Title

Measurement bias in caregiver-report of early childhood behavior problems across demographic factors in an ECHO-wide diverse sample

Permalink

https://escholarship.org/uc/item/7d23q41v

Journal

JCPP Advances, 4(1)

ISSN

2692-9384

Authors

Zheng, Shuting
Mansolf, Maxwell
McGrath, Monica
et al.

Publication Date

2024-03-01

DOI

10.1002/jcv2.12198

Received: 13 January 2023 | Accepted: 11 August 2023

DOI: 10.1002/jcv2.12198

ORIGINAL ARTICLE

JCPP Advances ACAMH

Measurement bias in caregiver-report of early childhood behavior problems across demographic factors in an ECHO-wide diverse sample

Shuting Zheng[1] | Maxwell Mansolf[2] | Monica McGrath[3] | Marie L. Churchill[3] | Traci A. Bekelman[4] | Patricia A. Brennan[5] | Amy E. Margolis[6] | Sara S. Nozadi[7] | Theresa M. Bastain[8] | Amy J. Elliott[9] | Kaja Z. LeWinn[1] | Julie A. Hofheimer[10] | Leslie D. Leve[11] | Brandon Rennie[12] | Emily Zimmerman[13] | Carmen A. Marable[14] | Cindy T. McEvoy[15] | Chang Liu[16] | Alexis Sullivan[17] | Tracey J. Woodruff[1] | Samiran Ghosh[18] | Bennett Leventhal[19] | Assiamira Ferrara[20] | Johnnye Lewis[21] | Somer Bishop[22] | on behalf of program collaborators for Environmental influences on Child Health Outcomes

[1]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA, USA

[2]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

[3]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[4]Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA

[5]Department of Psychology, Emory University, Atlanta, GA, USA

[6]Department of Psychiatry, Columbia University Irving Medical Center, New York State Psychiatric Institute, New York, NY, USA

[7]Community Environmental Health, College of Pharmacy, Health Sciences Center, University of New Mexico, Albuquerque, NM, USA

[8]Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA

[9]Avera Research Institute, Sioux Falls, SD, USA

[10]Department of Pediatrics, Division of Neonatal-Perinatal Medicine, North Carolina at Chapel Hill, Chapel Hill, NC, USA

[11]Prevention Science Institute, University of Oregon, Eugene, OR, USA

[12]Health Sciences Center, Department of Pediatrics, Center for Development and Disability, University of New Mexico, Navajo Birth Cohort Study, Albuquerque, NM, USA

[13]Communication Sciences & Disorders, Northeastern University, Boston, MA, USA

[14]School of Medicine, University of North Carolina at Chapel Hill, Neuroscience Curriculum, Chapel Hill, NC, USA

[15]Department of Pediatrics, Pape Pediatric Research Institute, Oregon Health & Science University, Portland, OR, USA

[16]Department of Psychology, Washington State University, Pullman, WA, USA

[17]Center for Health and Community, University of California, San Francisco, CA, USA

[18]Department of Biostatistics and Data Science & Coordinating Center for Clinical Trials (CCCT), University of Texas School of Public Health, Houston, TX, USA

[19]University of Chicago, Navajo Birth Cohort Study, University of New Mexico, Albuquerque, NM, USA

[20]Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA

[21]Navajo Birth Cohort Study, Community Environmental Health Program, College of Pharmacy, University of New Mexico, Albuquerque, NM, USA

[22]Department of Psychiatry and Behavioral Sciences, Weill Institute for Neurosciences, University of California, San Francisco, CA, USA

Shuting Zheng and Maxwell Mansolf contribute equally to this study.

See Acknowledgments for full listing of collaborators

Correspondence
Shuting Zheng, 675 18th Street, San Francisco, CA 94107, Department of Psychiatry and Behavioral Sciences, Weill Institute for Neurosciences, University of California, San Francisco;
Email: shuting.zheng@ucsf.edu

Funding information
Environmental influences on Child Health Outcomes program, Office of The Director, National Institutes of Health, Grant/Award Numbers: U24OD023319 with co-funding from the Office of Behavioral and Social Science Research (PROCore), U24OD023382 (Data Analysis Center), U2COD023375 (Coordinating Center), UH3 OD023248 (Dabelea), UH3 OD023249 (Stanford), UH3 OD023251 (Alshawabkeh), UH3 OD023271 (Karr), UH3 OD023272 (Schantz), UH3 OD023279 (Elliott), UH3 OD023285 (Kerver), UH3 OD023287 (Breton), UH3 OD023288 (McEvoy), UH3 OD023289 (Ferrara), UH3 OD023290 (Herbstman), UH3 OD023305 (Trasande), UH3 OD023313 (Koinis-Mitchell), UH3 OD023318 (Dunlop), UH3OD023320 (Aschner), UH3 OD023328 (Duarte), UH3 OD023337 (Wright), UH3 OD023347 (Lester), UH3 OD023349 (O'Connor), UH3 OD023389 (Leve), UH3OD023344 (MacKenzie & Lewis), UH3OD023348 (O'Shea)

Abstract

Background: Research and clinical practice rely heavily on caregiver-report measures, such as the Child Behavior Checklist 1.5–5 (CBCL/1.5-5), to gather information about early childhood behavior problems and to screen for child psychopathology. While studies have shown that demographic variables influence caregiver ratings of behavior problems, the extent to which the CBCL/1.5-5 functions equivalently at the item level across diverse samples is unknown.

Methods: Item-level data of CBCL/1.5-5 from a large sample of young children (N = 9087) were drawn from 26 cohorts in the Environmental influences on Child Health Outcomes program. Factor analyses and the alignment method were applied to examine measurement invariance (MI) and differential item functioning (DIF) across child (age, sex, bilingual status, and neurodevelopmental disorders), and caregiver (sex, education level, household income level, depression, and language version administered) characteristics. Child race was examined in sensitivity analyses.

Results: Items with the most impactful DIF across child and caregiver groupings were identified for Internalizing, Externalizing, and Total Problems. The robust item sets, excluding the high DIF items, showed good reliability and high correlation with the original Internalizing and Total Problems scales, with lower reliability for Externalizing. Language version of CBCL administration, education level and sex of the caregiver respondent showed the most significant impact on MI, followed by child age. Sensitivity analyses revealed that child race has a unique impact on DIF over and above socioeconomic status.

Conclusions: The CBCL/1.5-5, a caregiver-report measure of early childhood behavior problems, showed bias across demographic groups. Robust item sets with less DIF can measure Internalizing and Total Problems equally as well as the full item sets, with slightly lower reliability for Externalizing, and can be crosswalked to the metric of the full item set, enabling calculation of normed T scores based on more robust item sets.

KEYWORDS
behavior problems, behavioral measures, pre-school children, psychometrics

INTRODUCTION

Behavior problems, including noncompliance, emotional distress and outbursts, and disruptive behavior, are common during the toddler and preschool years (Wakschlag et al., 2007). While many such behavior problems are developmentally normative, some young children exhibit clinically significant behavior problems that disrupt child and family functioning (Keenan & Wakschlag, 2002). Population-based studies have reported prevalence estimates of elevated behavioral problems ranging from 7% to about 20% during the preschool years (Bayer et al., 2012; Briggs-gowan et al., 2001; Egger & Angold, 2006; Stülb et al., 2019). Importantly, clinically significant early behavior problems may be predictive of later psychopathology and adverse outcomes in adolescence and adulthood (Campbell, 1995; Eisenberg et al., 2009; Mathiesen & Sanson, 2000; Prior et al., 1992), underscoring the importance of early identification and intervention.

Key points

- Caregiver characteristics, especially language version, education level and sex of the caregiver respondent, greatly impacted the differential functioning of many CBCL/1.5-5 items.
- Child age also influenced the measurement of child behavior problems on the CBCL/1.5-5.
- SES variables (i.e., caregiver education and household income levels) cannot fully account for measurement bias related to child race.
- Robust item sets with less DIF can reliably capture Internalizing, Externalizing, and Total Problems with less measurement bias

Reliable and valid measures are central to assessment and treatment. For preschoolers, the Child Behavior Checklist for Ages 1.5–5 (CBCL/1.5–5) is one of the most commonly used measures for screening for behavior problems in both clinical and research settings (Ivanova et al., 2010; Medeiros et al., 2017), as it shows good reliability and validity in studies globally (Ivanova et al., 2010; Konold et al., 2003; Rescorla et al., 2011). In their initial validation, Achenbach and Rescorla (2001) described a factor structure of seven syndrome scales, which were then grouped under two second-order factors of *Internalizing* and *Externalizing Problems.* Subsequent confirmatory factor analyses have generally found this structure holds in samples with different compositions of nationality, culture, sex, and diagnosis (Ivanova et al., 2010; Konold et al., 2003; Koot et al., 1997; Medeiros et al., 2017; Tan et al., 2007). Accumulating research evidence supports the use of the two broad domains of *Internalizing* and *Externalizing* as transdiagnostic constructs for profiling clinically significant behavior problems in young children (Achenbach et al., 2016; Krueger & Markon, 2011).

Despite basic psychometric validations of the CBCL/1.5–5 (e.g., reliability, validity, structure), questions remain about whether it measures behavior problems equivalently across samples with different demographic compositions (i.e., does the same score reflect similar levels of problems regardless of child or caregiver demographic characteristics?). This warrants further investigation, as previous research has shown that measured/observed levels of child behavior problems can be affected by child factors, including age, sex, bilingual status, and developmental level (Carneiro et al., 2016; Chen, 2010; Sun et al., 2021; Wakschlag et al., 2017), as well as by caregiver characteristics, including informant's primary language, sex (father vs. mother), socioeconomic status (SES), and mental health (Davis & Qi, 2022; Flouri et al., 2017; Müller et al., 2011; Schroeder et al., 2010). Given that observed raw score differences across demographic groups could result from measurement bias and/or *true* group differences, it is necessary to ensure measurement equivalence or account for measurement bias if they exist before concluding that measured differences across groups reflect actual differences in levels of behavior problems (Vandenberg & Lance, 2000). For example, developmentally, younger children might use *whining* (Item 97) to communicate more often regardless of their levels of internalizing problems, in which case, high scores on this item might not be as reflective of the *true* levels of internalizing problems in young children as in older children.

