of items kept while ensuring adequate coverage across the target trait range. For efficient presentation, in Study 1, we focus on the common items identified in both Studies 1 and 2.

Reliability and psychometric properties of the selected items. We evaluated the reliability and psychometric properties of the scale using the selected items, including unidimensionality, internal consistency, test–retest reliability, interrater reliability, and overlap with the full scale.

Validation of the selected items. Our goal was to examine concurrent-criterion related validity of the selected items in relation to research diagnoses of conduct disorder. Because concurrent-criterion validity data were not available, we examined predictive-criterion related validity. We attempted to validate the selected items by comparing the classification accuracy of the full scale at ages 16 and 17 to the selected items from the CBCL in classifying later conduct disorder using the research diagnostic interview at age 18. Classification accuracy was measured by area under the curve (AUC) estimates from receiver operating characteristic curves, which examine the diagnostic utility of an assessment tool by evaluating the tradeoff between its sensitivity and specificity to predict the outcome. AUC represents the probability that a randomly selected person meeting the diagnostic threshold (for conduct disorder) will have a higher test result (i.e., more externalizing problems) than a randomly selected person who does not meet the cutoff. In general, a higher AUC represents a better performing diagnostic test (range: 0–1, chance = 0.5). AUC values were compared with DeLong's test (DeLong, DeLong, & Clarke-Pearson, 1988) in the pROC package (Robin et al., 2011) in R. We also provide other diagnostic accuracy estimates of the selected items, including sensitivity, specificity, positive predictive value, and negative predictive value (for definitions, see Akobeng, 2007a). We examined correlations of the full scale and selected items with research diagnoses of conduct disorder, and corrected for attenuation resulting from measurement error (Fan, 2003). For mother-reported externalizing problem scores, we used the estimate of test–retest reliability of the CBCL Externalizing Problems scale reported by Achenbach (r = .93; 1991a) to correct for attenuation. For DIS conduct disorder diagnoses, we used the estimate of test–retest reliability for conduct disorder on the DIS provided by Compton and Cottler (κ = .51; 2004) to correct for attenuation.

Results

IRT assumptions. First, we examined the degree to which the IRT assumptions were supported for the items of the Externalizing scale for each rater at each age. Table 2 presents (a) the percentage of variance accounted for by the first factor extracted from EFA of the Achenbach externalizing problem items (unidimensionality), and (b) the proportion of item correlations whose absolute value was greater than .20 after partialing out the latent externalizing factor (local independence). The assumption of unidimensionality was generally met, even though the first factor explained only slightly less than 20% of the variance in mothers' reports at ages 5, 6, and 7 (19.4%, 19.9%, and 19.4%, respectively). Prior research has also suggested that IRT parameter estimates are robust to violations of unidimensionality (Harrison, 1986). Moreover, the average of eigenvalues across years for the first four factors were: (a) 7.88, (b) 2.12, (c) 1.63, and (d) 1.43 for mothers' reports; and (a) 14.68, (b) 2.30, (c) 1.68, and (d) 1.48 for teachers' reports. The eigenvalues suggest that the first factor accounted for considerably more variance than additional factors, and that the items were "unidimensional enough" for IRT.

Regarding local independence, the proportion of correlations among items whose absolute value was greater than .20 after partialing out the first EFA factor ranged from .01 to .05 for mothers' ratings, depending on the year, and from .12 to .20 for teachers' ratings. This finding suggests that some items at some years for teachers' ratings, in particular, may have been related to each other in ways other than the externalizing factor exclusively. Nevertheless, IRT is robust to low and moderate violations of the local independence assumption (Fennessy, 1995). Despite evidence of potential local nonindependence, we examined the Achenbach scales with IRT to shed light on their item properties because the CBCL and TRF are among the most widely used checklists for children's behavior problems. We kept the same items across years for comparability.

We also examined items for monotonicity but found no violations in the expected order of score categories in terms of item threshold severity. Finally, we examined model fit. Models with different discrimination parameters across items fit better than models with fixed discrimination parameters across items (*p*s $< .001$) with the exception of mother-reported externalizing problems at age 10 (LRT $= 18.03$, $df = 31$, $p = .969$), so subsequent models allowed different items to have different discrimination parameters. Model fit was generally good according to RMSEA (ranging from .04 to .06, depending on the year) and CFI (.95 to .99).

Item discrimination and severity. We fit separate IRT models for mothers' and teachers' reports at each age. We then calculated robust averages of item properties, including discrimination and severity, across ages 5–8 and 9–13, and across all years and both raters (see Table 3). Some items' severity levels for mothers' reports were likely too high for adequate endorsement of *very or often true* (2) to provide severity estimates for this threshold: runs away (ages 5–8), truancy (5–8), use of alcohol/drugs (5–8), and vandalism (9–13).

There was an outlier in the parameter estimates in Table 3: the severity (b_1) for teachers' reports of alcohol and drug use from ages 5–8 (-19.95). Severity levels for this item were only available at ages 6 and 8 during this timeframe (the severity was likely too high at ages 5 and 7 to provide estimates for item severity). The severity (b_1) of alcohol and drug use was 4.64 at age 6 and -44.52 at age 8. The severity level at age 8 was an implausible outlier. Because only two values were available, the average for ages 5–8 was not a robust average but rather a simple mean, which is affected by outliers. The item's average severity across all years and raters (b_1: 4.44) was a robust average that was not as affected by outliers. The likely high severity of alcohol and drug use from ages 5–8 may reflect the possibility that it is a developmentally inappropriate item for young children.

Item information. The estimates of item information from 1.5 to 3 *SD* above the mean for mothers' and teachers' ratings from ages 5–8 and 9–13 are in Table 4. Several patterns are notable. First, in general, teachers' items provided more information in the target range than did mothers' items. Examination of Table 3 suggests that this may be because teachers' items generally had higher discrimination than did mothers' items. Higher discrimination of teachers' items than mothers' items may reflect the higher internal consistency and unidimensionality of teachers' items (see Table 2). Second, although most items' information stayed relatively constant from ages 5–8 to ages 9–13, some items became more informative for this target range at later ages (e.g., use of alcohol/drugs), and some became less informative at later ages (e.g., vandalism). Examination of Table 3 suggests that the use of alcohol and drugs (by mothers' report) became less severe with age and that vandalism became less discriminating and more severe with age.

Selecting a subset of useful items. The decision rule for selecting a subset of useful items resulted in selecting 11 items from the useful cluster: mean to others, destroys others' things, disobedient at school, fights, lies and cheats, attacks people, screams, swearing/obscene language, temper tantrums, threatens others, and loud. Of these 11 items, all but one (disobedience at school) were also identified as useful in the independent sample in Study 2, so we selected these 10 items for further analysis. Item operation characteristic curves for the 10 selected useful items, based on their average item properties across raters and years, are in Figure 1. Figure 1 shows that the items had good coverage of measurement precision across the target trait levels (1.5–3 *SD*).

Reliability and psychometric properties of the selected items. The percentage of variance accounted for by the first factor in the EFA for the mothers' selected items ranged from 23.2% to 36.3% ($M = 30.5\%$), depending on the year, and from 42.0% to 56.4% ($M = 49.1\%$) on the teachers' selected items. The internal consistency of scores on the mothers' selected items, measured by Cronbach's alpha, ranged from $\alpha = .747$ to $.849$ ($M = .809$), compared with $\alpha = .863$ to $.902$ ($M = .889$) for the full Externalizing scale. The internal consistency of scores on the teachers' selected items ranged from $\alpha = .872$ to $.926$ ($M = .902$), compared with $\alpha = .941$ to $.962$ ($M = .953$) for the full scale. The annual test–retest reliability of scores on the mothers' selected items, measured by Pearson correlation, ranged from $r = .575$ to $.752$ ($M = .670$, *p*s $< .001$), compared with $r = 659$ to $.774$ ($M = .719$) for the full scale. The annual test–retest reliability of scores on the teachers' selected items ranged from $r = .416$ to $.572$ ($M = .499$, *p*s $< .001$), compared with $r = .514$ to $.632$ ($M = .565$) for the full scale. The mother–teacher interrater reliability on the scale made from the selected items ranged from $r = .158$ to $.448$ ($M = .282$, *p*s $< .001$), compared with $r = .245$ to $.428$ ($M = .333$) for the full scale. The correlation of the selected items with the full scale ranged from $r = .880$ to $.908$ ($M = .892$, *p*s $< .001$) for the mothers' selected items and from $r = .898$ to $.922$ ($M = .905$, *p*s $< .001$) for the teachers' selected items.

Validation of the selected items. We examined the mothers' ratings on the full Externalizing scale compared to the selected items at ages 16 and 17 in predicting research diagnoses of conduct disorder at age 18. Conduct disorder diagnoses were associated with the full scale ($r[221] = .123$, $p = .067$; disattenuated: $r = .179$) and the selected items ($r[221] = .148$, $p = .027$; disattenu-

Table 3
Item Properties (Discrimination and Severity) of Items From the Child Behavior Checklist and Teacher's Report Form From Ages 5–13 (Study 1)

		Ages 5–8						Ages 9–13								
		Mother			Teacher			Mother			Teacher			Average		
Item	Short wording	a	b_1	b_2	a	b_1	b_2	a	b_1	b_2	a	b_1	b_2	a	b_1	b_2
3	Argues	1.67	−1.20	1.41	2.85	.78	2.01	1.89	−.99	1.15	3.06	.45	1.66	2.41	−.22	1.53
6	Defiant				2.69	1.44	2.48				3.31	.78	1.85	2.99	1.15	2.12
7	Brags	.99	−.22	3.02	1.57	1.40	3.00	1.35	−.11	2.72	1.88	1.14	2.43	1.47	.59	2.82
16	**Mean to others**	1.69	1.25	3.50	2.88	1.41	2.55	1.99	1.14	3.18	3.11	.95	2.09	2.43	1.20	2.85
19	Demands attention	1.47	−.22	1.85	2.03	.81	2.09	1.52	.03	2.23	2.18	.73	1.71	1.77	.38	2.02
20	Destroys own things	1.58	1.33	3.27	1.83	2.38	3.81	1.80	1.79	3.60	2.11	2.28	3.03	1.82	1.94	3.43
21	**Destroys others' things**	1.70	1.50	3.71	2.61	2.08	3.25	2.08	1.71	4.03	2.47	1.83	3.09	2.25	1.81	3.48
22	Disobedient at home	1.82	−.43	2.24				1.98	−.18	2.48				1.89	−.25	2.37
23	Disobedient at school	1.28	1.17	3.90	3.45	.95	2.15	1.77	.98	3.08	3.72	.69	1.84	2.56	.97	2.65
24	Disturbs others				3.16	.52	1.83				3.49	.37	1.56	3.34	.45	1.73
26	Lacks guilt	1.50	.76	2.63	2.39	1.15	2.18	1.41	.81	3.04	2.85	.83	1.75	2.05	.89	2.43
27	Jealous	1.36	.10	2.28	1.63	1.81	3.42	1.47	.24	2.41	1.69	1.42	3.14	1.54	.90	2.80
37	**Fights**	1.38	1.86	4.20	2.68	1.47	2.57	1.72	1.74	3.70	2.42	1.31	2.67	2.04	1.62	3.26
39	Bad companions	1.22	1.94	4.34	1.78	1.44	2.71	1.36	1.45	3.81	1.86	.92	2.29	1.58	1.42	3.24
43	**Lies and cheats**	1.64	.71	3.12	2.06	1.60	2.88	2.09	.75	2.85	2.11	1.21	2.75	2.01	1.10	2.87
53	Talks out of turn				2.38	.47	1.68				2.60	.31	1.43	2.41	.45	1.61
57	**Attacks people**	1.67	1.93	3.78	2.41	1.70	3.12	1.82	2.01	3.99	2.82	1.49	2.60	2.19	1.80	3.32
63	Prefers older kids	.61	−.53	4.11	.55	4.11	8.37	.79	.44	3.93	.72	3.24	6.18	.68	1.82	5.27
67	Runs away	1.00	4.68	*				1.36	3.93	4.40				1.24	4.15	4.40
67	Disrupts class				3.27	.91	1.90				3.65	.56	1.58	3.38	.78	1.79
68	**Screams**	1.66	1.25	3.10	2.51	2.41	3.35	1.74	1.23	2.86	2.61	2.08	2.96	2.06	1.71	3.07
72	Sets fires	1.08	5.04	*				1.53	3.67	4.58				1.32	4.06	4.58
72	Messy work				.89	1.08	3.12				1.01	.81	2.80	.94	.91	3.03
74	Shows off	1.56	−.48	1.98	2.20	1.05	2.31	1.55	−.15	2.24	2.48	.70	1.88	1.89	.35	2.16
76	Explosive				2.70	1.78	2.77				2.98	1.39	2.31	2.90	1.63	2.58
77	Easily frustrated				1.86	1.52	2.85				2.37	1.04	2.26	2.12	1.38	2.52
81	Steals at home	1.27	3.10	4.82				1.86	2.39	3.93				1.59	2.60	4.33
82	Steals outside home	1.31	3.29	4.72	1.99	2.47	3.17	1.76	2.84	4.57	1.68	2.65	3.73	1.75	2.77	3.96
86	Stubborn, irritable	1.73	−.20	2.03	2.01	1.31	2.70	1.89	−.24	1.96	2.20	.95	2.40	1.95	.46	2.26
87	Sudden mood changes	1.62	1.07	2.95	1.70	1.87	3.13	1.52	.63	3.12	2.01	1.31	2.53	1.74	1.19	2.98
90	**Swearing, obscenity**	1.54	2.01	4.00	2.17	2.46	3.70	1.78	1.57	3.46	2.46	1.74	3.07	1.99	1.92	3.55
93	Talks too much	1.28	−.12	1.91	2.00	.55	1.85	1.03	.22	2.71	2.46	.37	1.56	1.61	.28	2.00
94	Teases	1.34	.84	3.28	2.26	1.67	3.07	1.52	.61	2.94	2.52	1.05	2.23	1.91	1.01	2.94
95	**Temper tantrums**	2.03	.40	2.05	2.60	1.74	2.70	2.32	.49	2.03	2.92	1.35	2.19	2.44	.99	2.20
96	Thinks about sex too much	1.24	3.43	4.79				1.45	2.92	4.50				1.36	3.15	4.59
97	**Threatens others**	2.19	1.77	3.41	2.82	2.10	3.14	2.53	1.80	3.09	2.84	1.52	2.64	2.66	1.82	3.09
98	Tardy				.46	5.70	8.92				.83	3.64	5.15	.75	4.68	7.70
101	Truancy	1.53	4.49	*	.54	7.73	10.81	1.66	3.65	4.42	.70	4.05	5.43	1.09	4.28	5.98
104	**Loud**	1.78	.71	2.46	2.38	1.54	2.41	1.80	.74	2.41	2.74	1.08	2.04	2.13	1.02	2.38
105	Alcohol, drugs	8.50	5.22	*	.63	−19.95	4.94	5.38	4.01	5.16	1.39	5.38	9.78	1.38	4.44	6.26
106	Vandalism	3.29	2.89	3.68				2.26	3.39	*				2.75	3.16	3.68

Note. a = discrimination; b_1 = severity for endorsement of 1 on item; b_2 = severity for endorsement of 2 on item; * = no variability or not enough variability for convergence (high severity). Average represents the robust average across all years and both mothers' and teachers' ratings. Items in bold represent the 10 useful items selected.

ated: r = .215) at age 16, and with the full scale (r[393] = .201, $p < .0001$; disattenuated: r = .291) and the selected items (r[393] = .203, $p < .0001$; disattenuated: r = .295) at age 17. The full scale had an AUC of .705 at age 16 and .713 at age 17 in predicting later conduct disorder diagnosis. The selected items had an AUC of .703 at age 16 and .684 at age 17 in predicting later conduct disorder. The difference in AUCs between the selected items and full scale in predicting later conduct disorder was not significant at age 16 (z = 0.04, p = .968) or at age 17 (z = 0.84, p = .399). Thus, the selected items performed as well as the full Externalizing scale in predicting later conduct disorder. The diagnostic accuracy estimates at multiple cut points of the 10 selected items for predicting conduct disorder diagnosis are in Table 5.