Analyses of measurement invariance (MI) and differential item functioning (DIF) allow examination of measurement bias across groups to facilitate the estimation of actual group differences after accounting for measurement bias. Surprisingly, despite its widespread use, little MI/DIF work has investigated whether the CBCL/1.5–5 functions equivalently across groups that differ by child and caregiver characteristics. We only identified three studies that investigated MI/DIF of the CBCL/1.5–5 (Dovgan et al., 2019; Gross et al., 2006; Rescorla et al., 2019). Rescorla and colleagues reported that the CBCL/1.5–5 Autism Spectrum Subscale functioned equivalently across ages 18 months, 3 years, and 5 years, demonstrating its potential utility for tracking longitudinal changes across the preschool years. In a study of children with autism spectrum disorder (ASD), Dovgan et al. (2019) found that CBCL/1.5–5 syndrome subscales, emotional reactivity, anxious/depressed, and somatic complaints were non-invariant between those with and without intellectual disability (ID), raising concerns about the application of these syndrome scales in those with ASD and ID. When examining MI of *Internalizing* and *Externalizing* scales across groups defined by parent race/ethnicity, family income, and language version (English vs. Spanish), Gross et al. (2006) found differential functioning of items on both the *Internalizing* and *Externalizing* scales. However, their findings are limited by the selective set of grouping variables tested. In sum, studies of CBCL/1.5–5 support partial invariance of the scale across certain subgroups and further investigation is warranted to inform the use of CBCL/1.5–5 in diverse samples.

To date, no study has systematically assessed the MI/DIF of CBCL/1.5–5 across a broad range of child and caregiver-level characteristics, likely due to the lack of a large enough sample with sufficient diversity across multiple relevant characteristics. This type of MI/DIF analysis is necessary to ensure that the measure can be reliably and validly employed to measure behavior problems in diverse groups of young children. Data from the National Institutes of Health (NIH) Environmental influences on Child Health Outcomes (ECHO) program provide a unique opportunity for a comprehensive psychometric examination of CBCL/1.5–5 across multiple child and caregiver characteristics. Therefore, the current study leveraged these data to examine the configural invariance and item-level MI/DIF of CBCL/1.5–5 across a wide array of child and caregiver (i.e., respondent who completed the CBCL/1.5–5) characteristics that are available in the ECHO dataset and have been reported to influence caregiver-report of child behavior problems. Moreover, we also explore MI/DIF across child race groups in the context of SES variables as sensitivity analyses.

METHODS

Participants

Data for the current study were drawn from the NIH ECHO program. For more information on ECHO, see Blaisdell et al. (2021). The aggregated ECHO dataset was queried to identify cases from 26 out of 69 cohorts that met the inclusion criteria: (a) CBCL/1.5–5 was administered to caregivers of children aged 18–71 months; (b) CBCL/1.5–5 item-level data were available; (c) data were available on at least one of the child or caregiver-characteristic variables of interest. When multiple timepoints of CBCL/1.5–5 data were available for a single child, the first administration was taken, yielding a final sample size of 9087 CBCL/1.5–5 administrations. See Table 1 for distributions of child and caregiver characteristics. For each MI/DIF testing, only individuals with data on the assessed variable(s) were included in the corresponding analyses (i.e., individuals missing data on any of the specific variables were excluded for the specific set of analyses), allowing for the largest inclusion sample possible for each analysis (see Table 1 for proportions of missing data by child and caregiver characteristics).

TABLE 1 Demographic information of the analytical sample of caregiver-child dyads in Environmental influences on Child Health Outcomes (ECHO) (*N* = 9087).

Child Characteristics	
Female, *n* (%)	4327 (47.6%)
Age in months: M (SD)	39.2 (14.5)
Age subgroups, *n* (%)	
[18–27 m)	2784 (30.6%)
[27–36 m)	1585 (17.4%)
[36–45 m)	1408 (15.5%)
[45–54 m)	1259 (13.9%)
[54–72 m)	2051 (22.6%)
Race, *n* (%)	
White	4499 (49.5%)
Asian	494 (5.4%)
Black	1768 (19.5%)
Multiple race	1075 (11.8%)
Other	
American Indian or Alaska native[a]	146 (1.6%)
Native Hawaiian or other Pacific Islander[a]	42 (0.5%)
Other -participant self-report	691 (7.6%)
Missing	372 (4.1%)
Ethnicity, *n* (%)	
Hispanic	2923 (32.2%)
non-Hispanic	6019 (66.2%)
Missing	145 (1.6%)
Bilingual home environment, *n* (%)	
Yes	2931 (32.3%)
No	5056 (55.6%)
Missing	1100 (12.1%)
Any NDD, *n* (%)	
Yes	713 (7.8%)
No	4888 (53.8%)
Missing	3486 (38.4%)
Intellectual disability[b]	37 (0.4%)
Developmental delay[b]	225 (2.5%)
Autism, ASD, or Pervasive Developmental Disorder[b]	112 (1.2%)
Attention Deficit Disorder or Attention Deficit/Hyperactivity Disorder[b]	46 (0.5%)
Speech disorder[b]	568 (6.3%)
Learning disability[b]	84 (0.9%)
Respondent characteristics	
Female respondent, *n* (%)	8726 (96%)
Biological mother respondent, *n* (%)	8496 (93.5%)
CBCL language of administration, *n* (%)	
English	7596 (83.6%)
Spanish	923 (10.2%)
Missing	568 (6.3%)
Highest caregiver education, *n* (%)	
Less than high school	731 (8%)
High school degree or equivalent	1789 (19.8%)
Some college	2213 (24.4%)
Bachelor's degree	2148 (23.6%)
Master's degree or higher	2001 (22%)
Missing	196 (2.2%)
Household income, *n* (%)	
Less than $30,000	2084 (22.9%)
$30,000-$49,999	847 (9.3%)
$50,000-$74,999	1028 (11.3%)
$75,000-$99,999	748 (8.2%)
$100,000 or more	2837 (31.2%)
Missing	1543 (17%)
Caregiver dep. T score (M, SD)	46.5 (8.5)
Caregiver depression, *n* (%)	
$T < 60$	4709 (51.8%)
$T \geq 60$	333 (3.7%)
Missing	4045 (44.5%)

[a] For the MI/DIF analysis, categories of "American Indian or Alaska Native" and "Native Hawaiian or other Pacific Islander" were collapsed into "Other".

[b] Denominators used for calculating percentages of children with each NDD include missing values.

Measures

CBCL/1.5–5 (Achenbach & Rescorla, 2001) requires caregivers to rate their child's behaviors "now or within the past 2 months" on items describing behavior problems on a three-point scale: 0 "Not True", 1 "Somewhat or Sometimes True", to 2 "Very True or Often True". The ***Internalizing*** scale includes 36 items (score range: 0–72) across syndrome scales of Emotionally Reactive, Anxious/Depressed, Somatic Complaints, and Withdrawn; the ***Externalizing*** scale includes 24 items (score range: 0–48) across syndrome scales of Attention Problems and Aggressive Behaviors; and the ***Total Problems*** include all items from both the *Internalizing* and *Externalizing* problems, with additional items from syndrome scales of Sleep Problems and Other Problems, totaling 99 items (score range: 0–198). Frequencies of item endorsement are provided in Table S1. *Internalizing*, *Externalizing*, and *Total Problems* are the constructs of primary interest in the current analyses.

Grouping variables for measurement invariance/differential item functioning testing

We tested MI/DIF across univariate groupings defined by single child and caregiver characteristics separately, and across multivariate groupings defined by two child and caregiver variables in combination. Throughout, we use the term *grouping* to refer to subgroups defined by various levels of a given characteristic or combination of characteristics.

Univariate groupings

Child characteristics of interest included biological sex (male vs. female), age at the administration of CBCL/1.5–5 (18–27, 27–36, 36–45, 45–54, and 54–72 months, with children at the group boundaries assigned to the older group), bilingual/multilingual status (yes vs. no), and caregiver-reported diagnosis of a neurodevelopmental disorder (NDD) (none vs. any). For child bilingual and NDD variables, we included data from the same timepoint (i.e., within 6 months of CBCL/1.5–5 administration date) when available and otherwise used historic data from the closest previous timepoint to keep more cases for analysis. Bilingual status was determined using a harmonized indicator of bilingual/multilingual exposures, incorporating direct reports of bilingual exposure and language of administration for child tests within ECHO. We used the following NDD diagnoses reported on the ECHO medical history form: ASD, intellectual/developmental disorders, attention deficit/hyperactivity disorder, learning disability, and speech disorder. If the child was reported to have one or more of these diagnoses, they were classified as having NDD.

Caregiver characteristics included respondent sex (male vs. female), language version of CBCL administration (English vs. Spanish; the only two available in the ECHO data), annual household income levels (<$30,000; $30,000-$49,999; $50,000-$74,999; $75,000-$99,999; ≥$100,000), caregiver education level, and caregiver depression. Given the low number of male caregivers completing CBCL/1.5–5, caregiver educational level and depression status were only examined for female caregivers (mostly mothers). Education level was harmonized to derive four categories: less than high school; high school diploma or equivalent; some college; Bachelor's degree and above. Caregiver depression was indexed using the PROMIS® Depression T score (mean of 50 and standard deviation of 10 normed in reference to U.S. adults) as a common metric to which multiple instruments have been linked (Blackwell et al., 2021; Choi et al., 2014; Kaat et al., 2017). For the current analyses, caregiver depression was dichotomized to form groups according to the recommended threshold *T* score = 60 to distinguish cases with elevated depression levels. For caregiver depression status, only data from the same timepoint of CBCL/1.5–5 were used; for other caregiver demographic variables, historic data from the closest previous timepoint was used when the variable was not available from the same timepoint as CBCL/1.5–5.