Discussion

Study 1 identified 11 useful items (i.e., sensitive to clinical-range scores) that were administered to both mothers and teachers. Ten of these 11 items were also identified as useful in Study 2. Scores on a scale using these 10 items performed as well as scores on the full Externalizing scale in predicting later research diagnoses of conduct disorder.

Table 4
Item Information From 1.5 to 3 SD Above the Mean on the Latent Metric of Externalizing Problems From the Child Behavior Checklist and Teacher's Report Form From Ages 5–13 (Study 1)

		Ages 5–8		Ages 9–13			
Item	Short wording	Mother	Teacher	Mother	Teacher	Average	Cluster
16	**Mean to others**	.92	2.82	1.26	2.44	1.84	Useful
21	**Destroys others' things**	1.00	2.42	1.27	2.11	1.65	Useful
37	**Fights**	.69	2.65	1.00	2.01	1.58	Useful
43	**Lies and cheats**	.88	1.66	1.36	1.51	1.40	Useful
57	**Attacks people**	1.05	2.13	1.19	2.66	1.74	Useful
68	**Screams**	1.05	2.19	1.15	2.15	1.52	Useful
90	**Swearing, obscenity**	.87	1.49	1.14	2.28	1.41	Useful
95	**Temper tantrums**	1.32	2.66	1.56	2.53	2.01	Useful
97	**Threatens others**	1.67	2.72	2.36	2.81	2.47	Useful
104	**Loud**	1.17	2.12	1.17	1.99	1.60	Useful
6	Defiant		2.67		2.28	2.46	Useful
23	Disobedient at school	.55	3.00	1.05	2.64	1.88	Useful
24	Disturbs others		2.32		1.70	2.09	Useful
76	Explosive		2.86		2.72	2.77	Useful
77	Easily frustrated		1.40		1.55	1.48	Useful
67	Disrupts class		2.55		1.84	2.34	Useful
3	Argues	.55	2.40	.56	1.48	1.15	Low severity
26	Lacks guilt	.87	1.94	.73	1.66	1.31	Low severity
53	Talks out of turn		1.39		.97	1.27	Low severity
86	Stubborn, irritable	.96	1.54	1.06	1.50	1.26	Low severity
106	Vandalism	1.76		.73		1.12	High severity
20	Destroys own things	.94	1.25	1.21	1.39	1.19	Low discrimination
39	Bad companions	.57	1.29	.66	1.18	.95	Low discrimination
81	Steals at home	.51		1.16		.89	Low discrimination
82	Steals outside home	.45	1.36	1.00	1.01	1.10	Low discrimination
87	Sudden mood changes	.97	1.17	.86	1.42	1.10	Low discrimination
94	Teases	.63	1.93	.79	1.69	1.25	Low discrimination, borderline low severity
7	Brags	.36	1.01	.66	1.29	.84	Low discrimination, low severity
19	Demands attention	.68	1.36	.82	1.11	.97	Low discrimination, low severity
22	Disobedient at home	1.06		1.24		1.13	Low discrimination, low severity
27	Jealous	.69	1.09	.80	1.11	.90	Low discrimination, low severity
93	Talks too much	.58	1.13	.43	.97	.76	Low discrimination, low severity
72	Messy work		.34		.42	.38	Low discrimination, low severity
74	Shows off	.79	1.70	.78	1.29	1.12	Low discrimination, low severity
63	Prefers older kids	.15	.10	.25	.17	.17	Low discrimination, high severity
67	Runs away	.16		.30		.24	Low discrimination, high severity
72	Sets fires	.15		.41		.26	Low discrimination, high severity
96	Thinks about sex too much	.37		.64		.52	Low discrimination, high severity
98	Tardy		.05		.34	.27	Low discrimination, high severity
101	Truancy	.31	.07	.50	.15	.26	Low discrimination, high severity
105	Alcohol, drugs	.01	.00	2.10	.09	.22	Low discrimination, high severity

Note. Items ordered by cluster. Items in bold represent the 10 useful items selected. Average represents the robust average across all years and both mothers' and teachers' ratings.

Study 2

Method

Participants. Children were recruited for the Fast Track Project (Conduct Problems Prevention Research Group, 1992) from high-risk schools (in neighborhoods with high rates of crime, poverty, and low parental education) at one of four sites: Durham, NC; Nashville, TN; Seattle, WA; and central Pennsylvania. Schools were randomly assigned to intervention or control conditions. Children were screened and selected if they were in the top 10% of disruptive behavior based on parent and teacher report (for further details, see Lochman & The Conduct Problems Prevention Research Group, 1995). Children from schools assigned to the intervention condition numbered 445, and 446 were in the control condition. The sample also included a normative subsample of 379 children (composed of about 10 children within each decile of behavior problems at each school), of whom 79 were also part of the control condition. Among the full sample (N = 1,199), 63% were male and 49% were African American. The Hollingshead index of SES (M = 25.26, SD = 12.91, range: 4.5 to 66) reflected a broad range. Of the full sample, 53% came from single-parent families and 29% of mothers had not graduated from high school. The present study focuses on mothers' and teachers' annual ratings of children's externalizing problems from 5 to 13 years of age (though mothers did not provide ratings at ages 8 and 11).

Table 5
Diagnostic Accuracy Estimates of the 10 Selected Useful Items in Relation to Diagnoses of Conduct Disorder Based on a Research Diagnostic Interview

	Cutoff	Sensitivity	Specificity	PPV	NPV
Study 1					
Predictions from age 16 to adolescent DIS	1	.800	.455	.065	.980
	2	.700	.667	.090	.979
	3	.600	.765	.107	.976
Predictions from age 17 to adolescent DIS	1	.833	.504	.098	.979
	2	.458	.693	.088	.952
	3	.375	.787	.102	.951
Study 2					
Predictions from age 11 to child DISC	1	.913	.327	.072	.985
	2	.848	.447	.080	.981
	3	.783	.520	.085	.977
	4	.674	.588	.085	.969
	5	.609	.651	.090	.967
	6	.587	.696	.099	.967
	7	.587	.730	.110	.969
Predictions from age 11 to parent DISC	1	.919	.330	.095	.982
	2	.887	.454	.111	.981
	3	.806	.530	.117	.973
	4	.710	.597	.119	.964
	5	.661	.658	.129	.962
	6	.581	.701	.130	.956
	7	.516	.730	.128	.951

Note. PPV = positive predictive value; NPV = negative predictive value; DIS = Diagnostic Interview Schedule; DISC = Diagnostic Interview Schedule for Children. Research diagnoses of conduct disorder came from the DIS (Study 1) and DISC (Study 2). The DIS in Study 1 was administered to adolescents at age 18. The DISC in Study 2 was administered to children and their parents at age 11. The predictor scores were from mothers' reports on the 10 selected items in Study 1 and from teachers' reports on the selected items in Study 2.

Measures. As in Study 1, externalizing problems were measured by the Externalizing scale of the Achenbach scales (see Footnote 1). Mothers reported on the CBCL (Achenbach, 1991a) at ages 5, 6, 7, 9, 10, and 12. Teachers reported on the TRF (Achenbach, 1991b) annually from ages 5 to 13. At ages 6 and 7, the TRF was not administered to the normative subsample. At ages 11, 12, and 13, a partial set of TRF items was administered, and one item from the Externalizing scale was not administered (messy work). At ages 11, 12 and 13, some children (10%) had multiple teachers fill out the TRF. For these children, we used the maximum value for each item across teachers to avoid underreporting by a single informant, consistent with prior studies of conduct problems (e.g., Frick et al., 2003). Rates of missingness ranged from 1% to 16% (M = 9%) for mothers' reports, depending on the year, and from 13% to 28% for teachers' reports (M = 20%).

As a validation of the set of items selected via IRT, we tested how well the items from teachers' reports on the TRF classified diagnoses of conduct disorder based on parent and child interview at age 11 (the only year from ages 5 to 13 when diagnoses were available based on both parent and child interview). Conduct disorder was measured by the National Institute of Mental Health Diagnostic Interview Schedule for Children (DISC), an interview administered to both parents (Shaffer & Fisher, 1997) and children (Shaffer et al., 1996) using a laptop computer. Interviews took place in the child's home with the primary parent, usually the mother, during the summer following sixth grade. The interview asked children whether they experienced symptoms within the prior month, and asked parents about their child within the prior 6 months. Data on conduct disorder diagnoses were missing for 16% of the sample. Of the 84% with data, 7% met criteria for conduct disorder based on parent interview and 5% met criteria based on child interview, consistent with epidemiological studies (Costello et al., 2005). Scores on the DISC have good convergent validity with clinician's diagnoses (Schwab-Stone et al., 1996; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000) and reliability (Jensen et al., 1995).

Statistical analysis. Statistical analysis procedures were similar to those in Study 1.

Selecting a subset of useful items. Because item information estimates were lower, on average, than in Study 1, we lowered the threshold for item selection to keep items whose average information was greater than 0.9 (and administered to both raters). As in Study 1, this decision rule was used because (a) it retained items that were administered to both mothers and teachers, (b) it included most of the best performing items, and (c) it minimized the number of items kept while ensuring adequate coverage across the target trait range.

Validation of the selected items. We attempted to validate the selected items by comparing the classification accuracy of the full Externalizing scale to the selected items from the TRF in classifying conduct disorder diagnoses based on parent and child interview at age 11. For teacher-reported externalizing problem scores, we used the estimate of test–retest reliability of the TRF Externalizing Problems scale reported by Achenbach (r = .92; 1991b) to disattenuate correlations of the selected items and full scale with DISC conduct disorder diagnoses. For DISC conduct disorder diagnoses, we used the estimates of test–retest reliability for conduct disorder on the DISC based on parent (κ = .56) and child interview (κ = .64) provided by Schwab-Stone et al. (1996) to correct for attenuation.

Results

IRT assumptions. Estimates of internal consistency, unidimensionality, and local independence are presented in Table S5 of the supplementary appendix. We also examined monotonicity as in Study 1. We found no severe violations of assumptions, although (as in Study 1), there was evidence of possible local nonindependence for teachers' ratings. This suggests that some items at some years for teachers' ratings may have been related to each other in ways other than the externalizing factor exclusively. Nevertheless, IRT is robust to low and moderate violations of the local independence assumption (Fennessy, 1995). Model fit was generally good according to RMSEA (ranging from .05 to .08, depending on the year) and CFI (.95 to .98).

Selecting a subset of useful items. Item properties for every year and rater are in supplementary appendixes S6–S9 (discrimination, a_1: Table S6; severity, b_1: Table S7; severity, b_2: Table S8; information: Table S9). Robust averages of item properties across year and rater are in Table 6. The decision rule for selecting a subset of useful items resulted in selecting 12 items: mean to others, destroys own things, destroys others' things, fights, lies and cheats, attacks people, screams, swearing/obscene language, teases, temper tantrums, threatens others, and loud. Of these 12 items, all but two (destroy own things and teases) were also identified as useful in the independent sample in Study 1, so we selected these 10 items for further analysis.

Reliability and psychometric properties of the selected items. The percentage of variance accounted for by the first factor in the EFA for the mothers' selected items ranged from 31.1% to 39.6% (M = 36.2%), depending on the year, and from 48.9% to 52.4% (M = 50.7%) on the teachers' selected items. The internal consistency of scores on the mothers' selected items ranged from α = .816 to .866 (M = .847), compared with α = .890 to .929 (M = .911) for the full Externalizing scale. The internal consistency of scores on the teachers' selected items ranged from α = .903 to .915 (M = .909), compared with α = .954 to .961 (M = .958) for the full scale. The annual test–retest reliability of scores on the mothers' selected items ranged from r = .636 to .692 (M = .666, ps < .001), compared with r = .693 to .747 (M = .718) for the full scale. The annual test–retest reliability of scores on the teachers' items ranged from r = .347 to .544 (M = .484, ps < .001), compared with r = .394 to .602 (M = .545) for the full scale. The mother–teacher interrater reliability of scores on the selected items ranged from r = .231 to .326 (M = .264, ps < .001), compared with r = .277 to .359 (M = .303) for the full scale. The correlation of the selected items with the full Externalizing scale ranged from r = .912 to .932 (M = .926, ps < .001) for the mothers' selected items and from r = .927 to .943 (M = .936, ps < .001) for the teachers' selected items.

Validation of the selected items. We examined the teachers' ratings on the full Externalizing scale compared to the selected items in classifying diagnoses of conduct disorder at age 11. Conduct disorder diagnoses based on parent interview were associated with scores on the full scale (r[866] = .173, p < .0001; disattenuated: r = .241) and the selected items (r[866] = .178, p < .0001; disattenuated: r = .247). Conduct disorder diagnoses based on child interview were also associated with scores on the full scale (r[854] = .138, p < .0001; disattenuated: r = .180) and the selected items (r[854] = .166, p < .0001; disattenuated: r = .216). Scores on the full scale and selected items had an AUC of .698 and .705, respectively, in classifying conduct disorder based on parent interview. The difference in AUCs between the selected items and full scale in classifying conduct disorder based on parent interview was not significant (z = −0.48, p = .632). Scores on the full Externalizing scale and selected items had an AUC of .673 and

Table 6
Average Item Properties (Discrimination, Severity, and Information) Across Raters on the Child Behavior Checklist and Teacher's Report Form From Ages 5–13 (Study 2)

Item	Short wording	a	b_1	b_2	Information
3	Argues	2.28	−.71	.95	.42
6	Defiant	2.88	.03	1.15	.79
7	Brags	1.45	.10	2.06	.72
16	**Mean to others**	2.31	.30	1.84	1.26
19	Demands attention	1.73	−.32	1.06	.48
20	Destroys own things	1.60	1.20	2.65	1.02
21	**Destroys others' things**	2.03	1.00	2.43	1.49
22	Disobedient at home	2.10	−.55	1.92	1.32
23	Disobedient at school	2.73	−.28	1.40	.71
24	Disturbs others	2.75	−.65	.84	.40
26	Lacks guilt	1.78	−.02	1.69	.62
27	Jealous	1.23	.38	2.34	.59
37	**Fights**	2.13	.54	1.96	1.28
39	Bad companions	1.46	.16	1.84	.64
43	**Lies (and cheats)**	1.79	.22	1.90	1.00
53	Talks out of turn	2.06	−.62	.76	.36
57	**Attacks people**	2.18	1.01	2.41	1.57
63	Prefers older kids	.81	1.21	3.57	.27
67	Runs away	1.54	3.19	4.43	.61
67	Disrupts class	3.26	−.24	.92	.45
68	**Screams**	1.88	1.12	2.43	1.36
72	Sets fires	1.34	3.07	4.38	.50
72	Messy work	.70	−.40	2.30	.20
74	Shows off	1.70	−.22	1.43	.67
76	Explosive	2.66	.48	1.54	1.30
77	Easily frustrated	2.01	.19	1.40	.84
81	Steals at home	1.44	1.86	3.71	.83
82	Steals outside home	1.45	1.76	3.16	.84
86	Stubborn, irritable	1.84	−.30	1.46	.77
87	Sudden mood changes	1.65	.28	2.00	.88
90	**Swearing, obscenity**	1.77	.94	2.41	1.14
93	Talks too much	1.39	−.43	1.28	.39
94	Teases	1.87	.03	1.74	.93
95	**Temper tantrums**	2.27	.30	1.52	1.00
96	Thinks about sex too much	1.28	2.50	4.16	.57
97	**Threatens others**	2.70	.86	2.11	1.99
98	Tardy	.58	2.83	5.67	.14
101	Truancy	.72	4.17	6.22	.15
104	**Loud**	1.84	.39	1.75	.93
105	Alcohol, drugs	1.09	5.58	6.06	.19
106	Vandalism	1.98	2.48	3.62	1.29

Note. a = discrimination; b_1 = severity for endorsement of 1 on item; b_2 = severity for endorsement of 2 on item. Average represents the robust average across all years and both mothers' and teachers' ratings. "Information" represents item information from 1.5 to 3 *SD* above the mean. Items in bold represent the 10 useful items selected.

.702, respectively, in classifying conduct disorder based on child interview. The difference in AUCs between the selected items and full scale in predicting later conduct disorder was significant ($z = -2.90$, $p = .004$). Thus, scores on the selected items performed as well as scores on the full Externalizing scale in classifying conduct disorder based on parent interview, and performed better than scores on the full scale in classifying conduct disorder based on child interview. Scores on the selected items were moderately accurate in classifying conduct disorder (AUC $>$ .70; Akobeng, 2007b). The diagnostic accuracy estimates at multiple cut points of the 10 selected items for predicting conduct disorder diagnosis are in Table 5.

Discussion

Study 2 identified 12 useful items (i.e., sensitive to clinical-range scores) that were administered to both mothers and teachers. Ten of these 12 items were also identified as useful in Study 1. Scores on a scale using these 10 items performed as well as scores on the full Externalizing scale in classifying research diagnoses of conduct disorder based on parent interview. Moreover, scores on a scale using these 10 items performed *better* than scores on the full scale in classifying research diagnoses of conduct disorder based on child interview.

General Discussion

The present studies described the item properties from mother and teacher reports of externalizing problems on the Achenbach scales from ages 5 to 13 in two independent samples. We compared the useful items by rater and at different developmental eras. Based on the information provided by each item at subclinical to clinical trait levels, we selected an optimal subset of useful items across both samples that were sensitive to clinical-range scores on the Achenbach Externalizing scale.

The 10 items selected (externalizing problems that involve meanness to others, destroying others' things, fighting, lying and cheating, attacking people, screaming, swearing/obscene language, temper tantrums, threatening people, and loud) meet several requirements for clinical practicality for a screening assessment of externalizing disorders. The items are: (a) few, (b) developmentally appropriate, (c) rater appropriate, and (d) maximally informative for whether a child has clinical or subclinical levels of externalizing problems. Although the items may have a reduced representation of the construct of externalizing problems compared with the full scale, the items we identified appear to reflect some of the most overt externalizing behaviors. Overt behaviors may be advantageous for a brief screening device, and could result in greater interrater reliability of clinically significant behaviors.

Some of the useful items we identified are consistent with items identified from prior studies (though the exact wording of the items may differ slightly). Some items we identified were consistent with useful screening items for preschoolers' externalizing problems identified by Studts and van Zyl (2013): "fights," "mean to others," and "destroys others' things." Although Lambert et al. (2003) did not calculate items' information based on the goals of a diagnostic screening measure, several useful items that we identified had relatively high discrimination values ($a \geq 1$; i.e., they were highly relevant to the construct of externalizing behavior) among self-reports of 11- to 18-year-old Jamaicans: mean to others, destroys others' things, lies and cheats, threatens others, and loud.

In addition, researchers have developed a short form of the CBCL/TRF, the Brief Problem Monitor (BPM; Achenbach, McConaughy, Ivanova, & Rescorla, 2011), using factor analysis and IRT (Chorpita et al., 2010). Although Chorpita et al. (2010) did not calculate items' information based on the goals of a diagnostic screening measure, the following useful items that we identified were consistent with useful items for 8- to 12-year-old children: temper tantrums, threatens others, and destroys others' things. Nevertheless, the BPM was not developed with the purpose of identifying children with clinical-range scores, so the scale composed by the select set of items identified in the present study is nonredundant and useful because it serves a different purpose than the BPM. Different purposes (e.g., symptom severity target range) and/or different ages of the sample might account for differences in the items we identified compared with those identified in prior studies. Alternatively, the items selected could depend on the sample, so future studies might collectively examine the items that we and others have identified. Nevertheless, we identified considerably similar items in two large, independent samples using longitudinal data from both mother and teacher reports. In sum, there is prior support for many of the items we identified, using different samples, ages, and methods, suggesting that the behaviors tapped by these items are core to externalizing psychopathology in children.

When choosing items, we excluded items if they (a) were not administered to both parents and teachers (teacher-report only: explosive); (b) were not sensitive to clinical-range scores in both samples (Study 1: disobedient at school; Study 2: destroys own things and teases); or (c) did not provide much measurement precision in the subclinical to clinical range of externalizing problems. Reasons for less measurement precision in the subclinical to clinical range included: (a) low discrimination (e.g., talks too much); (b) low severity (e.g., argues); and/or (c) high severity (e.g., sets fires). Low discrimination would suggest that the item was not highly relevant to the construct of externalizing problems (insofar as it was measured). Low severity would suggest that the item was too common or normative to be informative for making diagnostic screening decisions. High severity would suggest that the item was too infrequent and non-normative for it to be informative for a brief screen of externalizing disorders. We also observed developmental changes in items' usefulness. The item on use of alcohol and drugs was more severe and less useful from ages 5–8 than from ages 9–13, whereas the item on vandalism was less severe and more useful from ages 5–8 than from ages 9–13. We also observed some differences in items' usefulness by rater. For example, as might be expected, the item reflecting disobedience at school had higher discrimination and greater usefulness for teachers' than mothers' ratings.

Smith, McCarthy, and Anderson (2000) described seven steps for short-form development: (a) ensure the full form has been validated for the intended purpose; (b) clarify the intended use of the short form and choose items that meet the goal; (c) compute an estimate of the short form's reliability; (d) compute an estimate of the overlap between the short form and the full form; (e) compute an estimate of the validity correlations of the short form with key criteria; (f) compute an estimate of the classification accuracy of the short form; and (g) compute estimates of the time saved and the validity lost.

Regarding Step 1, the CBCL is a widely used, researched, and validated assessment for externalizing behavior problems. Regarding Step 2, we chose items with the goal to develop a screening measure sensitive to clinical-range scores. Regarding Step 3, we demonstrated that scores on the mother- and teacher-reported short scales using the selected items have strong internal consistency, cross-age test–retest reliability, and interrater reliability (and are comparable with estimates from the full Externalizing scale). Cross-age test–retest reliability estimates ranged from $r = .35$ to .75, which is relatively high for behavioral constructs, especially given the expected developmental change in rank order during childhood and the long developmental lag between assessments (12 months). Cross-age test-retest reliability was stronger for mothers' reports than teachers' reports, possibly because different teachers provided ratings from year to year. Regarding Step 4, we showed that scores on the short scales are highly correlated with scores on the full Externalizing scale, and therefore have strong overlap.

Regarding Step 5, we showed that scores on the mother-reported short scale at ages 16 and 17 are associated with the key criterion of a research diagnosis of conduct disorder at age 18 (Study 1), and that scores on the teacher-reported short scale are associated with a research diagnosis of conduct disorder at age 11 (Study 2). Regarding Step 6, we demonstrated that scores on the mother-reported short scale at ages 16 and 17 performed as well as scores on the full Externalizing scale in predicting later DIS conduct disorder diagnoses at age 18, providing evidence of predictive validity. Moreover, scores on the teacher-reported short scale had moderate classification accuracy for DISC conduct disorder diagnoses at age 11, based on both parent and child interview, providing evidence of criterion validity. Scores on the teacher-reported short scale performed as well as scores on the full scale in classifying DISC conduct disorder diagnoses at age 11 based on parent interview, and performed *better* than scores on the full scale in classifying DISC conduct disorder diagnoses based on child interview.

Regarding Step 7, based on Achenbach and Rescorla's (2001) estimate that the full CBCL of 113 items typically takes about 10 min to complete, we estimate that the 10 items would typically take less than 1 min to complete. Thus, the scale met the criteria outlined by Smith et al. (2000) for short-form development, using two independent samples.

In addition to being quicker to administer, a shorter form has other benefits, including less missingness (missing responses to some items invalidate the sum scores; Lambert et al., 2003), less patient and participant burden, and potentially more measurement precision (Embretson & Reise, 2000). Smith et al. (2000) described IRT as a better approach to short-form development than classical test theory approaches because IRT can be used to develop shorter assessments that retain the measurement information for the trait levels of interest (as was done in the present study). Moreover, one can calculate more accurate estimates of externalizing problems from IRT factor scores based on response patterns from the selected items compared with unweighted item sums (Cole et al., 2011; Dumenci & Achenbach, 2008; Lindhiem, Bennett, Hipwell, & Pardini, 2015).

In terms of loss of validity for identifying clinical-range scores, we demonstrated that scores on the short scale were as accurate as, if not more accurate than, scores on the full Externalizing scale for classifying and predicting conduct disorder diagnoses based on a research diagnostic interview. Nevertheless, choosing whether to administer the select set of items or the full Externalizing scale depends on the purpose of the assessment. If one's goal is to understand normative variation in externalizing behavior across a wide range of severity, the full Externalizing scale is likely better than the select set of 10 items. On the other hand, if the goal is to efficiently screen for clinical-range externalizing behavior, the select set of items may be as useful as, and more efficient than, the full Externalizing scale for externalizing disorders. However, other sources of information in addition to questionnaires should be considered when making clinical diagnostic decisions, and comprehensive assessment would consider other domains in addition to externalizing problems.

Including additional assessment information is especially important given that the selected items had moderate accuracy for classifying conduct disorder. In particular, the 10 items had a low positive predictive value, indicating that many children who exceeded a given threshold did not meet criteria for conduct disorder (false positives). As a result, the 10 items might serve as an initial screen to indicate who should undergo further assessment for externalizing disorders. Sensitivity was higher at low cut points when the predictor and criterion were assessed in the same year (Study 2) compared with when the criterion was assessed a year or two later (Study 1). However, specificity was relatively low at low cut points. The optimal cutpoint for a measure depends on the goal (Treat & Viken, 2012). If the goal is to identify most at-risk children with a screening device, a lower cutpoint (e.g., 1 or 2) might be optimal. On the other hand, if the goal is to minimize false positives, a higher cutpoint might be better (e.g., 6 or 7). The moderate accuracy of the short scale for predicting conduct disorder is unsurprising because the CBCL is not a diagnostic checklist. Because research continues to demonstrate the dimensional nature of externalizing problems (Coghill & Sonuga-Barke, 2012; Krueger, Markon, Patrick, & Iacono, 2005; Markon & Krueger, 2005; Walton, Ormel, & Krueger, 2011), the field is moving from categorical approaches to more dimensional approaches to the assessment of externalizing psychopathology. The CBCL is a widely used dimensional approach to the assessment of externalizing problems, and more efficient sets of CBCL items might play an important role in increasing the practicality of its use in more research and practical contexts. Despite having moderate accuracy for classifying and predicting conduct disorder, the short scale may still be useful as an efficient screen for children with clinical-range externalizing problems who merit further attention.

Labeling and misidentification are potential concerns in the identification of children at risk for developing behavior problems. Some children may be deemed at risk but never develop problems (false positive). Giving these children an "at risk" label could be harmful if they receive, without a full assessment, a treatment with potentially serious adverse effects (e.g., medication). Nevertheless, some of the most effective treatments for externalizing problems include behavioral interventions such as parent training (West, 2013) that could still be beneficial for the child's development even in the case of a false positive. At the same time, sensitivity—detecting who is at risk of developing problems—is a bigger problem today than specificity because so many people go without important and effective services until it may be too late for cost-effective interventions (Insel, 2014). Ultimately, we view the short scale as a potential screen to indicate whether follow-up assessment is necessary, rather than as a diagnostic tool itself.

The present study had several strengths. First, we examined ratings of children's behavior problems from two different reporters: mothers and teachers. Having two raters allowed us to compare the useful items from different perspectives and contexts. Second, we collected mothers' and teachers' reports longitudinally from ages 5–13. Longitudinal reports of behavior problems allowed us to see the general stability of item properties and also the developmental changes over time in item functioning. From the item parameters across years and raters that we present in the tables, researchers could develop scales that target other ranges of severity in ways that are developmentally sensitive. Third, we validated the items selected for the short scale against DIS conduct disorder diagnoses. Fourth, we essentially replicated our findings in two independent samples, providing stronger evidence in support of the 10 useful items we identified.