Multivariate groupings

To understand the differential impact of variables on MI/DIF, we examined multivariate groupings defined by combinations of child or caregiver characteristics. Based on data availability and known confounding impacts on child behavior problems, we assessed MI/DIF across groupings defined by: (a) child sex and age; (b) child sex and NDD diagnosis (May et al., 2019); and (c) caregiver education and income levels (Braveman et al., 2005). To yield sufficient sample sizes, caregiver education was collapsed into two categories (less than Bachelor's degree, Bachelor's degree or higher) and income was collapsed into three categories (<$50,000, $50,000-$74,999, $75,000 or more).

Sensitivity Analyses. We conceptualized race as a social-cultural variable that is confounded with SES variables (Cheng, Goodman, & The Committee on Pediatric Research, 2015) and thus, conducted sensitivity analyses to evaluate the unique impact of child race/ethnicity on MI/DIF, as previous studies have shown the impact of race on the measurement of behavior problems and psychopathology in children (Gross et al., 2006; Vaughn-Coaxum et al., 2016). Race categories were collapsed to yield sufficient sample sizes: White, Black, Asian, Multiple Race, and Other (see Table 1 for all available racial groups). Ethnicity categories were not examined, given insufficient sample sizes. We chose to focus on child race, rather than the respondent race, to capture the full picture of the child's developmental context as it reflects the race of both parents. We tested combinations of race and caregiver education level, as well as race and income level (collapsed into two categories: <$50,000, $50,000 or more), to disentangle the impact of race in the context of SES variables.

Given language version variable is potentially confounded with other variables (i.e., bilingual status, race/ethnicity, caregiver education level, household income level), we conducted DIF testing across multivariate groupings between language version and bilingual status and caregiver education level where the sample sizes were sufficient for analyses (i.e., more than 100 cases within each level of combined groupings).

Statistical analyses

Separate analyses were conducted for each of the univariate and multivariate groupings for each latent construct of *Internalizing*, *Externalizing*, and *Total Problems*. All item response theory (IRT) models were estimated using the graded response model (Samejima, 1997) using the *mirt* package (Chalmers, 2012) in R 4.1.0 (R Core Team, 2022).

Configural invariance testing

Establishing configural invariance (similar number of factors and loading pattern) is the first step in MI/DIF examination. For *Internalizing* and *Externalizing*, we estimated a unidimensional model in each subsample. For *Total Problems*, we estimated both a unidimensional model and a bifactor model (i.e., *Internalizing* and *Externalizing* items loaded on two separate factors and all items loaded on the general factor representing *Total Problems*, with all factors mutually orthogonal). Model fit was evaluated using the standardized root mean squared residual (SRMSR) and root mean squared error of approximation (RMSEA), given their demonstrated performance advantage in IRT models (Maydeu-Olivares & Joe, 2014) and the general incomparability of other estimable fit measures (e.g., comparative fit index [CFI], Tucker-Lewis index [TLI]) from IRT to commonly used criteria (Yuan & Chan, 2005). Standardized root mean squared residual below 0.08 was used as the criterion for adequate fit (Hu & Bentler, 1999), while RMSEA is reported as a secondary index of fit. For *Total*

Problems, the unidimensional and bifactor models were compared with respect to Akaike information criterion (AIC) and sample size-adjusted Bayesian Information Criterion (SABIC), where a difference greater than 10 indicated a superior fit for the model with lower deviance (Raftery, 1995).

Alignment method for measurement invariance/differential item functioning estimation

We applied the alignment method for MI/DIF testing in the current study to accommodate the considerable number of items included in the analyses and the need to examine DIF across univariate and multivariate groupings. In brief, alignment involves estimating the configural model separately in each group and then estimating factor means and variances across groups within each grouping such that DIF is minimized for that grouping. Alignment yields DIF-adjusted factor mean and variance estimates, as well as "aligned" item parameter estimates and standard errors, which can be used to test for DIF and judge its impact on item responses. We refer interested readers to Appendix S1, Mansolf et al. (2020), and Muthén and Asparouhov (2014) for technical details of the alignment method.

Statistical testing of measurement invariance/differential item functioning

We tested DIF for statistical significance using ANOVA, treating parameter estimates and standard errors analogously to sample statistics (mean and standard error, respectively) in conventional ANOVA, to screen for significance to examine the impact of DIF. For each grouping, an item was determined to have statistically significant DIF if any Bonferroni-corrected *p* value was less than 0.05, where the correction was applied for all item parameters. For multivariate groupings, two-way ANOVA was used, allowing DIF by each variable and the interaction between variables to be partitioned. To evaluate the differential impact of variables on MI/DIF, η^2 statistics were calculated for the main and interaction effect of DIF on each item, allowing comparisons of impact on the magnitude of DIF among the three sources. See Appendix S1 for additional details.

Assessing the impact of measurement invariance/differential item functioning

To steer away from reliance on significance testing, we used the unsigned item difference in the sample (UIDS; Meade, 2010) to quantify the impact of DIF on CBCL items and construct scores (i.e., *Internalizing, Externalizing, and Total Problems*). Unsigned Item Difference in the Sample is calculated by comparing expected item scores using model parameters estimated from different groups, and the magnitude of UIDS reflects the impact of differences in item parameter estimates on scores and can be interpreted on the scale of raw item scores. Using a threshold of UIDS >0.1, representing one-tenth of a point of score difference on the item, we identified items with the most impactful DIF for each construct, operationalized as those with UIDS >0.1 across more than one univariate grouping for *Internalizing* and *Externalizing*, and more than two for *Total Problems*. Thus, removing these items yielded a more measurement-invariant (i.e., robust) item set for each construct. Moreover, to assess the aggregated impact of item-level DIF, Signed Test Difference in the Sample (STDS) was calculated to represent differential test functioning. See Appendix S1 for additional details on UIDS and STDS calculation.

The effect of removing these items on reliability was assessed by plotting IRT reliability of measurement, calculated as one minus the reciprocal of test information, as a function of the latent trait, comparing values from the full item set to the robust item set.

Linking of robust and full item sets

For each domain, we used equipercentile equating (Kolen & Brennan, 2004), implemented in the *equate* package in R (Albano, 2016) to derive crosswalk tables which can be used to link total scores from the robust item sets to the total score metric of the original full item set. Then, the linked total scores can be used to derive the corresponding T scores using conversion tables available from the CBCL scoring manual. Correlations and mean differences between the scores from the full (i.e., originally published scale) and the linked scores from the robust item sets were computed to assess the reliability and bias, respectively, of the resulting linkages. Lastly, we conducted group comparisons of T scores based on the robust item sets across levels of each child and caregiver characteristic to examine whether substantive group differences remain after removing items with impactful measurement bias.

RESULTS

Configural invariance of CBCL/1.5–5

Both AIC and SABIC indicated that the bifactor model fit better than the unidimensional model across all subsamples for *Total Problems*. We, therefore, used the bifactor model parameters for DIF testing. Configural invariance was tested across all groupings of interest. Figure 1 shows the distribution of SRMSR and RMSEA values across groupings for each latent construct, estimated using the unidimensional model for *Internalizing* and *Externalizing* and the bifactor model for *Total Problems*. While most models across groupings and constructs showed adequate fit, model fit was better on average for *Externalizing* (mean SRMSR = 0.054) than for *Internalizing* (0.065) or *Total Problems* (0.066). Full sets of fit statistics and deviance can be found in the Open Science Framework (OSF) Repository: https://tinyurl.com/CBCLosffiles, due to space limit. Syndrome Scale analyses results are presented in the Appendix S2 and Supplementary Tables S7 to S11 and Figures S3, S4, and S6 to S11 for interested readers.

Significance and meaningfulness of item-level differential item functioning

Given the large number of results generated and the limited manuscript space, below we focus on the main takeaways of our

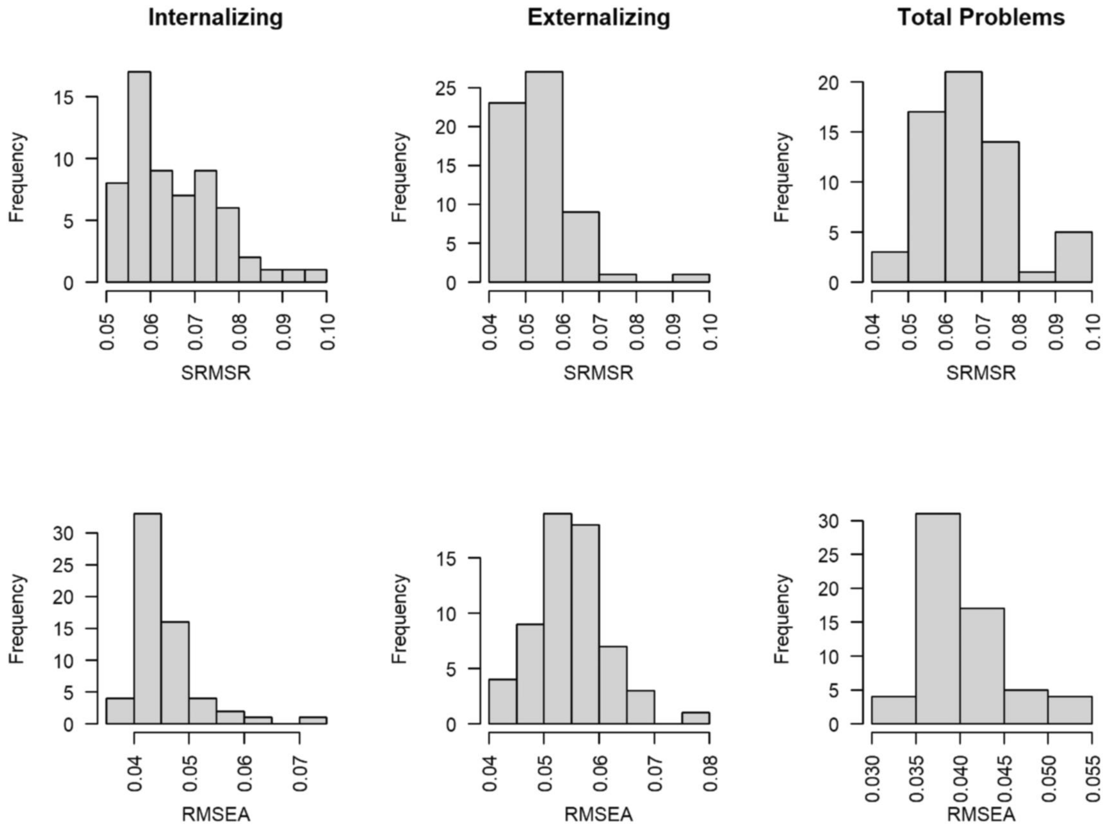

FIGURE 1 Model fit distributions for factor analysis models across three latent constructs. SRMSR, Standardized Root Mean Squared Residual; RMSEA, Root Mean Squared Error of Approximation.