The present study also had limitations. First, the IRT models did not meet strict assumptions of unidimensionality and local independence for both raters at all years, suggesting that the CBCL captures multiple subdimensions of externalizing problems. The subset of 10 items could not assess all of the aspects of externalizing problems that would be relevant in a clinical context. For example, "sets fires" could be an essential item to a clinician even though it is not included in the 10 items we identified. Other IRT studies of the CBCL also demonstrated violations of unidimensionality and/or local independence (Cheong & Raudenbush, 2000; Lambert et al., 2003). Nevertheless, IRT parameter estimates are robust to minor and moderate violations of unidimensionality (Harrison, 1986) and local independence (Fennessy, 1995). In any case, the CBCL and TRF are two of the most widely used assessments for externalizing problems, both in clinical and research contexts, so it is informative to know their test and item properties for clinical utility (especially because the Externalizing scale is often used as if it were unidimensional). Thus, although we feel the item properties are informative, we present them with caution. Future studies might be able to estimate more accurate parameter estimates of the Achenbach Externalizing scale using multidimensional IRT approaches (Reckase, 2009). The general stability of item properties across years, raters, and samples provides evidence for the utility of the identified items (but there were some differences, as described earlier). Moreover, many of the useful items are consistent with findings from prior studies (possible reasons for differences in the items we identified were described earlier). The ultimate question, however, is not whether the IRT assumptions are fully met and the IRT parameters are accurate, but rather how well the selected items perform in classifying externalizing psychopathology. In both studies, scores on the selected items performed as well as, if not better than, scores on the full Externalizing scale in classifying or predicting DIS conduct disorder diagnoses, suggesting that the IRT item properties were meaningful.

Second, although the study included a community sample (Study 1) and a high-risk sample (Study 2), it remains to be seen whether the findings would generalize to a clinical sample. Other steps will be necessary to validate the items identified here (Smith McCarthy, & Anderson, 2000). The selected items should be replicated with independent samples and, ideally, as a standalone measure because item responses can be influenced by the presence of other items. Nevertheless, IRT analyses provide theoretically unbiased measurement estimates on a common metric that are independent of the sample and items (Lambert et al., 2003). In other words, IRT item parameters closely approximate the true item parameters even if the sample is unrepresentative (Embretson, 1996), assuming similar symptom severity target range (when calculating information estimates) and developmental stage, providing further confidence in our results. We view the present study as an important step toward developing a clinically useful and efficient screen for externalizing disorders. Future studies should attempt to replicate and extend these findings longitudinally with multiple raters and larger, independent samples, and should establish population norms for a scale using these items. Of particular interest may be replication with clinical samples and samples with a different ethnic or cultural composition. Another complex but potentially important future approach may be using longitudinal IRT and tests of differential item functioning to examine whether items change in discrimination, severity, or usefulness across time. Such an approach would help ensure that item properties can be compared on the same scale across years and raters. Future studies might also identify the most useful items for assessing particular subdimensions of externalizing problems (e.g., physical aggression).

Future studies might also examine other ways to increase the efficiency of assessment. One potential way to increase efficiency may be through the use of computerized adaptive testing (CAT). Unlike fixed forms, CAT tailors the items administered to each person by updating an estimate of the person's trait level based on whether an item is endorsed and choosing which item to administer next (i.e., the item that would be most informative for updating the trait level estimate). This iterative process continues until confidence about the person's trait level or diagnostic status reaches a certain threshold. CAT could allow fewer items to be administered, and these individually tailored items might be just as informative as the full fixed form. The Patient-Reported Outcomes Measurement Information System (PROMIS) is an initiative of the National Institutes of Health that uses CAT based on IRT analysis for a wide range of medical problems, including depression and anxiety (Reeve et al., 2007).

In conclusion, the present study identified 10 items that were sensitive to clinical-range scores of externalizing problems and were predictive of conduct disorder diagnoses. These 10 items appear to be core to externalizing psychopathology. The items were generally informative for clinical levels of externalizing problems across two raters (mothers and teachers), developmental eras (ages 5–8 and 9–13), and independent samples. Moreover, scores on the select set of items showed criterion and predictive validity, and may be as informative as, and more efficient than, scores on the full Externalizing scale for screening externalizing disorders.

References

Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Teacher's Report Form and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., McConaughy, S. H., Ivanova, M. Y., & Rescorla, L. A. (2011). *Manual for the ASEBA Brief Problem Monitor™(BPM)*. Burlington, VT: ASEBA.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Akobeng, A. K. (2007a). Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. *Acta Paediatrica, 96,* 338–341. http://dx.doi.org/10.1111/j.1651-2227.2006.00180.x

Akobeng, A. K. (2007b). Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatrica, 96,* 644–647. http://dx.doi.org/10.1111/j.1651-2227.2006.00178.x

Bérubé, R. L., & Achenbach, T. M. (2014). *Bibliography of published studies using the ASEBA*. Retrieved from www.aseba.org

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods, 5,* 477–495. http://dx.doi.org/10.1037/1082-989X.5.4.477

Chorpita, B. F., Reise, S., Weisz, J. R., Grubbs, K., Becker, K. D., Krull, J. L., & the Research Network on Youth Mental Health. (2010). Evaluation of the Brief Problem Checklist: Child and caregiver interviews to measure clinical progress. *Journal of Consulting and Clinical Psychology, 78,* 526–536. http://dx.doi.org/10.1037/a0019602

Coghill, D., & Sonuga-Barke, J. S. (2012). Annual research review: Categories versus dimensions in the classification and conceptualisation of child and adolescent mental disorders: Implications of recent empirical study. *Journal of Child Psychology and Psychiatry, 53,* 469–489. http://dx.doi.org/10.1111/j.1469-7610.2011.02511.x

Cole, D. A., Cai, L., Martin, N. C., Findling, R. L., Youngstrom, E. A., Garber, J., . . . Forehand, R. (2011). Structure and measurement of depression in youths: Applying item response theory to clinical data. *Psychological Assessment, 23,* 819–833. http://dx.doi.org/10.1037/a0023518

Compton, W. M., & Cottler, L. B. (2004). The Diagnostic Interview Schedule (DIS). In M. Hersen (Ed.), *Comprehensive handbook of psychological assessment: Vol. 2. Personality assessment* (pp. 153–162). Hoboken, NJ: Wiley.

Conduct Problems Prevention Research Group. (1992). A developmental and clinical model for the prevention of conduct disorder: The FAST Track Program. *Development and Psychopathology, 4,* 509–527. http://dx.doi.org/10.1017/S0954579400004855

Costello, E. J., Egger, H., & Angold, A. (2005). 10-year research update review: The epidemiology of child and adolescent psychiatric disorders: I. Methods and public health burden. *Journal of the American Academy of Child and Adolescent Psychiatry, 44,* 972–986. http://dx.doi.org/10.1097/01.chi.0000172552.41596.6f

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44,* 837–845. http://dx.doi.org/10.2307/2531595

Dodd, B. G., De Ayala, R., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19,* 5–22. http://dx.doi.org/10.1177/014662169501900103

Dodge, K. A., Bates, J. E., & Pettit, G. S. (1990). Mechanisms in the cycle of violence. *Science, 250,* 1678–1683. http://dx.doi.org/10.1126/science.2270481

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment, 20,* 55–62. http://dx.doi.org/10.1037/1040-3590.20.1.55

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341–349. http://dx.doi.org/10.1037/1040-3590.8.4.341

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (Vol. 4). Mahwah, NJ: Erlbaum.

Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement, 63,* 915–930. http://dx.doi.org/10.1177/0013164403251319

Fantoni-Salvador, P., & Rogers, R. (1997). Spanish versions of the MMPI-2 and PAI: An investigation of concurrent validity with Hispanic patients. *Assessment, 4,* 29–39.

Fennessy, L. M. (1995). *The impact of local dependencies on various IRT outcomes* (Doctoral dissertation). Available from ProQuest Dissertations & Theses database. (UMI No. 9524701)

Frick, P. J., Cornell, A. H., Bodin, S. D., Dane, H. E., Barry, C. T., & Loney, B. R. (2003). Callous-unemotional traits and developmental pathways to severe conduct problems. *Developmental Psychology, 39,* 246–260. http://dx.doi.org/10.1037/0012-1649.39.2.246

Gomez, R. (2008). Item response theory analyses of the parent and teacher ratings of the *DSM–IV* ADHD rating scale. *Journal of Abnormal Child Psychology, 36,* 865–885. http://dx.doi.org/10.1007/s10802-008-9218-8

Gumpel, T., Wilson, M., & Shalev, R. (1998). An item response theory analysis of the Conners Teacher's Rating Scale. *Journal of Learning Disabilities, 31,* 525–532. http://dx.doi.org/10.1177/002221949803100602

Harford, T. C., Chen, C. M., Saha, T. D., Smith, S. M., Hasin, D. S., & Grant, B. F. (2013). An item response theory analysis of *DSM–IV* diagnostic criteria for personality disorders: Findings from the national epidemiologic survey on alcohol and related conditions. *Personality Disorders, 4,* 43–54. http://dx.doi.org/10.1037/a0027416

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11,* 91–115. http://dx.doi.org/10.3102/10769986011002091

Hodges, J. L., Jr., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics, 34,* 598–611. http://dx.doi.org/10.1214/aoms/1177704172

Insel, T. R. (2014). Mental disorders in childhood: Shifting the focus from behavioral symptoms to neurodevelopmental trajectories. *Journal of the American Medical Association, 311,* 1727–1728. http://dx.doi.org/10.1001/jama.2014.1193

Jensen, P., Roper, M., Fisher, P., Piacentini, J., Canino, G., Richters, J., . . . Schwab-Stone, M. (1995). Test-retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1). Parent, child, and combined algorithms. *Archives of General Psychiatry, 52,* 61–71. http://dx.doi.org/10.1001/archpsyc.1995.03950130061007

Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacono, W. G. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology, 114,* 537–550. http://dx.doi.org/10.1037/0021-843x.114.4.537

Krueger, R. F., Nichol, P. E., Hicks, B. M., Markon, K. E., Patrick, C. J., Iacono, W. G., & McGue, M. (2004). Using latent trait modeling to conceptualize an alcohol problems continuum. *Psychological Assessment, 16,* 107–119.

Lambert, M. C., Schmitt, N., Samms-Vaughan, M. E., An, J. S., Fairclough, M., & Nutter, C. A. (2003). Is it prudent to administer all items for each Child Behavior Checklist cross-informant syndrome? Evaluating the psychometric properties of the Youth Self-Report dimensions with confirmatory factor analysis and item response theory. *Psychological Assessment, 15,* 550–568. http://dx.doi.org/10.1037/1040-3590.15.4.550

Lindhiem, O., Bennett, C. B., Hipwell, A. E., & Pardini, D. A. (2015). Beyond symptom counts for diagnosing oppositional defiant disorder and conduct disorder? *Journal of Abnormal Child Psychology*. Advance online publication. http://dx.doi.org/10.1007/s10802-015-0007-x

Lochman, J. E., & The Conduct Problems Prevention Research Group. (1995). Screening of child behavior problems for prevention programs at school entry. *Journal of Consulting and Clinical Psychology, 63,* 549–559. http://dx.doi.org/10.1037/0022-006X.63.4.549

Markon, K. E., & Krueger, R. F. (2005). Categorical and continuous models of liability to externalizing disorders: A direct comparison in NESARC. *Archives of General Psychiatry, 62,* 1352–1359. http://dx.doi.org/10.1001/archpsyc.62.12.1352

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176. http://dx.doi.org/10.1177/014662169201600206

Muthén, L. K., & Muthén, B. O. (2010). *Mplus Version 6.1*. Los Angeles, CA: Muthén & Muthén.

Patrick, C. J., Kramer, M. D., Krueger, R. F., & Markon, K. E. (2013). Optimizing efficiency of psychopathology assessment through quantitative modeling: Development of a brief form of the Externalizing Spectrum Inventory. *Psychological Assessment, 25,* 1332–1348. http://dx.doi.org/10.1037/a0034864

Petersen, I. T., Bates, J. E., Dodge, K. A., Lansford, J. E., & Pettit, G. S. (2015). Describing and predicting developmental profiles of externalizing problems from childhood to adulthood. *Development and Psychopathology, 27,* 791–818. http://dx.doi.org/10.1017/S0954579414000789

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavior problems in children. *Journal of Marriage and the Family, 48,* 295–307. http://dx.doi.org/10.2307/352397

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230. http://dx.doi.org/10.2307/1164671

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-89976-3

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45,* S22–S31. http://dx.doi.org/10.1097/01.mlr.0000250483.85507.04

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17,* 1–25.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12,* 77. http://dx.doi.org/10.1186/1471-2105-12-77

Robins, L. N., Cottler, L. B., Bucholz, K. K., Compton, W. M., North, C. S., & Rourke, K. M. (1999). *Diagnostic Interview Schedule for DSM–IV (DIS-IV)*. St. Louis, MO: Washington University School of Medicine.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. *Archives of General Psychiatry, 38,* 381–389. http://dx.doi.org/10.1001/archpsyc.1981.01780290015001

Robins, L. N., Helzer, J. E., Ratcliff, K. S., & Seyfried, W. (1982). Validity of the diagnostic interview schedule, version II: *DSM–III* diagnoses. *Psychological Medicine, 12,* 855–870. http://dx.doi.org/10.1017/S0033291700049151

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88,* 1273–1283. http://dx.doi.org/10.1080/01621459.1993.10476408

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 34,* 100.

Sattler, J. M., & Hoge, R. D. (2006). *Assessment of children: Behavioral, social, and clinical foundations* (5th ed.). San Diego, CA: Jerome M. Sattler, Publisher, Inc.

Schwab-Stone, M. E., Shaffer, D., Dulcan, M. K., Jensen, P. S., Fisher, P., Bird, H. R., . . . Rae, D. S. (1996). Criterion validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3). *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 878–888. http://dx.doi.org/10.1097/00004583-199607000-00013

Shaffer, D., & Fisher, P. (1997). *NIMH—Diagnostic Interview Schedule for Children: Parent informant*. New York, NY: New York State Psychiatric Institute.

Shaffer, D., Fisher, P., Dulcan, M. K., Davies, M., Piacentini, J., Schwab-Stone, M. E., . . . Regier, D. A. (1996). The NIMH Diagnostic Interview Schedule for Children version 2.3 (DISC-2.3): Description, acceptability, prevalence rates, and performance in the MECA study. *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 865–877. http://dx.doi.org/10.1097/00004583-199607000-00012

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 28–38. http://dx.doi.org/10.1097/00004583-200001000-00014

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12,* 102–111. http://dx.doi.org/10.1037/1040-3590.12.1.102

Studts, C. R., & van Zyl, M. A. (2013). Identification of developmentally appropriate screening items for disruptive behavior problems in preschoolers. *Journal of Abnormal Child Psychology, 41,* 851–863. http://dx.doi.org/10.1007/s10802-013-9738-8

Treat, T. A., & Viken, R. J. (2012). Measuring test performance with signal detection theory techniques. In H. Cooper (Ed.), *Handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (Vol. 1, pp. 723–744). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/13619-038

Wakschlag, L. S., Briggs-Gowan, M. J., Choi, S. W., Nichols, S. R., Kestler, J., Burns, J. L., . . . Henry, D. (2014). Advancing a multidimensional, developmental spectrum approach to preschool disruptive behavior. *Journal of the American Academy of Child and Adolescent Psychiatry, 53,* 82–96.e3. http://dx.doi.org/10.1016/j.jaac.2013.10.011

Walton, K. E., Ormel, J., & Krueger, R. F. (2011). The dimensional nature of externalizing behaviors in adolescence: Evidence from a direct comparison of categorical, dimensional, and hybrid models. *Journal of Abnormal Child Psychology, 39,* 553–561. http://dx.doi.org/10.1007/s10802-010-9478-y

West, A. E. (2013). Review: Psychosocial interventions improve early disruptive behaviour in very young children. *Evidence-Based Mental Health, 16,* 70. http://dx.doi.org/10.1136/eb-2013-101252

Zill, N. (1990). *Behavior Problems Index based on parent report*. Unpublished measure. Washington, DC: Child Trends.

Received December 9, 2014
Revision received May 19, 2015
Accepted May 20, 2015 ■

Table S1. Item discrimination parameters by age and rater (Study 1).

				Age 5		Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12		Age 13	
Number	Item	CBCL	TRF	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T
3	Argues	x	x	1.39	2.41	1.70	2.82	1.57	3.11	2.42	2.95	1.80	2.88	2.40	2.48	1.89	3.63	1.65	2.46	1.96	3.69
6	Defiant		x		2.67		2.75		2.91		2.38		3.31		3.19		3.42		2.67		3.55
7	Brags	x	x	0.88	1.73	0.98	1.66	1.00	1.43	1.08	1.46	1.23	1.76	1.59	1.85	1.47	1.99	1.09	1.55	1.47	2.24
16	Mean to others	x	x	1.71	2.82	1.54	3.14	1.66	2.64	1.85	2.91	2.06	3.35	1.92	2.87	1.98	3.62	1.75	2.21	2.34	3.26
19	Demands attention	x	x	1.36	2.08	1.53	2.13	1.26	1.97	1.71	1.92	1.39	2.18	1.45	1.98	1.50	2.38	1.59	1.54	1.67	3.18
20	Destroys own things	x	x	1.28	2.45	1.31	1.89	1.72	1.42	1.99	1.65	2.13	2.11	1.57	2.61	1.45	2.54	1.46	1.38	2.33	1.77
21	Destroys others' things	x	x	1.24	2.96	1.53	2.37	1.88	2.20	2.14	2.91	2.38	2.59	2.04	2.68	1.78	3.00	1.54	1.93	3.11	2.19
22	Disobedient at home	x		1.68		1.89		1.56		2.13		1.87		2.01		2.34		1.62		2.01	
23	Disobedient at school	x	x	0.83	3.63	1.26	3.23	1.28	3.25	2.01	3.67	1.79	3.85	1.30	3.67	1.88	3.76	1.66	3.39	1.92	3.76
24	Disturbs others		x		3.65		3.17		2.77		3.11		3.01		3.42		3.73		3.25		4.19
26	Lacks guilt	x	x	1.08	2.40	1.51	2.23	1.46	2.35	2.09	2.96	1.45	3.43	1.39	2.27	1.53	3.32	1.20	2.35	1.41	3.03
27	Jealous	x	x	1.32	1.56	1.44	1.99	1.05	1.73	1.64	1.25	1.61	1.72	1.25	1.36	1.38	1.82	1.56	1.54	1.53	2.01
37	Fights	x	x	1.29	3.14	1.59	2.42	1.39	2.49	1.33	2.74	1.68	2.71	1.85	2.13	1.78	2.75	1.59	2.09	1.65	2.21
39	Bad companions	x	x	1.23	1.98	1.34	2.01	1.19	1.54	1.12	1.58	1.85	2.15	1.30	1.58	1.32	1.94	1.35	1.44	1.39	2.13
43	Lies (and cheats)	x	x	1.42	2.37	1.43	1.82	1.71	2.17	2.26	1.87	2.39	2.44	1.78	1.57	1.96	2.54	1.99	1.52	2.27	2.70
53	Talks out of turn		x		2.41		2.61		2.36		2.13		2.39		2.18		2.91		2.29		3.54
57	Attacks people	x	x	1.84	2.93	1.64	2.06	1.54	1.92	1.68	2.74	1.77	2.54	1.51	3.01	1.83	3.10	1.83	3.14	2.13	2.19
63	Prefers older kids	x	x	0.51	0.68	0.73	0.42	0.64	0.41	0.57	0.68	0.73	0.69	0.77	0.64	0.85	0.85	0.84	0.81	0.72	0.59
67	Runs away	x		1.06		2.26		0.74		0.82		1.36		1.23		1.09		1.96		1.38	
67	Disrupts class		x		3.38		3.38		3.20		3.14		3.73		3.30		4.00		2.73		4.14
68	Screams	x	x	1.73	2.56	1.31	1.75	1.63	3.65	1.87	2.37	1.54	2.37	2.04	2.89	1.88	3.14	1.43	2.35	1.92	2.08
72	Sets fires	x		1.33		1.47		0.96		0.54		1.56		1.06		1.62		1.31		2.00	
72	Messy work		x		0.90		0.96		0.75		0.88		1.13		0.89		1.01		0.91		1.08
74	Shows off	x	x	1.34	2.38	1.57	2.33	1.52	2.09	1.96	2.01	1.62	1.95	1.83	2.11	1.27	3.02	1.32	2.01	1.60	2.95
76	Explosive		x		2.60		2.98		1.96		3.26		2.69		2.98		3.55		2.81		3.06
77	Easily frustrated		x		1.86		2.14		1.86		1.64		2.18		2.12		2.62		1.90		3.09
81	Steals at home	x		0.98		1.38		1.16		1.56		1.77		1.86		1.62		2.00		2.01	
82	Steals outside home	x	x	0.72	2.46	2.65	1.71	1.21	1.95	1.34	2.00	1.74	1.93	1.53	1.68	0.91	2.26	2.60	1.30	2.20	1.29
86	Stubborn, irritable	x	x	1.66	1.89	1.80	1.93	1.52	2.33	1.95	2.03	1.62	2.53	2.03	2.03	2.20	2.59	1.78	1.81	1.74	2.11
87	Sudden mood changes	x	x	1.64	1.42	1.57	1.65	1.77	1.79	1.50	1.92	1.14	2.18	1.80	1.79	1.90	2.66	1.17	1.85	1.69	1.85
90	Swearing, obscenity	x	x	1.48	2.51	1.26	1.94	1.91	2.19	1.56	2.12	1.78	2.59	2.04	1.90	1.66	3.49	2.03	2.40	1.31	2.32
93	Talks too much	x	x	1.19	1.99	1.20	2.14	1.30	2.02	1.45	1.59	1.15	2.00	0.99	1.89	0.91	2.92	0.89	2.12	1.26	3.11
94	Teases	x	x	1.31	2.30	1.34	2.70	1.34	1.94	1.83	2.13	1.53	2.60	1.52	1.99	1.30	3.27	1.42	1.96	1.80	2.52
95	Temper tantrums	x	x	1.85	3.04	2.11	2.70	1.85	2.18	2.26	2.48	2.00	2.60	2.32	2.69	2.38	3.24	2.31	3.44	2.39	2.59
96	Thinks about sex too much	x		0.83		1.37		1.20		1.36		1.23		1.48		1.46		1.41		1.63	
97	Threatens others	x	x	2.50	3.38	1.89	2.82	2.34	2.68	2.03	2.82	2.35	2.93	4.01	2.75	2.83	3.33	2.16	2.83	2.23	2.52
98	Tardy		x		0.51		0.47		0.39		0.45		0.31		0.46		1.12		1.19		1.23
101	Truancy	x	x	a	0.25	1.74	0.40	1.93	0.90	0.70	0.62	1.77	0.62	2.10	0.77	1.77	0.54	1.10	1.04	1.55	0.64
104	Loud	x	x	1.84	2.60	1.43	2.57	1.68	2.22	2.15	2.11	1.46	2.39	1.99	2.03	1.69	3.46	1.72	2.56	2.14	3.09
105	Alcohol, drugs	x	x	0.87	a	a	1.37	16.12	a	a	-0.12	a	1.50	a	a	1.29	2.32	17.63	1.17	1.30	0.57
106	Vandalism	x		3.28		3.30		2.13		3.34		2.23		1.18		2.31		3.30		a	

a = no variability or not enough variability for convergence
M = mother-reported
T = teacher-reported

Table S2. Item severity (b1) parameters by age and rater (Study 1).

				Age 5		Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12		Age 13	
Number	Item	CBCL	TRF	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T
3	Argues	x	x	-1.44	0.71	-1.28	0.91	-1.16	0.80	-0.90	0.71	-1.05	0.45	-1.13	0.61	-0.93	0.15	-0.88	0.81	-0.95	0.19
6	Defiant		x		1.43		1.45		1.40		1.47		0.90		0.97		0.54		1.20		0.35
7	Brags	x	x	-0.39	1.39	-0.23	1.62	-0.21	1.39	0.12	1.41	-0.15	1.14	-0.15	1.25	-0.07	0.74	-0.08	1.75	-0.15	0.71
16	Mean to others	x	x	0.98	1.30	1.15	1.55	1.28	1.39	1.62	1.42	1.11	0.93	1.11	1.18	1.17	0.71	1.29	1.74	1.09	0.70
19	Demands attention	x	x	-0.65	0.74	-0.39	0.80	-0.07	0.83	0.22	0.86	0.06	0.73	-0.07	0.96	0.11	0.51	0.05	1.18	-0.01	0.27
20	Destroys own things	x	x	1.01	2.18	1.32	2.25	1.36	2.63	1.58	2.45	1.54	2.28	1.68	2.29	1.96	1.80	2.27	4.24	1.62	1.91
21	Destroys others' things	x	x	1.38	1.82	1.51	2.05	1.46	2.17	1.68	2.28	1.53	1.76	1.59	2.09	1.88	1.57	2.30	2.92	1.48	1.55
22	Disobedient at home	x		-0.75		-0.45		-0.42		0.01		-0.15		-0.28		-0.07		-0.02		-0.35	
23	Disobedient at school	x	x	1.66	0.82	1.12	1.03	1.02	0.93	1.18	1.01	0.92	0.76	1.19	0.89	0.98	0.49	1.06	1.09	0.59	0.29
24	Disturbs others		x		0.56		0.46		0.47		0.57		0.40		0.44		0.19		0.67		0.07
26	Lacks guilt	x	x	0.39	1.10	0.68	1.15	1.01	1.20	0.96	1.20	0.93	0.72	0.64	1.06	0.91	0.59	1.10	1.38	0.52	0.36
27	Jealous	x	x	-0.24	1.86	-0.13	1.55	0.25	1.77	0.51	1.86	0.22	1.42	0.24	1.58	0.18	1.11	0.33	2.18	0.25	1.17
37	Fights	x	x	1.63	1.25	1.62	1.44	2.04	1.61	2.13	1.55	1.70	1.16	1.24	1.45	1.82	1.04	2.07	2.21	1.74	1.18
39	Bad companions	x	x	1.98	1.38	1.81	1.42	1.86	1.50	2.09	1.46	1.36	1.02	1.45	1.17	1.63	0.69	1.56	1.15	1.19	0.28
43	Lies (and cheats)	x	x	0.54	1.48	0.72	1.68	0.68	1.53	1.02	1.70	0.72	1.12	0.57	1.51	0.74	0.91	0.95	2.09	0.77	0.71
53	Talks out of turn		x		0.44		0.48		0.41		0.63		0.48		0.46		0.08		0.60		0.01
57	Attacks people	x	x	1.39	1.46	1.88	1.62	2.07	1.97	2.12	1.73	2.01	1.49	2.02	1.63	1.94	1.12	2.19	2.33	1.90	1.29
63	Prefers older kids	x	x	-1.11	2.88	-0.63	5.09	-0.22	4.59	-0.23	3.88	0.13	3.53	0.42	3.82	0.56	2.35	0.75	3.08	0.41	3.24
67	Runs away	x		3.99		2.81		6.18		5.75		4.11		4.31		4.30		3.56		3.36	
67	Disrupts class		x		0.88		0.84		0.92		1.02		0.70		0.71		0.40		0.86		0.12
68	Screams	x	x	0.91	1.97	1.38	2.49	1.21	2.33	1.48	2.85	1.24	2.16	0.78	2.40	1.27	1.52	1.41	2.63	1.18	1.69
72	Sets fires	x		3.87		3.88		5.43		10.03		4.23		4.25		3.10		3.85		3.05	
72	Messy work		x		1.07		0.63		1.12		1.16		0.71		0.91		0.65		1.14		0.81
74	Shows off	x	x	-0.85	1.03	-0.64	1.05	-0.42	0.98	0.15	1.14	-0.19	0.93	-0.23	0.91	-0.01	0.37	-0.19	1.11	-0.11	0.28
76	Explosive		x		1.70		1.60		1.99		1.84		1.50		1.65		1.11		1.96		0.82
77	Easily frustrated		x		1.44		1.41		1.58		1.65		1.24		1.36		0.61		1.58		0.49
81	Steals at home	x		3.71		2.76		3.27		2.67		2.37		2.46		2.53		2.34		2.25	
82	Steals outside home	x	x	5.32	2.21	2.19	2.57	3.34	2.44	3.15	2.55	2.79	2.11	3.23	2.65	4.56	2.07	2.45	3.74	2.43	2.85
86	Stubborn, irritable	x	x	-0.44	1.16	-0.27	1.35	-0.13	1.36	0.03	1.30	-0.02	0.96	-0.27	0.95	-0.12	0.54	-0.42	1.50	-0.36	0.56
87	Sudden mood changes	x	x	0.83	2.10	0.96	1.75	1.11	1.83	1.41	1.88	1.12	1.36	0.61	1.47	0.64	0.90	0.63	1.71	0.35	0.96
90	Swearing, obscenity	x	x	1.66	2.03	2.29	2.68	1.77	2.44	2.30	2.51	1.74	1.70	1.57	2.15	1.81	1.34	1.56	2.28	1.27	1.20
93	Talks too much	x	x	-0.40	0.61	-0.19	0.48	-0.10	0.41	0.37	0.68	0.14	0.46	0.24	0.48	0.30	0.19	0.32	0.66	0.07	0.07
94	Teases	x	x	0.73	1.70	0.75	1.66	0.90	1.71	0.97	1.53	0.87	1.07	0.48	1.47	0.75	0.62	0.46	1.39	0.37	0.68
95	Temper tantrums	x	x	0.05	1.67	0.33	1.69	0.47	2.01	0.74	1.76	0.59	1.39	0.12	1.48	0.47	0.90	0.63	1.79	0.49	0.91
96	Thinks about sex too much	x		4.52		3.00		3.50		3.21		3.37		2.59		3.08		3.14		2.46	
97	Threatens others	x	x	1.46	1.88	1.91	2.25	1.72	2.08	1.98	2.17	1.85	1.52	1.51	1.79	1.80	1.22	2.11	2.21	1.69	1.16
98	Tardy		x		5.34		5.50		6.18		5.77		7.30		5.22		2.05		2.34		1.25
101	Truancy	x	x	a	15.04	3.47	8.48	3.46	4.39	7.57	5.46	3.67	5.07	2.94	3.71	3.65	4.93	4.52	3.14	3.09	3.16
104	Loud	x	x	0.55	1.26	0.57	1.48	0.75	1.56	1.09	1.85	0.84	1.23	0.61	1.39	0.86	0.77	0.88	1.56	0.61	0.58
105	Alcohol, drugs	x	x	7.66	a	a	4.64	2.77	a	a	-44.53	a	4.61	a	a	4.82	3.11	3.07	6.09	4.07	7.69
106	Vandalism	x		2.65		2.61		3.15		3.16		3.29		4.13		3.42		3.14		a	

a = no variability or not enough variability for convergence
M = mother-reported
T = teacher-reported