TABLE 2 Items without any significant differential item functioning (DIF) across domains and characteristics.

Latent construct	Characteristics	Item	CBCL subdomain
Externalizing	Child	27 lacks guilt	Aggressive behavior
		53 attacks people	Aggressive behavior
Internalizing	Child	07 Can't stand things out of place	Somatic complaints
		70 little affection	Withdrawn
		86 too concerned with neatness	Somatic complaints
		51 panics	Emotionally reactive
		62 refuses active games	Withdrawn
		93 vomits	Somatic complaints
		98 withdrawn	Withdrawn
Total problems	Child	41 holds breath	Other problems
		62 refuses active games	Withdrawn
		93 vomits	Somatic complaints

findings, summarizing alignment results with respect to specific assessments of DIF in the CBCL 1.5–5. Detailed results of item parameter estimates, significance, effect sizes (η^2) of DIF, and UIDS statistics across all groupings and constructs can be found in the OSF repository.

Across the child and caregiver groupings we tested, we observed significant DIF in most of the items across groupings. Only a small set of items showed no DIF across child characteristics tested (see Table 2), while all items showed DIF across caregiver characteristics tested. Because this number of items was so small and not representative of the breadth of the full CBCL 1.5/5, we instead sought to identify which items had the *most* DIF and, when removed, would yield a robust item set with scores comparable across diverse samples.

To better understand the meaningfulness of DIF in the CBCL 1.5/5, Figure 2 presents items and groupings in which DIF was found to be significant and with UIDS greater than 0.1. For *Internalizing*, most of the impactful DIF (i.e., UIDS >0.1) was aggregated on items 10 *too dependent*, 33 *feelings hurt*, 37 *upset by separation*, 68 *self-conscious*, and 97 *whining*. For *Externalizing*, impactful DIF was observed on

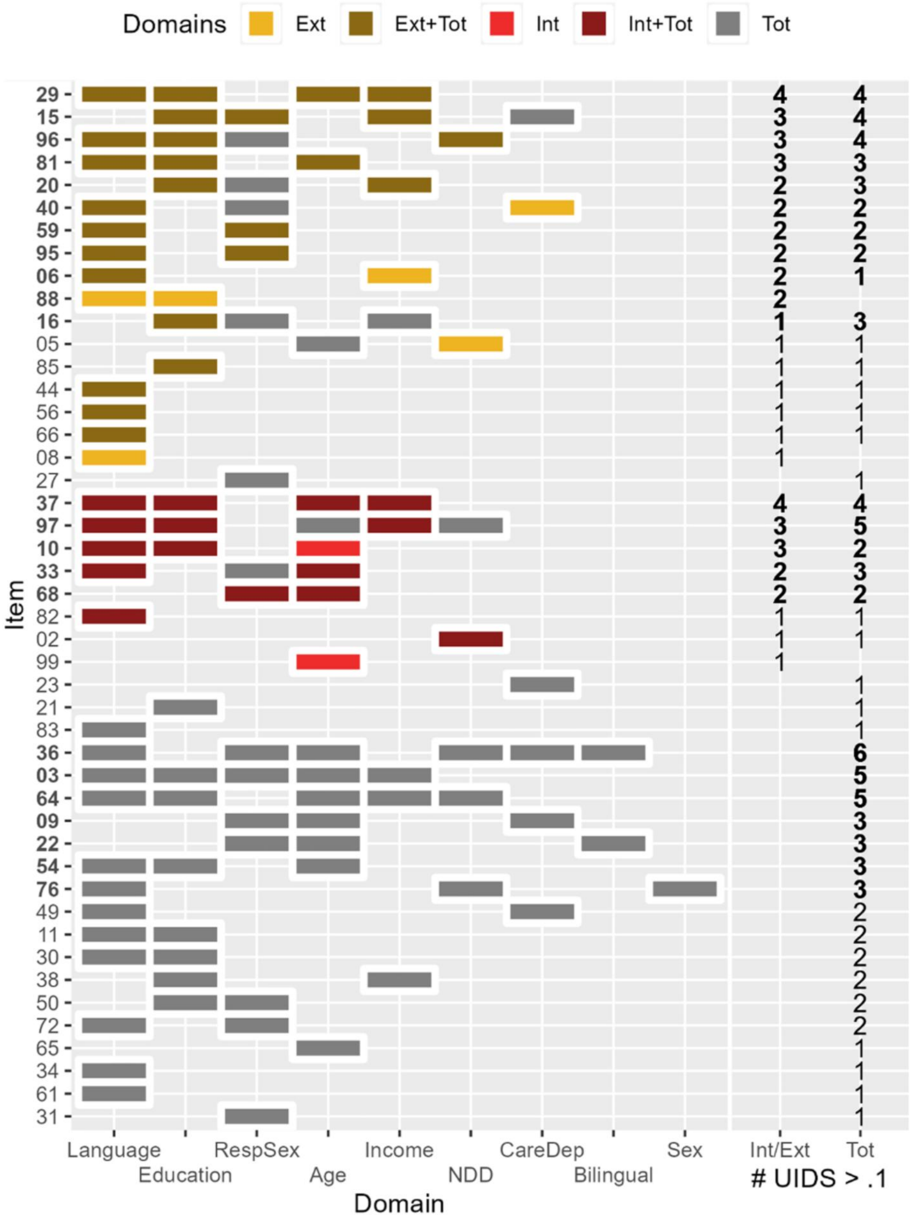

FIGURE 2 Items with Unsigned Item Difference in the Sample (UIDS) > 0.1 in Univariate Groupings across Latent Constructs. *Ext*, Externalizing domain; *Ext + Tot*, Externlizing and Total Problems; *Int*, Internalizing domain; *Int + Tot*, Internalizing and Total Problems; *Tot*, Total Problems; *UIDS*, Unsigned Item Difference in the Sample; *RespSex*, Respondent Sex; *NDD*, Neurodevelopmental Disorder. Items with no significant differential item functioning (DIF) or no UIDS >0.1 are excluded from the figure. Items are grouped by domains and sorted in decreasing order of the number of groupings with UIDS >0.1 of items; groupings are sorted by the number of items with UIDS >0.1 within the grouping. The numbers on the two right-hand columns show the number of groupings with UIDS >0.1 for the specific items within domains (Internalizing or Externalizing, and Total Problems).

TABLE 3 CBCL/1.5-5 robust item sets with less differential item functioning for internalizing, externalizing, and total problems.

Internalizing	Externalizing	Total problems
01, 02, 04, 07, 12, 19, 21, 23, 24, 39, 43, 45, 46, 47, 51, 52, 62, 67, 70, 71, 78, 79, 82, 83, 86, 87, 90, 92, 93, 98, 99	05, 08, 16, 18, 27, 35, 42, 44, 53, 56, 58, 66, 69, 85	01, 02, 04, 05, 06, 07, 08, 10, 11, 12, 13, 14, 17, 18, 19, 21, 23, 24, 25, 26, 27, 28, 30, 31, 32, 34, 35, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 55, 56, 57, 58, 59, 60, 61, 62, 63, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 77, 78, 79, 80, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 98, 99

items 06 *can't sit still*, 15 *defiant*, 20 *Disobedient*, 29 *frustrated*, 40 *Hits Others*, 59 *quickly shifts*, 81 *stubborn*, 88 *uncooperative*, 95 *wanders away*, and 96 *wants attention*. For *Total Problems*, impactful DIF was observed across 16 items (03, 09, 15, 16, 20, 22, 29, 33, 36, 37, 54, 64, 76, 81, 96, 97) showing UIDS >0.1 on at least three out of nine univariate groupings tested. Taken together, the impact of DIF was

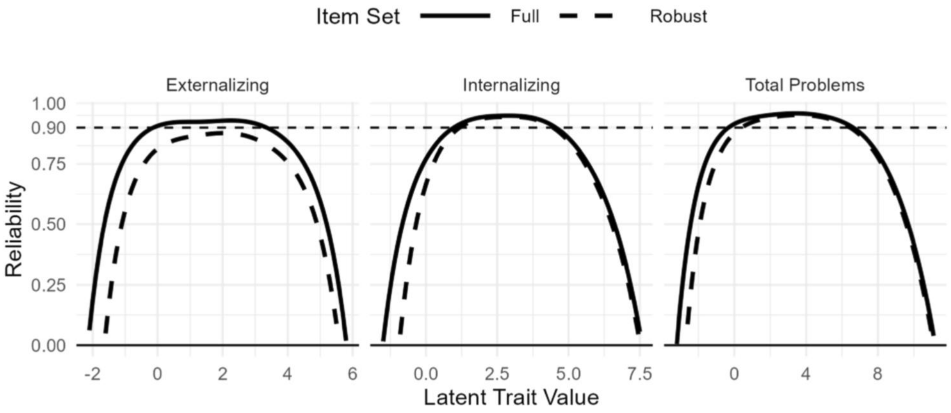

FIGURE 3 Item response theory (IRT) Reliability for Full and Robust Item Sets. *Full*, full CBCL item set; *Robust*, item set with high-DIF items removed.