Table S3. Item severity (b2) parameters by age and rater (Study 1).

				Age 5		Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12		Age 13	
Number	Item	CBCL	TRF	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T
3	Argues	x	x	1.45	1.86	1.12	2.07	1.28	1.94	1.78	2.15	1.15	1.40	0.92	1.92	1.13	1.08	1.21	2.25	1.20	1.47
6	Defiant		x		2.43		2.49		2.36		2.71		1.75		2.06		1.64		2.47		1.22
7	Brags	x	x	3.12	2.82	2.90	3.01	3.40	2.98	2.67	3.26	2.64	2.43	2.44	2.61	2.42	2.10	3.22	3.22	3.00	1.73
16	Mean to others	x	x	3.26	2.31	3.37	2.63	3.60	2.50	3.76	2.77	2.90	1.90	3.58	2.41	3.15	1.76	3.45	3.19	2.84	1.70
19	Demands attention	x	x	1.53	1.88	1.71	1.91	2.25	2.28	1.92	2.28	2.34	1.68	2.21	2.17	2.44	1.51	2.12	2.51	2.07	0.90
20	Destroys own things	x	x	3.27	3.19	3.76	3.37	3.27	4.41	3.24	4.25	2.94	a	3.64	3.58	a	2.40	4.45	a	3.48	3.06
21	Destroys others' things	x	x	3.77	3.03	4.19	3.36	3.52	3.44	3.45	3.18	3.36	3.50	3.70	3.54	4.29	2.58	4.76	a	a	2.73
22	Disobedient at home	x		2.38		2.13		2.58		1.87		2.61		2.09		2.37		2.88		2.35	
23	Disobedient at school	x	x	5.68	1.86	3.75	2.19	3.95	2.15	3.26	2.16	2.94	1.70	3.73	2.17	3.08	1.53	3.14	2.49	2.61	1.19
24	Disturbs others		x		1.74		1.72		1.86		2.21		1.60		1.83		1.29		2.09		0.93
26	Lacks guilt	x	x	3.09	2.15	2.39	2.19	2.76	2.12	2.27	2.25	3.33	1.57	2.74	2.20	2.74	1.40	3.48	2.34	2.93	1.15
27	Jealous	x	x	1.93	3.51	2.01	2.59	3.03	3.39	2.37	3.74	2.13	3.21	3.01	3.93	2.31	2.58	2.26	3.49	2.55	2.34
37	Fights	x	x	4.19	2.28	3.68	2.51	4.59	2.82	4.21	2.66	3.44	2.31	3.95	3.08	3.31	2.16	3.53	3.44	4.31	2.26
39	Bad companions	x	x	4.55	2.48	3.71	2.36	4.19	3.16	4.92	2.84	2.78	2.22	3.81	2.69	3.99	1.88	3.78	3.03	4.07	1.76
43	Lies (and cheats)	x	x	3.41	2.60	3.43	2.84	2.95	3.02	2.69	2.99	2.55	2.26	3.11	3.43	2.79	2.07	3.14	3.77	2.68	2.04
53	Talks out of turn		x		1.45		1.64		1.69		2.01		1.58		1.77		1.08		2.04		0.78
57	Attacks people	x	x	3.13	2.49	3.30	3.06	4.25	3.29	4.43	3.32	3.91	2.49	4.37	2.89	3.60	1.98	4.34	3.26	3.65	2.30
63	Prefers older kids	x	x	4.20	5.41	3.21	11.33	4.05	9.67	4.97	7.06	3.93	6.54	4.25	7.11	3.49	4.66	3.21	5.74	4.64	6.61
67	Runs away	x		a		a		a		a		a		a		a		a		4.40	
67	Disrupts class		x		1.79		1.79		1.90		2.09		1.57		2.11		1.33		2.25		0.91
68	Screams	x	x	2.74	3.06	3.25	3.56	2.96	3.14	3.44	3.62	3.18	2.96	2.40	4.16	2.64	2.58	3.32	3.28	2.80	2.63
72	Sets fires	x		a		a		a		a		a		a		a		5.37		3.78	
72	Messy work		x		3.14		2.62		3.79		3.06		2.40		2.69		2.56		3.77		3.03
74	Shows off	x	x	1.83	2.17	1.83	2.15	2.13	2.35	2.13	2.83	2.29	2.45	1.75	2.18	2.57	1.30	2.28	2.48	2.19	1.09
76	Explosive		x		2.82		2.60		2.96		2.71		2.44		2.47		1.78		3.00		1.62
77	Easily frustrated		x		2.93		2.52		2.89		2.84		2.32		2.47		1.65		3.11		1.41
81	Steals at home	x		6.88		4.59		4.89		4.21		4.34		a		4.09		3.54		3.74	
82	Steals outside home	x	x	a	2.84	3.40	3.29	5.65	3.13	4.91	3.35	4.37	3.15	4.64	3.73	6.40	2.82	3.36	5.12	a	4.17
86	Stubborn, irritable	x	x	1.72	2.71	1.91	2.61	2.41	2.69	2.09	2.80	2.51	2.36	1.81	2.45	1.96	1.53	1.81	3.05	2.00	2.40
87	Sudden mood changes	x	x	2.88	4.67	2.65	3.02	2.97	3.17	3.34	2.91	3.81	2.43	3.12	2.90	2.65	1.65	3.30	3.40	2.87	2.19
90	Swearing, obscenity	x	x	3.83	2.95	4.89	3.61	3.72	4.01	4.05	3.98	3.64	2.59	2.91	3.83	3.46	2.59	3.11	3.55	4.28	2.37
93	Talks too much	x	x	1.76	1.73	1.88	1.67	2.01	1.91	1.99	2.10	2.34	1.56	2.37	2.04	3.05	1.26	3.21	2.08	2.54	0.79
94	Teases	x	x	3.17	2.87	3.32	3.00	3.61	3.22	3.15	3.17	2.94	2.34	3.16	2.70	3.56	1.56	2.72	2.78	2.44	1.75
95	Temper tantrums	x	x	1.77	2.57	1.99	2.59	2.18	3.06	2.27	2.73	2.07	2.18	1.81	2.58	2.13	1.78	2.07	2.82	1.99	1.80
96	Thinks about sex too much	x		a		4.59		5.71		4.25		5.53		4.05		4.84		4.58		3.46	
97	Threatens others	x	x	3.08	3.01	3.47	3.12	3.34	3.15	3.73	3.40	3.21	2.32	2.33	2.95	2.95	2.29	3.36	3.52	3.23	2.37
98	Tardy		x		7.88		9.00		9.83		8.89		13.52		7.51		2.77		4.01		2.78
101	Truancy	x	x	a	17.91	a	11.86	a	6.11	a	7.65	a	5.73	3.46	5.13	a	6.74	6.16	4.38	4.03	5.29
104	Loud	x	x	2.34	2.14	2.89	2.29	2.47	2.52	2.41	2.70	2.61	2.17	1.94	2.40	2.66	1.67	2.64	2.57	2.18	1.46
105	Alcohol, drugs	x	x	a	a	a	4.94	a	a	a	a	a	a	a	a	a	a	a	a	5.16	9.78
106	Vandalism	x		a		a		3.84		3.52		a		a		a		a		a	

a = no variability or not enough variability for convergence
M = mother-reported
T = teacher-reported

Table S4. Item information by age and rater (Study 1).

				Age 5		Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12		Age 13	
Number	Item	CBCL	TRF	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T
3	Argues	x	x	0.53	1.63	0.51	2.34	0.55	2.48	1.49	2.48	0.56	1.21	0.47	1.74	0.57	0.65	0.55	1.96	0.64	1.73
6	Defiant		x		2.63		2.78		2.97		2.17		2.34		2.80		2.09		2.46		0.95
7	Brags	x	x	0.29	1.22	0.36	1.13	0.34	0.85	0.46	0.84	0.55	1.22	0.88	1.36	0.78	1.35	0.41	0.99	0.66	1.33
16	Mean to others	x	x	0.93	2.74	0.83	3.53	0.88	2.57	1.12	2.84	1.45	2.72	0.94	2.74	1.26	2.62	1.00	1.80	1.76	2.15
19	Demands attention	x	x	0.54	1.32	0.71	1.40	0.59	1.39	0.96	1.34	0.72	1.24	0.76	1.40	0.82	1.15	0.88	0.97	0.95	0.41
20	Destroys own things	x	x	0.63	2.19	0.62	1.35	1.07	0.68	1.44	0.94	1.74	1.39	0.93	2.18	0.70	2.48	0.77	0.18	1.65	1.29
21	Destroys others' things	x	x	0.58	3.08	0.75	2.02	1.14	1.75	1.54	2.83	1.71	2.03	1.27	2.31	1.05	3.32	0.83	0.92	1.48	1.90
22	Disobedient at home	x		0.95		1.16		0.82		1.31		1.10		1.28		1.66		0.81		1.29	
23	Disobedient at school	x	x	0.26	2.90	0.56	3.02	0.52	2.94	1.21	3.53	1.10	2.66	0.60	3.43	1.13	1.99	0.96	3.29	1.25	0.89
24	Disturbs others		x		2.57		2.09		1.95		2.65		1.70		2.55		1.16		2.77		0.35
26	Lacks guilt	x	x	0.45	1.96	0.88	1.77	0.85	1.89	1.56	2.84	0.72	1.94	0.74	1.80	0.91	1.38	0.55	2.07	0.71	0.76
27	Jealous	x	x	0.61	0.96	0.72	1.61	0.43	1.16	0.99	0.63	0.91	1.11	0.56	0.68	0.71	1.29	0.89	0.94	0.84	1.53
37	Fights	x	x	0.61	3.16	0.92	2.24	0.70	2.36	0.67	2.82	1.08	2.47	0.84	1.61	1.25	2.40	1.00	1.60	0.89	1.78
39	Bad companions	x	x	0.57	1.56	0.72	1.59	0.56	0.95	0.48	1.04	1.36	1.64	0.63	1.01	0.65	1.18	0.69	0.82	0.60	1.25
43	Lies (and cheats)	x	x	0.60	2.18	0.64	1.38	0.96	1.74	1.74	1.42	1.81	2.07	0.90	0.93	1.25	2.00	1.16	0.89	1.62	2.12
53	Talks out of turn		x		1.08		1.50		1.37		1.45		1.25		1.30		0.64		1.63		0.25
57	Attacks people	x	x	1.25	3.09	1.09	1.63	0.86	1.44	0.99	2.34	1.13	2.47	0.82	3.14	1.25	2.66	1.14	3.07	1.55	1.80
63	Prefers older kids	x	x	0.10	0.17	0.21	0.05	0.16	0.05	0.12	0.13	0.21	0.15	0.22	0.12	0.29	0.29	0.30	0.22	0.19	0.12
67	Runs away	x		0.20		1.25		0.04		0.05		0.21		0.17		0.16		0.46		0.43	
67	Disrupts class		x		2.49		2.49		2.56		2.81		2.12		2.87		1.34		2.33		0.33
68	Screams	x	x	1.11	2.47	0.70	1.07	1.03	3.96	1.18	1.36	0.90	2.15	1.47	2.31	1.39	3.52	0.80	1.65	1.38	1.78
72	Sets fires	x		0.26		0.27		0.06		0.01		0.18		0.17		0.63		0.27		0.88	
72	Messy work		x		0.35		0.39		0.24		0.34		0.52		0.34		0.42		0.34		0.48
74	Shows off	x	x	0.60	1.92	0.77	1.84	0.80	1.59	1.27	1.44	0.91	1.40	0.96	1.55	0.60	1.05	0.64	1.52	0.89	0.67
76	Explosive		x		2.60		3.31		1.56		3.97		2.71		3.33		2.72		2.92		1.83
77	Easily frustrated		x		1.36		1.81		1.41		1.14		1.77		1.76		1.53		1.39		1.33
81	Steals at home	x		0.23		0.62		0.37		0.80		1.06		1.10		0.90		1.42		1.42	
82	Steals outside home	x	x	0.07	2.26	2.36	1.01	0.37	1.36	0.48	1.34	0.89	1.47	0.54	0.93	0.12	2.02	2.13	0.29	1.46	0.54
86	Stubborn, irritable	x	x	0.81	1.36	1.01	1.49	0.82	2.02	1.23	1.56	0.90	2.13	1.17	1.50	1.43	1.31	0.95	1.28	0.99	1.50
87	Sudden mood changes	x	x	0.97	0.72	0.97	1.13	1.13	1.29	0.87	1.52	0.48	1.84	0.92	1.28	1.23	1.57	0.50	1.27	0.91	1.26
90	Swearing, obscenity	x	x	0.81	2.43	0.58	1.16	1.25	1.52	0.89	1.41	1.14	2.63	1.63	1.29	1.07	3.75	1.54	1.93	0.54	1.96
93	Talks too much	x	x	0.48	1.12	0.51	1.17	0.60	1.25	0.75	0.92	0.52	0.97	0.40	1.20	0.33	0.96	0.32	1.47	0.60	0.31
94	Teases	x	x	0.63	2.09	0.62	2.56	0.59	1.45	1.05	1.61	0.86	2.28	0.72	1.59	0.55	1.78	0.74	1.52	1.13	1.60
95	Temper tantrums	x	x	1.00	3.47	1.38	2.84	1.19	1.87	1.71	2.45	1.34	2.35	1.44	2.71	1.75	2.35	1.67	4.20	1.68	1.76
96	Thinks about sex too much	x		0.12		0.54		0.33		0.48		0.37		0.76		0.57		0.52		0.97	
97	Threatens others	x	x	2.02	3.85	1.34	2.76	1.86	2.60	1.45	2.59	2.02	3.07	5.03	2.81	2.94	3.40	1.73	2.52	1.78	2.23
98	Tardy		x		0.06		0.05		0.03		0.04		0.02		0.05		0.51		0.55		0.65
101	Truancy	x	x	a	0.00	0.48	0.02	0.52	0.14	0.02	0.06	0.38	0.07	1.06	0.17	0.39	0.07	0.14	0.33	0.61	0.15
104	Loud	x	x	1.20	2.27	0.74	2.42	1.06	1.97	1.70	1.83	0.84	1.99	1.26	1.62	1.07	2.25	1.11	2.53	1.52	1.44
105	Alcohol, drugs	x	x	0.01	a	a	0.11	15.71	a	a	0.00	a	0.11	a	a	0.10	0.96	4.10	0.03	0.22	0.02
106	Vandalism	x		2.42		2.50		0.86		1.24		0.73		0.19		0.61		1.26		a	

a = no variability or not enough variability for convergence

M = mother-reported

T = teacher-reported

Table S5. Evaluating the assumptions of IRT: Unidimensonality and local independence (Study 2).

Rater	Age	Cronbach’s Alpha	Unidimensionality	Local Independence
Mother	5	0.89	21.8	0.01
Mother	6	0.90	24.7	0.01
Mother	7	0.91	24.6	0.01
Mother	9	0.92	27.3	0.02
Mother	10	0.92	28.3	0.03
Mother	12	0.93	29.2	0.04
Teacher	5	0.96	42.2	0.11
Teacher	6	0.95	41.1	0.12
Teacher	7	0.95	41.0	0.09
Teacher	8	0.96	44.1	0.10
Teacher	9	0.96	42.8	0.09
Teacher	10	0.96	43.2	0.07
Teacher	11	0.96	43.8	0.11
Teacher	12	0.96	42.9	0.14
Teacher	13	0.96	43.8	0.16

Table S6. Item discrimination parameters by age and rater (Study 2).

				Age 5		Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12		Age 13	
Number	Item	CBCL	TRF	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T
3	Argues	x	x	1.37	2.37	1.50	2.41	1.61	2.29		2.98	1.52	2.68	1.76	3.00		2.72	1.46	2.62		3.06
6	Defiant		x		2.59		2.48		2.51		2.94		2.83		3.13		2.94		3.18		3.25
7	Brags	x	x	0.94	1.44	1.20	1.32	1.34	1.34		1.49	1.21	1.75	1.38	1.78		1.62	1.44	1.45		1.66
16	Mean to others	x	x	1.91	2.72	1.78	2.30	2.29	2.29		2.60	2.05	2.65	2.45	2.56		2.27	2.04	2.11		2.33
19	Demands attention	x	x	1.19	2.06	1.38	1.55	1.42	1.85		1.83	1.56	1.76	1.45	1.98		1.92	1.49	1.92		2.13
20	Destroys own things	x	x	1.59	1.88	1.55	1.68	1.44	1.83		1.70	1.55	1.66	1.48	1.44		1.50	1.72	1.38		1.60
21	Destroys others' things	x	x	1.89	2.17	1.77	2.14	1.84	2.28		2.23	2.12	2.23	1.72	1.88		1.98	2.11	1.91		2.03
22	Disobedient at home	x		1.70		1.85		1.83				2.07		2.65				2.35			
23	Disobedient at school	x	x	1.27	3.23	1.26	3.30	1.48	2.98		3.54	1.49	3.54	1.75	3.71		3.66	2.22	3.41		3.62
24	Disturbs others		x		3.09		2.74		2.62		2.65		2.63		2.82		2.89		2.71		2.76
26	Lacks guilt	x	x	1.08	2.36	0.96	1.77	1.19	2.02		2.15	1.32	2.24	1.36	2.37		2.11	1.39	1.94		2.08
27	Jealous	x	x	1.02	1.23	1.16	1.11	1.08	1.08		1.26	1.31	1.24	1.22	1.31		1.34	1.21	1.24		1.42
37	Fights	x	x	1.70	2.68	1.90	2.12	1.79	2.42		2.35	2.11	2.36	1.89	2.55		2.21	1.81	1.84		2.12
39	Bad companions	x	x	1.30	1.64	1.27	1.55	1.41	1.35		1.72	1.41	1.60	1.30	1.70		1.37	1.52	1.17		1.50
43	Lies (and cheats)	x	x	1.61	1.81	1.95	1.59	1.75	1.73		1.65	1.86	1.82	1.79	1.74		1.84	2.10	1.67		1.86
53	Talks out of turn		x		2.09		1.86		1.71		2.35		1.91		1.94		2.35		2.13		2.21
57	Attacks people	x	x	1.69	2.45	1.98	2.07	1.64	2.45		2.43	2.19	2.25	2.25	2.09		2.00	2.41	2.01		2.12
63	Prefers older kids	x	x	0.87	0.84	0.86	0.72	0.79	0.50		0.87	0.99	0.93	0.82	0.90		0.68	0.99	0.67		0.67
67	Runs away	x		1.27		1.77		1.57				1.41		1.24				1.68			
67	Disrupts class		x		3.07		3.15		2.93		3.45		2.95		3.74		3.45		3.45		3.24
68	Screams	x	x	1.34	2.34	1.48	2.10	1.45	1.79		2.14	1.70	2.06	1.85	1.82		1.96	1.59	2.20		1.80
72	Sets fires	x		0.87		0.69		1.21				1.58		1.34				1.58			
72	Messy work		x		0.82		0.61		0.71		0.70		0.65		0.70		b		b		b
74	Shows off	x	x	1.35	2.03	1.53	1.76	1.49	1.52		1.74	1.61	1.73	1.48	1.76		1.88	1.63	1.79		2.06
76	Explosive		x		2.59		2.44		2.38		2.88		2.95		2.52		2.59		2.69		2.90
77	Easily frustrated		x		1.97		1.82		2.11		1.91		1.94		1.98		2.04		2.23		2.11
81	Steals at home	x		1.52		1.59		1.30				1.38		1.24				1.71			
82	Steals outside home	x	x	1.22	1.47	1.60	1.25	1.44	1.38		1.37	1.39	1.33	1.50	1.46		1.22	1.82	1.58		1.59
86	Stubborn, irritable	x	x	1.32	1.84	1.66	1.80	1.38	1.84		2.23	1.85	1.98	1.93	2.00		1.87	1.55	1.89		1.79
87	Sudden mood changes	x	x	1.11	1.64	1.50	1.58	1.36	1.78		1.99	1.48	2.09	1.58	1.75		1.52	1.33	1.86		1.73
90	Swearing, obscenity	x	x	1.27	1.69	1.29	1.77	1.41	1.72		1.93	1.59	1.99	1.36	1.85		2.15	1.77	2.05		2.20
93	Talks too much	x	x	0.97	1.58	0.97	1.52	0.86	1.39		1.51	1.03	1.36	1.11	1.59		1.78	0.93	1.88		1.86
94	Teases	x	x	1.54	1.90	1.78	1.92	1.59	1.68		2.09	1.61	2.05	1.85	2.17		2.07	1.69	1.80		1.97
95	Temper tantrums	x	x	1.39	2.45	1.76	2.11	1.63	2.32		2.72	1.86	2.59	1.86	2.62		2.61	1.93	2.96		2.44
96	Thinks about sex too much	x		1.14		0.99		1.31				1.20		1.36				1.40			
97	Threatens others	x	x	2.07	2.66	2.01	2.70	2.21	2.23		3.04	2.61	3.09	3.00	2.88		2.76	3.08	2.74		2.55
98	Tardy		x		0.47		0.15		0.29		0.47		0.44		0.58		0.77		1.06		1.07
101	Truancy	x	x	1.02	0.39	0.89	0.32	1.37	0.32		0.41	0.83	0.38	1.30	0.56		0.64	1.40	0.86		0.80
104	Loud	x	x	1.37	1.96	1.61	1.94	1.60	1.72		2.05	1.63	1.71	1.73	1.78		2.12	1.58	2.26		2.10
105	Alcohol, drugs	x	x	1.09	1.04	0.66	0.94	1.36	0.49		2.49	1.21	0.89	a	0.97		0.90	1.57	1.27		1.24
106	Vandalism	x		1.57		1.68		1.98				2.05		1.98				2.00			

a = no variability or not enough variability for convergence
b = item not administered
M = mother-reported
T = teacher-reported

Table S7. Item severity (b1) parameters by age and rater (Study 2).

				Age 5		Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12		Age 13	
Number	Item	CBCL	TRF	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T
3	Argues	x	x	-1.55	-0.64	-1.60	-0.52	-1.38	-0.48		-0.37	-1.35	-0.29	-1.13	-0.43		-0.47	-0.93	-0.21		-0.52
6	Defiant		x		-0.04		0.02		0.12		0.14		0.14		-0.09		-0.06		0.17		-0.18
7	Brags	x	x	-0.39	0.17	-0.63	0.50	-0.73	0.48		0.46	-0.55	0.59	-0.30	0.26		0.25	-0.13	0.55		0.14
16	Mean to others	x	x	0.40	-0.11	0.29	-0.03	0.36	0.04		0.23	0.58	0.39	0.57	0.14		0.21	0.71	0.62		0.27
19	Demands attention	x	x	-1.14	-0.48	-1.12	-0.65	-0.75	-0.35		-0.21	-0.28	0.05	0.02	-0.14		-0.47	0.16	-0.20		-0.38
20	Destroys own things	x	x	0.23	0.98	0.40	0.71	0.59	1.03		1.38	0.93	1.57	1.13	1.67		1.67	1.25	2.08		1.66
21	Destroys others' things	x	x	0.42	0.59	0.50	0.57	0.55	0.90		1.25	0.78	1.46	0.98	1.22		1.30	1.04	1.67		1.23
22	Disobedient at home	x		-1.07		-0.94		-0.73				-0.44		-0.38				-0.29			
23	Disobedient at school	x	x	-0.36	-0.66	-0.48	-0.67	-0.35	-0.43		-0.22	-0.13	-0.12	-0.10	-0.25		-0.28	-0.06	0.01		-0.33
24	Disturbs others		x		-0.88		-1.11		-0.84		-0.55		-0.42		-0.51		-0.70		-0.38		-0.62
26	Lacks guilt	x	x	-0.10	-0.26	-0.31	-0.20	-0.22	-0.08		0.07	0.07	0.22	0.23	-0.02		0.02	0.13	0.20		-0.17
27	Jealous	x	x	-1.00	0.28	-0.96	0.45	-0.72	0.59		0.74	-0.35	1.11	-0.11	0.86		0.65	0.10	1.28		0.95
37	Fights	x	x	0.59	-0.14	0.41	0.13	0.55	0.23		0.38	0.75	0.53	0.85	0.37		0.56	1.07	1.08		0.80
39	Bad companions	x	x	0.76	-0.11	0.52	-0.20	0.52	0.02		0.15	0.64	0.24	0.64	-0.02		-0.27	0.34	0.03		-0.28
43	Lies (and cheats)	x	x	-0.17	-0.02	-0.34	-0.13	-0.03	0.11		0.42	0.19	0.54	0.27	0.34		0.35	0.21	0.69		0.39
53	Talks out of turn		x		-0.96		-1.03		-0.81		-0.50		-0.32		-0.61		-0.62		-0.36		-0.58
57	Attacks people	x	x	1.36	0.22	1.03	0.39	1.49	0.54		0.77	1.36	0.92	1.49	0.78		0.94	1.55	1.36		1.15
63	Prefers older kids	x	x	-0.50	1.64	-0.72	2.48	-0.56	3.44		2.15	-0.12	2.36	0.07	1.89		1.65	0.02	1.77		1.54
67	Runs away	x		3.50		2.95		3.25				3.43		3.74				2.48			
67	Disrupts class		x		-0.56		-0.58		-0.37		-0.14		0.01		-0.20		-0.21		-0.03		-0.24
68	Screams	x	x	0.61	0.95	0.31	1.05	0.65	1.47		1.43	0.76	1.71	0.84	1.44		1.21	1.09	1.41		1.16
72	Sets fires	x		3.42		4.99		3.09				2.67		3.07				2.88			
72	Messy work		x		-0.68		-1.12		-0.40		-0.16		-0.12		-0.34		b		b		b
74	Shows off	x	x	-1.14	-0.19	-1.00	-0.18	-0.88	0.00		0.03	-0.61	0.23	-0.38	-0.03		-0.21	-0.39	-0.03		-0.27
76	Explosive		x		0.29		0.44		0.55		0.59		0.75		0.49		0.28		0.67		0.28
77	Easily frustrated		x		0.10		0.08		0.21		0.33		0.53		0.30		-0.12		0.22		-0.09
81	Steals at home	x		1.73		1.54		1.93				1.95		2.17				1.70			
82	Steals outside home	x	x	2.25	0.99	1.84	0.85	2.08	1.09		1.51	2.33	1.70	2.38	1.48		2.04	1.79	2.28		1.90
86	Stubborn, irritable	x	x	-1.10	-0.30	-1.09	-0.26	-0.85	-0.22		-0.01	-0.52	0.14	-0.33	-0.12		-0.39	-0.37	0.04		-0.31
87	Sudden mood changes	x	x	0.20	0.33	-0.29	0.28	0.00	0.43		0.50	0.19	0.80	0.28	0.52		-0.03	0.29	0.40		0.07
90	Swearing, obscenity	x	x	0.94	0.79	0.91	0.94	1.09	1.03		1.18	0.98	1.13	1.21	0.81		0.74	0.66	1.08		0.50
93	Talks too much	x	x	-0.92	-0.58	-1.08	-0.56	-1.22	-0.55		-0.24	-0.25	-0.12	-0.07	-0.30		-0.59	0.09	-0.39		-0.63
94	Teases	x	x	-0.14	0.00	-0.31	0.11	-0.26	0.25		0.30	-0.18	0.44	-0.06	0.10		-0.18	0.03	0.26		-0.06
95	Temper tantrums	x	x	-0.44	0.36	-0.46	0.37	-0.21	0.53		0.54	0.08	0.63	0.29	0.33		0.17	0.28	0.57		0.17
96	Thinks about sex too much	x		2.91		2.66		2.07				2.93		2.76				2.22			
97	Threatens others	x	x	1.16	0.64	1.01	0.68	1.08	0.79		0.82	1.13	0.85	1.25	0.55		0.48	1.17	0.97		0.69
98	Tardy		x		2.96		7.86		4.47		3.32		3.43		2.33		1.19		1.02		0.59
101	Truancy	x	x	4.83	4.35	5.08	6.13	4.17	6.14		5.29	5.71	5.25	3.62	3.51		2.69	2.39	2.13		1.79
104	Loud	x	x	-0.09	0.17	-0.16	0.29	-0.03	0.53		0.61	0.24	0.90	0.42	0.58		0.35	0.55	0.57		0.22
105	Alcohol, drugs	x	x	6.41	6.16	8.93	6.42	5.03	10.33		3.68	6.37	6.61	a	5.57		4.80	2.88	3.43		2.76
106	Vandalism	x		2.51		2.42		2.48				2.55		2.65				2.35			

a = no variability or not enough variability for convergence
b = item not administered
M = mother-reported
T = teacher-reported

Table S8. Item severity (b2) parameters by age and rater (Study 2).