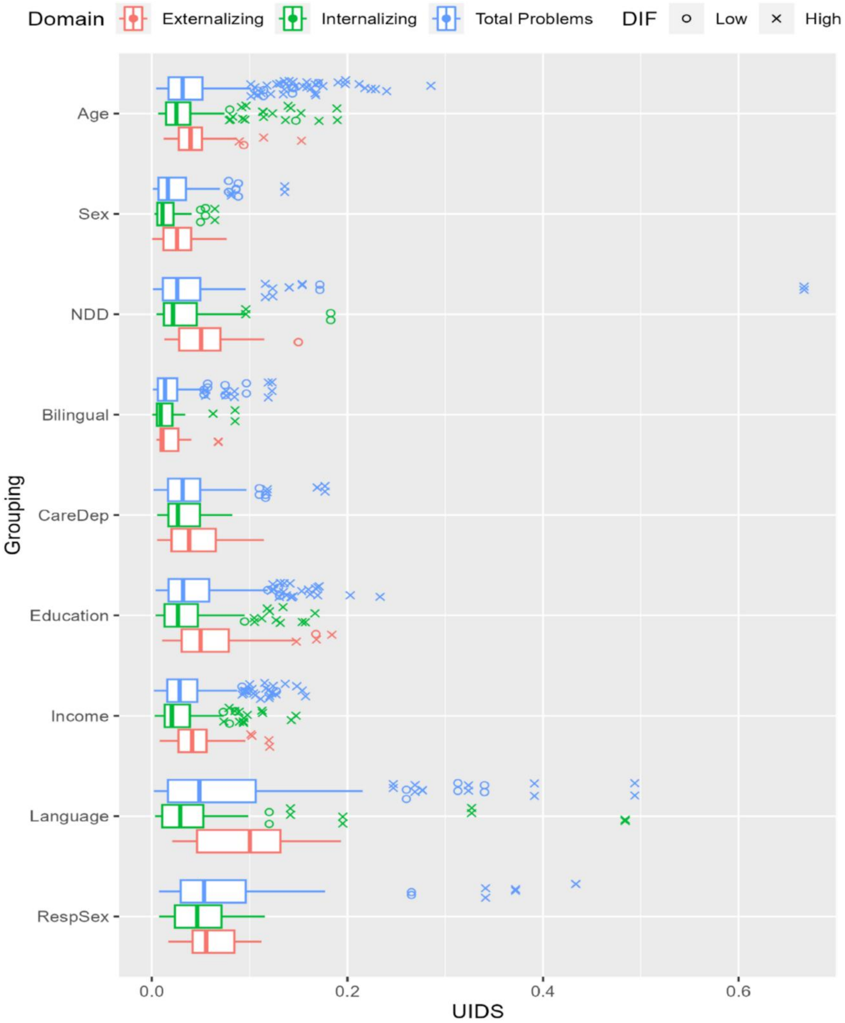

FIGURE 4 Distribution of Unsigned Item Difference in the Sample (UIDS) by Grouping and Latent Constructs. *NDD*, Any Neurodevelopment Disorders; *CareDep*, Caregiver Depression with a cutoff of T-score ≥ 60 for clinical range; *RespSex*, Respondent Sex.

higher for *Externalizing* than *Internalizing* and was the highest for *Total Problems*. Distributions of median UIDS of all items can be found in Figure S1.

Table 3 contains the robust item sets for each construct (i.e., with high DIF items removed): *Internalizing* (n_{item} = 31), *Externalizing* (n_{item} = 14), and *Total Problems* (n_{item} = 83). Reliabilities for the full and the robust items were all near or above 0.9 and in nearly identical ranges for Internalizing and Total Problems (see Figure 3 with the solid and dotted lines closely overlapping).

To understand the impact of different groupings, Figure 4 presents the distribution of UIDS by univariate groupings and domains for items with significant DIF. Notably, the impactful DIF across items concentrated on language version, caregiver education level, caregiver sex, and child age. The largest sources of DIF for caregiver

TABLE 4 Median and maximum effect sizes (eta-squared) of main and interaction effects on differential item functioning (DIF).

		Internalizing		Externalizing		Total problems	
	Term	Median	Max	Median	Max	Median	Max
Caregiver education-income	Education	0.091	0.672	0.072	0.680	0.093	0.648
	Income	0.164	0.817	0.267	0.699	0.208	0.845
	Education: Income	0.140	0.623	0.089	0.517	0.140	0.629
Child sex-age	Age	0.286	0.891	0.334	0.858	0.377	0.970
	Sex	0.035	0.660	0.037	0.489	0.035	0.736
	Sex: Age	0.130	0.397	0.126	0.436	0.138	0.421
Child Sex-NDD	Sex	0.073	0.648	0.108	0.557	0.092	0.784
	NDD	0.261	0.902	0.145	0.864	0.175	0.979
	Sex: NDD	0.053	0.317	0.062	0.546	0.061	0.552

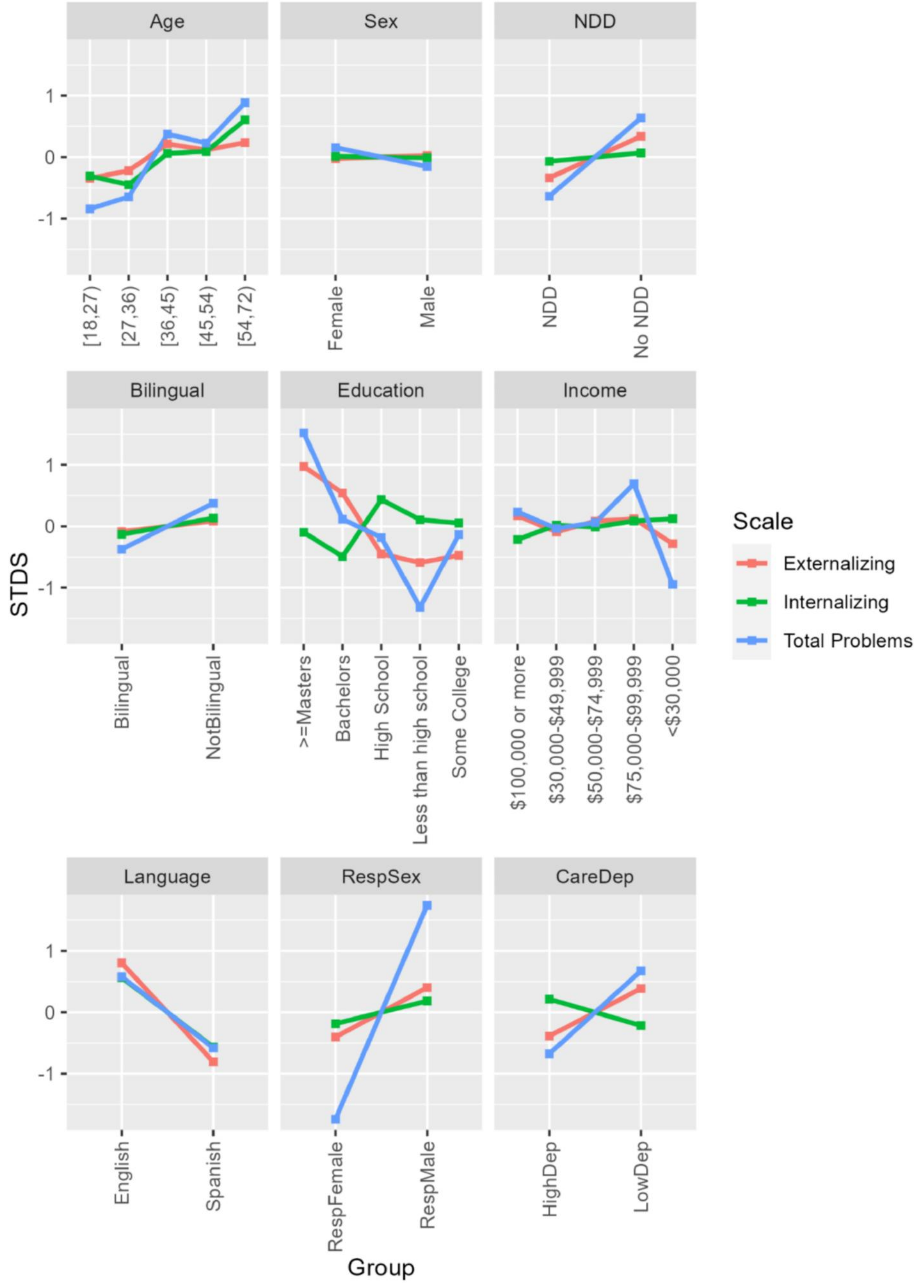

FIGURE 5 Signed Test Difference in the Sample (STDS) for Univariate Groupings. *RespSex*, Respondent Sex; *RespFemale*, Female Respondent; *RespMale*, Male Respondent; *CareDep*, Caregiver Depression with clinical threshold of $T = 60$; *HighDep*, Depression T score greater than or equal to 60; *LowDep*, Depression T score less than 60; *NDD*, Neurodevelopmental Disorder.

characteristics were language version of CBCL administration, followed by caregiver sex and education. The largest source of DIF for child characteristics was child age. Effect sizes of the multivariate groupings on DIF (Table 4) showed main effects of the groupings were associated with larger magnitudes of DIF than the interactions. Notably, child age and NDD status were associated with larger DIF than child sex or their interactions with child sex. One outlier item (Item 76, *speech problem*) showed the largest magnitudes of DIF between those with and without NDD. Household income level showed larger magnitudes of DIF effect sizes than caregiver education level, especially for *Externalizing Problems*. For differential test functioning, respondent sex showed the largest STDS on the measurement of *Total Problems*, followed by caregiver education levels on *Total Problems* (see Figure 5).

Sensitivity analyses with child race groupings showed impactful DIF across race categories on multiple items on each latent construct, with impactful DIF aggregated on similar items as identified above. While child race, caregiver education, and income all showed unique associations with DIF, child race was associated with larger magnitudes of DIF than the SES variables after adjusting for each other and the interaction, indicating that measurement bias across racial groups cannot be fully accounted for by SES variables (See Table S2). Similarly, the multivariate analyses with language version revealed that language version was associated with more DIF than child bilingual status but with less DIF than caregiver education level (See Table S3 for effect sizes, Figure S2 for Model Fit, and Table S6 for items without significant DIF and Figure S5 for items with impactful DIF).

T scores based on the robust item sets and tests of group differences

Correlations between linked scores (i.e., linking summed scores for the robust item sets to the metric of the full CBCL) and the observed summed scores of the full item sets, were 0.963 for *Internalizing*, 0.965 for *Externalizing*, and 0.989 for *Total Problems*. The raw mean differences between linked and observed scores, an indicator of linking bias, were −0.004, 0.008, and 0.007 for *Total Problems*, *Internalizing*, and *Externalizing*, respectively, indicating very minimal bias (See Table S4). Group means and standard deviations of T scores based on robust item sets, and ANOVA tests of mean differences in each grouping, are presented in Table 5 (group comparisons of the full item set are shown in Table S5). We observed the largest group differences in T scores based on robust item sets across Caregiver Education levels ($\eta^2 \geq 0.023$) on all three latent constructs, followed by income ($\eta^2 \geq 0.020$) and Caregiver Depression ($\eta^2 \geq 0.018$), with the smallest differences in *Internalizing* and the largest differences in *Total Problems*. For child characteristics, Child NDD status ($\eta^2 \geq 0.021$) showed small to medium effects, with the smallest differences in *Externalizing*.

DISCUSSION

The CBCL/1.5–5 has been widely applied in clinical practice and research, but little information is available on its measurement equivalence across subgroups defined by various child and caregiver characteristics. We conducted the largest study of MI/DIF of CBCL/1.5–5 (N = 9087) to date and identified multiple sources of measurement bias across child and caregiver characteristics. Our findings directly inform the use of CBCL/1.5–5 and have implications for the measurement of childhood behavior problems more broadly.

Results of the factor analyses confirmed the unidimensional structure of *Internalizing* and *Externalizing* broad domains and the bifactor structure of *Total Problems* across child and caregiver groupings, demonstrating that both symptom clusters and the general psychopathology ("*p*") factor account for distinct sources of variance in ratings of child psychopathology. These findings are consistent with prior conceptualizations of the CBCL/1.5–5 (Achenbach & Rescorla, 2001) and previous studies in different samples (Achenbach et al., 2016), providing further evidence for the utility of broad-domain constructs.

Item-level analyses identified multiple sources of DIF, raising concerns about the ubiquitous application of CBCL/1.5-5 in diverse samples without adjusting for measurement bias. Caregiver demographic variables were associated with a larger magnitude of measurement bias than child characteristics across all three constructs as indicated by UIDS and STDS, with the greatest DIF/Differential Test Functioning (DTF) arising from language version, caregiver education, and respondent sex. Among them, language version showed the largest impact of MI/DIF across all constructs. It is possible that both the information loss due to translation and differences in cultural expectations/interpretations of child behavior problems contribute to the noninvariance of the English and Spanish versions of the CBCL/1.5–5. For example, as discussed by Gross et al., 2006, Item 97 *whining* is translated to *queja* in the Spanish version, which means *complain* in English and may have different connotations for parents. This underscores the need to consider validation efforts separately from simple translation of measurement tools in order to ensure equivalence across language translations.

We further observed an impact of respondent sex and caregiver education on DIF. It is possible that parents' ethnotheories of desirable and maladaptive behaviors in children (Olson et al., 2019), and their expectations about child development, influence how parents perceive and rate their child's behavior. Hence, in line with recommendations to consider how caregiver informants' background might influence the assessment of child psychopathology (De Los Reyes & Kazdin, 2005), these findings provide empirical evidence of the potential impact of caregiver characteristics on the measurement of specific child behavior problems. Notably, we did not observe much impactful DIF related to caregiver depression, despite the depression-distortion hypothesis suggesting that depressed caregivers perceive more problems in their children (De Los Reyes & Kazdin, 2005). Since only a small proportion of caregivers met the clinical cutoff of T-score ≥60 (3.4%), our analyses might have lacked the power to detect possible bias across caregiver depression status. Nevertheless, we found small-to-moderate differences in *Internalizing, Externalizing, and Total Problems* across caregiver depression groups using T-scores derived from robust item sets, indicating that observed differences across caregiver depression groups were not attributable to item bias (Chi & Hinshaw, 2002; Gartstein et al., 2009).

Group comparison of T scores using robust item sets showed that differences between groups of language version and respondent sex were close to zero, while the original scores of *Internalizing* differed significantly across language version and respondent sex (see Table S5). These findings further underscore the need to assess

TABLE 5 Group comparisons of T-scores of robust item set across child and caregiver characteristics across domains.

Grouping variable	Group	N	Externalizing	Internalizing	Total problems
Child characteristics					
Age (months)			F(4,7380) = 21.918; p = 0.000, η^2 = .012	F(4,7380) = 18.866; p = 0.000, η^2 = .010	F(4,7380) = 21.329; p = 0.000, η^2 = .011
	[18,27)	2241	45.18 (9.54)	42.65 (9.55)	44.00 (9.57)
	[27,36)	1328	46.46 (10.15)	45.20 (10.36)	46.28 (10.26)
	[36,45)	1155	45.96 (10.58)	45.16 (10.49)	45.68 (10.77)
	[45,54)	983	44.14 (10.30)	43.88 (10.15)	43.69 (10.13)
	[54,72)	1678	43.43 (10.07)	44.36 (10.45)	43.48 (10.31)
Child sex			F(1,7381) = 41.844; p = 0.000, η^2 = .006	F(1,7381) = 12.967; p = 0.000, η^2 = .002	F(1,7381) = 29.021; p = 0.000, η^2 = .004
	Female	3527	44.21 (9.97)	43.61 (10.15)	43.85 (10.13)
	Male	3856	45.72 (10.16)	44.46 (10.20)	45.12 (10.21)
Bilingual			F(1,6507) = 1.951; p = 0.162, $\eta^2 < .001$	F(1,6507) = 0.352; p = 0.553, $\eta^2 < 0.001$	F(1,6507) = 0.003; p = 0.954, $\eta^2 < 0.001$
	Bilingual	2345	44.75 (10.27)	44.02 (10.22)	44.45 (10.33)
	Not bilingual	4164	45.12 (10.06)	43.87 (10.18)	44.46 (10.15)
NDD			F(1,4491) = 95.068; p = 0.000, η^2 = .021	F(1,4491) = 126.232; p = 0.000, η^2 = .027	F(1,4491) = 138.176; p = 0.000, η^2 = .030
	NDD	552	49.26 (12.42)	49.07 (12.26)	49.65 (12.48)
	No NDD	3941	44.76 (9.79)	43.91 (9.77)	44.25 (9.74)
Caregiver characteristics					
Caregiver depression			F(1,4038) = 136.745; p = 0.000, η^2 = .033	F(1,4038) = 75.894; p = 0.000, η^2 = .018	F(1,4038) = 143.201; p = 0.000, η^2 = .034
	HighDep	247	52.75 (11.78)	49.85 (11.56)	52.49 (11.46)
	LowDep	3793	44.91 (10.09)	43.97 (10.19)	44.40 (10.21)
Income			F(4,6148) = 41.263; p = 0.000, η^2 = .026	F(4,6148) = 30.816; p = 0.000, η^2 = .020	F(4,6148) = 67.041; p = 0.000, η^2 = .042
	<$30,000	1583	42.97 (9.16)	42.33 (9.14)	41.90 (8.90)
	$30,000-$49,999	681	46.17 (10.67)	44.62 (10.68)	45.63 (10.77)
	$50,000-$74,999	838	45.27 (9.73)	43.98 (9.86)	44.69 (9.81)
	$75,000-$99,999	602	44.34 (9.60)	42.89 (9.68)	43.54 (9.50)
	$100,000 or more	2301	46.86 (11.00)	45.76 (11.06)	46.94 (11.18)
Language version			F(1,6956) = 18.961; p = 0.000, η^2 = .003	F(1,6956) = 17.236; p = 0.000, η^2 = .002	F(1,6956) = 47.130; p = 0.000, η^2 = .007
	English	6193	44.68 (10.07)	43.79 (10.05)	44.08 (10.12)
	Spanish	765	46.36 (10.10)	45.40 (10.89)	46.75 (10.37)
Maternal education			F(4,7227) = 63.130; p = 0.000, η^2 = .034	F(4,7227) = 42.248; p = 0.000, η^2 = .023	F(4,7227) = 93.252; p = 0.000, η^2 = .049
	<high school	545	43.92 (9.48)	43.12 (9.74)	43.11 (9.50)
	High school	1451	46.51 (10.44)	45.08 (10.57)	46.33 (10.61)
	Some college	1741	48.08 (11.19)	47.38 (11.46)	48.76 (11.50)

TABLE 5 (Continued)

Grouping variable	Group	N	Externalizing	Internalizing	Total problems
	Bachelor's degree	1762	42.45 (8.80)	42.07 (9.20)	41.53 (8.61)
	>=Master's degree	1676	46.38 (10.57)	44.95 (10.30)	45.98 (10.49)
Respondent sex			$F(1,7383) = 2.865$; $p = 0.091$, $\eta^2 < 0.001$	$F(1,7383) = 0.108$; $p = 0.743$, $\eta^2 < 0.001$	$F(1,7383) = 1.342$; $p = 0.247$, $\eta^2 < 0.001$
	Female	7061	44.96 (10.15)	44.06 (10.24)	44.48 (10.25)
	Male	324	45.93 (8.85)	43.87 (8.88)	45.15 (8.79)

and account for measurement bias resulting from informant characteristics when interpreting results from caregiver-report measures. For example, father-mother discrepancies in perceived levels of child behavior problems have been reported by many previous studies but were not significant in our sample after accounting for measurement bias. This is likely because the observed discrepancies are at least partially due to differences in how caregivers report specific items. Thus, while DIF is not always problematic and, in fact, often informative for measure development and application by showing group-specific patterns of item responding, accounting for DIF across these factors is necessary for teasing apart measurement bias from true differences in the construct of interests (e.g., behavior problems) between groups.

Child age showed a larger impact on DIF than child sex, indicating the importance of developmental considerations when measuring psychopathology even within the relatively narrow age range of early childhood (18–71 months) (Wakschlag et al., 2010). These findings suggest that some child behavior problems might not be informative in caregiver-report of psychopathology for children of different ages. Therefore, normative ranges of child behaviors should be clarified by research, and care should be taken when administering identical item sets across age groups. Moreover, our sensitivity analyses showed that the DIF associated with child race cannot be fully accounted for by SES variables (caregiver education and income), indicating a need for further investigation of potential measurement bias associated with race (e.g., race-related cultural differences, trauma experiences).