				Age 5		Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12		Age 13	
Number	Item	CBCL	TRF	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T
3	Argues	x	x	1.09	0.50	1.16	0.72	1.20	0.81		0.91	1.00	1.10	0.89	0.85		0.87	1.14	1.11		0.80
6	Defiant		x		0.93		1.15		1.23		1.28		1.37		1.03		1.13		1.32		0.97
7	Brags	x	x	2.82	1.82	2.17	2.21	2.11	2.29		2.14	2.31	2.01	2.26	1.71		1.78	2.14	2.07		1.48
16	Mean to others	x	x	2.66	0.95	2.50	1.28	2.26	1.39		1.54	2.43	1.62	2.37	1.41		1.64	2.59	2.04		1.56
19	Demands attention	x	x	1.04	0.53	0.84	0.74	1.11	0.86		1.09	1.39	1.32	1.80	1.08		0.78	1.90	1.04		0.78
20	Destroys own things	x	x	1.96	2.18	2.13	1.94	2.37	2.26		2.56	2.65	3.01	2.78	3.23		2.93	2.66	3.77		2.82
21	Destroys others' things	x	x	2.10	1.87	2.25	1.88	2.24	2.25		2.51	2.29	2.82	2.79	2.80		2.59	2.37	2.91		2.57
22	Disobedient at home	x		1.78		1.88		2.10				2.04		1.74				1.82			
23	Disobedient at school	x	x	2.58	0.47	2.36	0.76	2.21	0.85		1.19	2.34	1.30	2.11	1.01		1.04	1.76	1.35		1.00
24	Disturbs others		x		0.32		0.51		0.72		0.98		1.14		0.87		0.88		1.17		0.79
26	Lacks guilt	x	x	2.55	0.75	2.77	1.13	2.66	1.04		1.30	2.37	1.45	2.27	1.11		1.07	2.58	1.43		0.98
27	Jealous	x	x	1.70	2.06	1.50	2.54	2.04	2.44		2.58	1.81	2.97	2.09	2.47		2.23	2.34	3.24		2.46
37	Fights	x	x	2.43	1.00	2.27	1.39	2.41	1.54		1.60	2.30	1.87	2.50	1.57		1.78	2.54	2.49		1.97
39	Bad companions	x	x	2.65	1.17	2.40	1.35	2.35	1.61		1.57	2.59	1.70	2.80	1.32		1.36	2.32	1.70		1.15
43	Lies (and cheats)	x	x	2.23	1.41	1.80	1.55	2.24	1.70		2.08	2.26	1.98	2.12	1.82		1.77	2.03	2.20		1.56
53	Talks out of turn		x		0.29		0.48		0.66		0.85		1.21		0.87		0.70		1.01		0.76
57	Attacks people	x	x	3.70	1.33	2.79	1.75	3.09	1.89		2.09	3.07	2.37	2.66	2.23		2.20	3.09	2.59		2.25
63	Prefers older kids	x	x	2.25	3.77	1.95	4.47	2.37	6.75		4.15	2.34	4.20	2.95	3.78		4.12	2.32	4.64		4.18
67	Runs away	x		5.09		4.19		4.51				a		6.27				3.55			
67	Disrupts class		x		0.37		0.62		0.81		1.02		1.31		0.92		0.97		1.24		0.86
68	Screams	x	x	2.47	1.91	2.13	2.43	2.55	2.88		2.49	2.37	2.92	2.24	2.66		2.32	2.77	2.32		2.32
72	Sets fires	x		6.15		a		4.52				4.16		4.53				4.29			
72	Messy work		x		1.56		1.75		2.30		2.49		2.86		2.43		b		b		b
74	Shows off	x	x	1.46	0.98	1.49	1.19	1.62	1.47		1.48	1.70	1.71	1.87	1.39		1.07	1.67	1.33		0.97
76	Explosive		x		1.31		1.53		1.53		1.59		1.84		1.62		1.46		1.65		1.28
77	Easily frustrated		x		1.30		1.30		1.33		1.66		1.86		1.59		1.15		1.43		1.20
81	Steals at home	x		3.52		3.17		4.22				3.88		4.33				3.10			
82	Steals outside home	x	x	4.54	2.31	3.48	2.36	4.31	2.59		2.90	4.29	3.27	4.07	2.88		3.29	3.06	3.12		2.62
86	Stubborn, irritable	x	x	1.54	0.98	1.30	1.14	1.73	1.31		1.42	1.62	1.80	1.62	1.43		1.24	1.91	1.59		1.34
87	Sudden mood changes	x	x	2.87	1.62	1.94	1.77	2.53	1.84		1.95	2.41	2.15	2.30	2.00		1.70	2.58	1.90		1.62
90	Swearing, obscenity	x	x	3.61	1.99	2.91	2.24	3.01	2.68		2.37	2.71	2.54	3.00	2.20		1.91	2.41	2.06		1.64
93	Talks too much	x	x	1.39	0.62	1.37	0.78	1.71	1.02		1.36	1.91	1.54	2.20	1.14		0.81	2.77	1.01		0.80
94	Teases	x	x	2.11	1.37	1.91	1.57	2.08	1.81		1.66	2.02	1.94	2.00	1.45		1.44	2.03	1.74		1.24
95	Temper tantrums	x	x	1.59	1.26	1.47	1.34	1.74	1.56		1.53	1.54	1.77	1.80	1.29		1.25	1.79	1.55		1.31
96	Thinks about sex too much	x		4.80		4.81		3.88				4.47		4.16				3.36			
97	Threatens others	x	x	2.93	1.62	3.01	1.87	2.82	2.12		1.90	2.48	2.07	2.50	1.72		1.54	2.52	1.97		1.75
98	Tardy		x		5.67		16.59		8.91		6.30		6.89		4.49		2.90		2.42		2.06
101	Truancy	x	x	a	6.95	7.63	8.90	5.30	9.39		7.47	8.00	8.11	5.50	5.54		4.18	3.36	3.22		3.12
104	Loud	x	x	1.91	1.16	1.82	1.37	1.87	1.67		1.81	2.11	2.20	2.13	1.72		1.47	2.28	1.62		1.30
105	Alcohol, drugs	x	x	a	a	a	7.91	a	12.91		4.12	a	8.42	a	7.57		6.61	4.12	4.22		3.67
106	Vandalism	x		4.07		4.59		3.60				3.43		3.81				3.28			

a = no variability or not enough variability for convergence
b = item not administered
M = mother-reported
T = teacher-reported

Table S9. Item information by age and rater (Study 2).

				Age 5		Age 6		Age 7		Age 8		Age 9		Age 10		Age 11		Age 12		Age 13	
Number	Item	CBCL	TRF	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T	M	T
3	Argues	x	x	0.40	0.20	0.47	0.31	0.53	0.38		0.43	0.42	0.66	0.41	0.37		0.41	0.46	0.67		0.32
6	Defiant		x		0.47		0.72		0.81		1.00		1.13		0.58		0.73		1.14		0.49
7	Brags	x	x	0.34	0.69	0.52	0.67	0.63	0.70		0.82	0.54	1.05	0.69	0.92		0.83	0.74	0.77		0.71
16	Mean to others	x	x	1.18	0.49	1.09	0.83	1.67	0.95		1.31	1.43	1.48	1.89	1.09		1.23	1.42	1.44		1.18
19	Demands attention	x	x	0.33	0.23	0.33	0.32	0.43	0.40		0.54	0.60	0.66	0.69	0.56		0.36	0.75	0.52		0.36
20	Destroys own things	x	x	0.86	1.29	0.87	0.97	0.80	1.26		1.19	0.94	1.13	0.89	0.86		0.96	1.19	0.76		1.09
21	Destroys others' things	x	x	1.21	1.39	1.11	1.37	1.20	1.78		1.87	1.55	1.89	1.11	1.35		1.54	1.64	1.49		1.58
22	Disobedient at home	x		0.87		1.03		1.09				1.32		1.65				1.46			
23	Disobedient at school	x	x	0.58	0.11	0.58	0.26	0.77	0.37		0.88	0.79	1.15	1.03	0.52		0.57	1.30	1.26		0.51
24	Disturbs others		x		0.08		0.17		0.29		0.52		0.72		0.40		0.41		0.77		0.33
26	Lacks guilt	x	x	0.45	0.33	0.35	0.55	0.53	0.54		0.79	0.65	1.00	0.70	0.65		0.58	0.70	0.82		0.50
27	Jealous	x	x	0.36	0.56	0.41	0.50	0.43	0.48		0.64	0.58	0.64	0.55	0.70		0.70	0.56	0.63		0.81
37	Fights	x	x	1.06	0.54	1.24	0.88	1.15	1.22		1.25	1.55	1.56	1.31	1.34		1.34	1.26	1.31		1.44
39	Bad companions	x	x	0.68	0.53	0.64	0.58	0.76	0.56		0.80	0.77	0.77	0.66	0.64		0.50	0.86	0.46		0.47
43	Lies (and cheats)	x	x	0.90	0.74	1.09	0.70	1.05	0.87		0.96	1.18	1.10	1.09	0.94		1.00	1.38	1.02		0.89
53	Talks out of turn		x		0.15		0.23		0.30		0.40		0.63		0.41		0.30		0.53		0.35
57	Attacks people	x	x	0.90	0.94	1.38	1.18	1.07	1.67		1.85	1.62	1.77	1.99	1.51		1.43	2.00	1.59		1.64
63	Prefers older kids	x	x	0.30	0.30	0.28	0.21	0.25	0.09		0.31	0.39	0.35	0.27	0.34		0.20	0.39	0.19		0.19
67	Runs away	x		0.35		0.82		0.55				0.41		0.28				1.01			
67	Disrupts class		x		0.09		0.18		0.34		0.55		1.05		0.38		0.47		0.98		0.36
68	Screams	x	x	0.71	1.64	0.80	1.62	0.82	1.29		1.81	1.09	1.72	1.25	1.35		1.48	1.00	1.84		1.26
72	Sets fires	x		0.22		0.08		0.43				0.82		0.50				0.72			
72	Messy work		x		0.24		0.15		0.20		0.20		0.18		0.20		b		b		b
74	Shows off	x	x	0.51	0.49	0.62	0.58	0.65	0.62		0.75	0.77	0.88	0.72	0.71		0.53	0.77	0.68		0.49
76	Explosive		x		0.95		1.22		1.19		1.61		2.14		1.41		1.19		1.58		0.99
77	Easily frustrated		x		0.73		0.67		0.82		0.98		1.15		0.98		0.63		0.98		0.69
81	Steals at home	x		0.91		1.01		0.65				0.75		0.59				1.19			
82	Steals outside home	x	x	0.56	0.86	1.01	0.64	0.77	0.79		0.80	0.71	0.74	0.83	0.89		0.63	1.35	0.99		1.04
86	Stubborn, irritable	x	x	0.51	0.47	0.60	0.56	0.60	0.68		0.96	0.90	1.13	0.96	0.85		0.65	0.79	0.92		0.68
87	Sudden mood changes	x	x	0.47	0.77	0.75	0.79	0.68	0.99		1.25	0.80	1.49	0.91	1.04		0.71	0.67	1.10		0.83
90	Swearing, obscenity	x	x	0.56	1.00	0.67	1.17	0.78	1.14		1.44	0.99	1.51	0.76	1.24		1.42	1.15	1.45		1.19
93	Talks too much	x	x	0.30	0.28	0.29	0.33	0.26	0.39		0.57	0.39	0.55	0.47	0.50		0.37	0.35	0.49		0.37
94	Teases	x	x	0.82	0.76	0.99	0.92	0.86	0.88		1.12	0.87	1.30	1.10	0.97		0.89	0.96	0.95		0.68
95	Temper tantrums	x	x	0.57	0.84	0.75	0.82	0.80	1.19		1.38	0.86	1.69	1.04	0.93		0.87	1.09	1.56		0.91
96	Thinks about sex too much	x		0.42		0.35		0.68				0.46		0.61				0.79			
97	Threatens others	x	x	1.47	1.51	1.28	1.92	1.63	1.64		2.36	2.33	2.64	2.97	1.85		1.42	2.97	2.16		1.63
98	Tardy		x		0.09		0.01		0.03		0.08		0.07		0.14		0.26		0.47		0.44
101	Truancy	x	x	0.10	0.05	0.09	0.03	0.20	0.03		0.04	0.05	0.04	0.33	0.11		0.16	0.75	0.31		0.28
104	Loud	x	x	0.65	0.62	0.82	0.78	0.83	0.86		1.24	0.93	1.09	1.05	0.94		0.96	0.93	1.22		0.78
105	Alcohol, drugs	x	x	0.02	0.03	0.01	0.03	0.07	0.01		0.38	0.02	0.03	a	0.06		0.10	0.72	0.37		0.53
106	Vandalism	x		0.86		0.96		1.33				1.38		1.21				1.45			

a = no variability or not enough variability for convergence
b = item not administered
M = mother-reported
T = teacher-reported