In terms of measuring *Internalizing problems*, items with the most measurement bias were from the *Anxious/Depressed* and *Emotionally Reactive* subscales. Studies on anxiety and depression in young children have repeatedly emphasized the challenge of generalizing adult diagnostic criteria to young children (Carter et al., 2004; Luby et al., 2002; Tandon et al., 2009), and relying heavily on caregivers to infer the internal states (e.g., feelings) of young children. As for *Externalizing*, almost half of the items showed impactful DIF (10 out of 24), with three out of five items on the *Attention Problems* scale showing high DIF. Removing a large amount of high-DIF items to construct the robust item set might result in changes in the construct validity of the *Externalizing* scale. *Externalizing* items with the most DIF are adjectives that may imply character judgment (e.g., *defiant, stubborn*). Taken together, many items with impactful DIF lack clear descriptions of observable behaviors, which might be more susceptible to bias driven by child and caregiver characteristics. Thus, measurement of early behavior problems may be enhanced by carefully operationalizing specific behaviors relevant to constructs under investigation to reduce the risk of bias (Merrell, 2001).

Implications for using the CBCL/1.5–5

Given the large amount of DIF identified in our analyses of CBCL/1.5 items, it is important to consider the effects of measurement bias when using the CBCL/1.5–5. Practically, DIF could be handled in two ways: (a) by applying group-specific parameters for score calculation, or (b) by removing items with impactful DIF. Given the technical requirements to apply group-specific algorithms, we recommend CBCL/1.5–5 consumers follow the below guideline to administer the robust item sets with reduced DIF: (1) administer robust item sets listed in Table 3; (2) calculate summed scores for the desired domain; (3) use the crosswalk table on OSF repository for this study to convert robust item set sum scores to approximate full item set sum scores; (4) convert sum scores to T scores using CBCL score conversion tables. Researchers and clinicians can then apply the recommended cut-offs based on the CBCL manual: 60–63 as borderline range, and above 63 as the clinical range. Changes in construct validity notwithstanding, the resulting T scores are expected to be comparable to the T score from the original CBCL but with fewer biases across groups that differ on the child and caregiver characteristics detailed above.

Limitations and strengths

We only highlighted the main takeaways of our study and did not present all results in full detail in the manuscript. Decisions on how to winnow these results were made with attention to presentability while maintaining statistical rigor and transparency. First, although different decisions, such as the threshold of "impactful" DIF, could have been made, we believe the core conclusions would remain the same. We share item-level parameter estimates and standard errors via OSF Repository for interested readers to conduct their own queries. Second, we had chosen to focus on the broad domains of *Internalizing*, *Externalizing*, and *Total Problems*. Thus, findings should not be generalized to the syndrome scales or researcher-defined item subsets (MI/DIF analyses for syndrome scales are included in Appendix S2, given the emphasis here on broadband domains). Additionally, the current analysis could not distinguish between uniform DIF (i.e., DIF affecting all items in the same way) and differences in latent mean and variance. Future work including multiple raters can help clarify this distinction. Finally, the current analysis did not examine item discrimination and severity based on the criteria of clinical concerns due to the lack of such information, so robust item sets are not optimized for identifying clinical-range problems.

Despite the abovementioned limitations, our study has many methodological strengths, including an unprecedently large and diverse sample from across the United States, with high representation across race, language, caregiver education, and income. Furthermore, the large sample size allowed us to perform analyses with many subdivisions of the data, yielding a highly multifaceted assessment of MI/DIF. The multivariate examinations also revealed the differential impact of certain variables in the presence of possible confounders. Relatedly, the application of the alignment approaches to MI/DIF allowed the efficient investigations of large numbers of groups and items while accommodating mismatches between item sets and observed item categories across groups.

Conclusions

The CBCL/1.5–5 is a commonly used measure of early childhood behavior problems and risk for psychopathology. This study systematically applied a rigorous method to examine MI/DIF in a large, diverse sample and found measurement bias related to language version, caregiver education level and sex, and child age. Our identification of robust item sets with the least DIF that offer similar levels of information as the full item sets and could be applied across diverse samples to reduce measurement bias of the CBCL/1.5–5. Future work in child assessment should carefully consider the bias of caregiver-report measures and devise methods to account for measurement bias when possible.

AUTHOR CONTRIBUTIONS

Shuting Zheng: Conceptualization; Investigation; Methodology; Project administration; Resources; Supervision; Writing – original draft; Writing – review & editing. **Maxwell Mansolf**: Formal analysis; Methodology; Visualization; Writing – original draft; Writing – review & editing. **Monica McGrath**: Data curation; Supervision; Writing – review & editing. **Marie L. Churchill**: Data curation; Formal analysis; Project administration; Writing – review & editing. **Traci A. Bekelman**: Data curation; Funding acquisition; Writing – review & editing. **Patricia A. Brennan**: Data curation; Funding acquisition; Writing – review & editing. **Amy E. Margolis**: Data curation; Funding acquisition; Writing – review & editing. **Sara S. Nozadi**: Data curation; Writing – review & editing. **Theresa M. Bastain**: Data curation; Writing – review & editing. **Amy J. Elliott**: Data curation; Funding acquisition; Writing – review & editing. **Kaja Z. LeWinn**: Data curation; Funding acquisition; Writing – review & editing. **Julie A. Hofheimer**: Data curation; Funding acquisition; Writing – review & editing. **Leslie D. Leve**: Data curation; Funding acquisition; Writing – review & editing. **Brandon Rennie**: Data curation; Writing – review & editing. **Emily Zimmerman**: Data curation; Writing – review & editing. **Carmen A. Marable**: Data curation; Writing – review & editing. **Cindy T. McEvoy**: Data curation; Funding acquisition; Writing – review & editing. **Chang Liu**: Data curation; Writing – review & editing. **Alexis Sullivan**: Data curation; Writing – review & editing. **Tracey J. Woodruff**: Data curation; Funding acquisition; Writing – review & editing. **Samiran Ghosh**: Data curation; Funding acquisition; Writing – review & editing. **Bennett Leventhal**: Data curation; Funding acquisition; Writing – review & editing. **Assiamira Ferrara**: Data curation; Funding acquisition; Writing – review & editing. **Johnnye Lewis**: Data curation; Funding acquisition; Writing – review & editing. **Somer Bishop**: Conceptualization; Investigation; Supervision; Writing – review & editing.

ACKNOWLEDGMENTS

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Research reported in this publication was supported by the ECHO program, Office of The Director, NIH, under Award Numbers U2COD023375 (Coordinating Center), U24OD023382 (Data Analysis Center), U24OD023319 with co-funding from the Office of Behavioral and Social Science Research (PRO Core), UH3 OD023271 (Karr), UH3 OD023305 (Trasande), UH3 OD023289 (Ferrara), UH3 OD023347 (Lester), UH3 OD023389 (Leve), UH3 OD023248 (Dabelea), UH3 OD023337 (Wright), UH3 OD023272 (Schantz), UH3 OD023328 (Duarte), UH3 OD023318 (Dunlop), UH3 OD023251 (Alshawabkeh), UH3 OD023279 (Elliott), UH3 OD023285 (Kerver), UH3 OD023349 (O'Connor), UH3 OD023287 (Breton), UH3 OD023288 (McEvoy), UH3 OD023320 (Aschner), UH3 OD023313 (Koinis-Mitchell), UH3 OD023249 (Stanford), UH3 OD023290 (Herbstman), UH3OD023344 (MacKenzie & Lewis), UH3OD023348 (O'Shea).

The authors wish to thank our ECHO colleagues; the medical, nursing, and program staff; and the children and families participating in the ECHO cohorts. We also acknowledge the contribution of the following ECHO program collaborators:

Environmental influences on Child Health Outcomes Components—Coordinating Center: Duke Clinical Research Institute, Durham, North Carolina: Smith PB, Newby KL, Data Analysis Center: Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland: Jacobson LP; Research Triangle Institute, Durham, North Carolina: Catellier DJ; Person-Reported Outcomes Core: Northwestern University, Evanston, Illinois: Gershon R, Cella D.

Environmental influences on Child Health Outcomes Awardees and Cohorts— Northeastern University, Boston, Massachusetts: Alshawabkeh, AN; Icahn School of Medicine at Mount Sinai, New York, NY: Teitelbaum SL; Cohen Children's Medical Center, Northwell Health: Stroustrup A; Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio: Merhar S; Children's Hospital and Clinic Minnesota, Minneapolis, MN: Lampland A; University of Buffalo, Buffalo, NY: Reynolds A; University of Florida, College of Medicine, Jacksonville, FL: Hudak M; University of Rochester Medical Center, Rochester, NY: Pryhuber G; Vanderbilt Children's Hospital, Nashville, TN: Moore P; Wake Forest University School of Medicine, Winston Salem, NC: Washburn L; University of Colorado Denver, Denver, CO: Dabelea D; Rhode Island Hospital, Providence RI: Koinis-Mitchell D; University of Puerto Rico, San Jaun, PR: Canino G; Emory University, Atlanta, GA: Dunlop A; Avera Health Rapid City, Rapid City, SD: Elliott A; Kaiser Permanente Northern California Division of Research, Oakland, CA: Croen L; University of Southern California, Los Angeles, CA: Bastain T, Farzan S, Habre R; University of Washington, Department of Environmental and Occupational Health Sciences, Seattle, WA: Karr C; University of Tennessee Health Science Center, Memphis, TN: Mason A; Seattle Children's Research Institute, Seattle, WA: Sathyanarayana S; Children's Mercy, Kansas City, MO: Carter B; Emory University, Atlanta, GA: Marsit C; Helen DeVos Children's Hospital, Grand Rapids, MI: Pastyrnak S; Kapiolani Medical

Center for Women and Children, Providence, RI: Neal C; Los Angeles Biomedical Research Institute at Harbour-UCLA Medical Center, Los Angeles CA: Smith L; Wake Forest University School of Medicine, Winston Salem, NC: Helderman J; George Washington University, Washington, DC: Ganiban J; Pennsylvania State University, University Park, PA: Neiderhiser J; University of New Mexico Community Environmental Health Program, Albuquerque, NM: MacKenzie D & Lewis J; Indiana University, Riley Hospital for Children: Indianapolis, IN, Tepper R; University of Rochester Medical Center Rochester, NY: O'Connor T; University of Pittsburgh Medical Center, Magee Women's Hospital, Pittsburgh, PA: Simhan H; Michigan State University, East Lansing, MI: Kerver J; Henry Ford Health System, Detroit, MI: Barone, C; Michigan Department of Health and Human Services, Lansing, MI: McKane, P; Michigan State University, East Lansing, MI: Paneth N; University of Michigan, Ann Arbor, MI: Elliott, M; Columbia University Medical Center, New York, NY: Herbstman J; University of Utah, Salt Lake City, UT: Stanford J; New York School of Medicine, New York, NY: Trasande L; Icahn School of Medicine at Mount Sinai, New York, NY: Wright R; Boston Children's Hospital, Boston MA: Bosquet-Enlow M.

CONFLICT OF INTEREST STATEMENT

The authors have declared that they have no competing or potential conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from NIH ECHO program to ECHO-affiliated researchers.

ETHICAL CONSIDERATIONS

Study protocols of individual cohorts in the ECHO program were reviewed and approved by site-specific institutional review boards. Participants provided informed consent for themselves and their children to be in the studies and have their data be shared for research use in the ECHO program.

ORCID

Shuting Zheng https://orcid.org/0000-0002-0821-2716

REFERENCES

Achenbach, T. M., Ivanova, M. Y., Rescorla, L. A., Turner, L. V., & Althoff, R. R. (2016). Internalizing/Externalizing problems: Review and recommendations for clinical and research applications. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(8), 647–656. https://doi.org/10.1016/j.jaac.2016.05.012

Achenbach, T. M., & Rescorla, L. A. (2001). *Child behavior checklist 1.5 to 5*. University of Vermont, Research Center for Children, Youth, & Families.

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*(8), 1–36. https://doi.org/10.18637/jss.v074.i08

Bayer, J. K., Ukoumunne, O. C., Mathers, M., Wake, M., Abdi, N., & Hiscock, H. (2012). Development of children's internalising and externalising problems from infancy to five years of age. *Australian and New Zealand Journal of Psychiatry*, *46*(7), 659–668. SAGE Publications Ltd. https://doi.org/10.1177/0004867412450076

Blackwell, C. K., Tang, X., Elliott, A. J., Thomes, T., Louwagie, H., Gershon, R., Schalet, B. D., & Cella, D. (2021). (Vol. *33*, pp. 610–618). American Psychological Association. https://doi.org/10.1037/pas0001009. Developing a common metric for depression across adulthood: Linking PROMIS depression with the Edinburgh Postnatal Depression Scale*Psychological Assessment*7

Blaisdell, C. J., Park, C., Hanspal, M., Roary, M., Arteaga, S. S., Laessig, S., Luetkemeier, E., & Gillman, M. W. (2021). The NIH ECHO Program: Investigating how early environmental influences affect child health. *Pediatric Research*, *1–2*(5), 1215–1216. Nature Publishing Group. https://doi.org/10.1038/s41390-021-01574-8

Braveman, P. A., Cubbin, C., Egerter, S., Chideya, S., Marchi, K. S., Metzler, M., & Posner, S. (2005). Socioeconomic status in health research one size does not fit all. *JAMA*, *294*(22), 2879–2888. https://doi.org/10.1001/jama.294.22.2879

Briggs-gowan, M. J., Carter, A. S., Skuban, E. M., & Horwitz, S. M. (2001). Prevalence of social-emotional and behavioral problems in a community sample of 1- and 2-year-old children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(7), 811–819. https://doi.org/10.1097/00004583-200107000-00016

Campbell, S. B. (1995). Behavior problems in preschool children: A review of recent research. *Journal of Child Psychology and Psychiatry*, *36*(1), 113–149. https://doi.org/10.1111/j.1469-7610.1995.tb01657.x

Carneiro, A., Dias, P., & Soares, I. (2016). Risk factors for internalizing and externalizing problems in the preschool years: Systematic literature review based on the child behavior checklist 1½–5. *Journal of Child and Family Studies*, *25*(10), 2941–2953. https://doi.org/10.1007/s10826-016-0456-z

Carter, A. S., Briggs-Gowan, M. J., & Davis, N. O. (2004). Assessment of young children's social-emotional development and psychopathology: Recent advances and recommendations for practice. *Journal of Child Psychology and Psychiatry*, *45*(1), 109–134. https://doi.org/10.1046/j.0021-9630.2003.00316.x

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, J. J. (2010). Gender differences in externalising problems among preschool children: Implications for early childhood educators. *Early Child Development and Care*, *180*(4), 463–474. Routledge.

Cheng, T. L., Goodman, E., Cheng, T. L., Bogue, C. W., Chien, A. T., Dean, J. M., Kharbanda, A. B., Peeples, E. S., & Scheindlin, B. (2015). Race, ethnicity, and socioeconomic status in research on child health. *Pediatrics*, *135*(1), e225–e237. https://doi.org/10.1542/peds.2014-3109

Chi, T. C., & Hinshaw, S. P. (2002). Mother–child relationships of children with ADHD: The role of maternal depressive symptoms and depression-related distortions. *Journal of Abnormal Child Psychology*, *30*(4), 387–400. https://doi.org/10.1023/a:1015770025043

Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment*, *26*(2), 513–527. https://doi.org/10.1037/a0035768

Davis, A. N., & Qi, C. H. (2022). Associations between maternal depressive symptoms and child internalizing problems. *Early Child Development and Care*, *192*(5), 697–709. Routledge. https://doi.org/10.1080/03004430.2020.1794860

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483–509. American Psychological Association. https://doi.org/10.1037/0033-2909.131.4.483

Dovgan, K., Mazurek, M. O., & Hansen, J. (2019). Measurement invariance of the child behavior checklist in children with autism spectrum disorder with and without intellectual disability: Follow-up study. *Research in Autism Spectrum Disorders*, *58*, 19–29. https://doi.org/10.1016/j.rasd.2018.11.009

Egger, H. L., & Angold, A. (2006). Common emotional and behavioral disorders in preschool children: Presentation, nosology, and epidemiology. *Journal of Child Psychology and Psychiatry*, *47*(3–4), 313–337. https://doi.org/10.1111/j.1469-7610.2006.01618.x

Eisenberg, N., Valiente, C., Spinrad, T. L., Cumberland, A., Liew, J., Reiser, M., Zhou, Q., & Losoya, S. H. (2009). Longitudinal relations of children's effortful control, impulsivity, and negative emotionality to their externalizing, internalizing, and co-occurring behavior problems. *Developmental Psychology*, *45*(4), 988–1008. Washington, US: American Psychological Association. https://doi.org/10.1037/a0016213

Flouri, E., Ruddy, A., & Midouhas, E. (2017). Maternal depression and trajectories of child internalizing and externalizing problems: The roles of child decision making and working memory. *Psychological Medicine*, *47*(6), 1138–1148. Cambridge University Press. https://doi.org/10.1017/s0033291716003226

Gartstein, M. A., Bridgett, D. J., Dishion, T. J., & Kaufman, N. K. (2009). Depressed mood and maternal report of child behavior problems: Another look at the depression–distortion hypothesis. *Journal of Applied Developmental Psychology*, *30*(2), 149–160. https://doi.org/10.1016/j.appdev.2008.12.001

Gross, D., Fogg, L., Young, M., Ridge, A., Cowell, J. M., Richardson, R., & Sivan, A. (2006). The equivalence of the Child Behavior Checklist/1 1/2-5 across parent race/ethnicity, income level, and language. *Psychological Assessment*, *18*(3), 313–323. Washington, US: American Psychological Association. https://doi.org/10.1037/1040-3590.18.3.313

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. Routledge. https://doi.org/10.1080/10705519909540118

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Harder, V. S., Ang, R. P., Bilenberg, N., Bjarnadottir, G., Capron, C., De Pauw, S. S., Dias, P., Dobrean, A., Doepfner, M., Duyme, M., Eapen, V., Erol, N., Esmaeili, E. M., Ezpeleta, L., Frigerio, A., Gonçalves, M. M., & Verhulst, F. C. (2010). Preschool psychopathology reported by parents in 23 societies: Testing the seven-syndrome model of the child behavior checklist for ages 1.5–5. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*(12), 1215–1224. https://doi.org/10.1016/j.jaac.2010.08.019

Kaat, A. J., Newcomb, M. E., Ryan, D. T., & Mustanski, B. (2017). Expanding a common metric for depression reporting: Linking two scales to PROMIS® depression. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *26*(5), 1119–1128. https://doi.org/10.1007/s11136-016-1450-z

Keenan, K., & Wakschlag, L. S. (2002). Can a valid diagnosis of disruptive behavior disorder Be made in preschool children? *American Journal of Psychiatry*, *159*(3), 351–358. https://doi.org/10.1176/appi.ajp.159.3.351

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. Springer. Retrieved June 3, 2023, from http://link.springer.com/10.1007/978-1-4757-4310-4

Konold, T. R., Hamre, B. K., & Pianta, R. C. (2003). Measuring problem behaviors in young children. *Behavioral Disorders*, *28*(2), 111–123. SAGE Publications Inc. https://doi.org/10.1177/019874290302800202

Koot, H. M., Van Den Oord, E. J. C. G., Verhulst, F. C., & Boomsma, D. I. (1997). Behavioral and emotional problems in young preschoolers: Cross-cultural testing of the validity of the child behavior checklist/2-3. *Journal of Abnormal Child Psychology*, *25*(3), 183–196. https://doi.org/10.1023/a:1025791814893

Krueger, R. F., & Markon, K. E. (2011). A dimensional-spectrum model of psychopathology: Progress and opportunities. *Archives of General Psychiatry*, *68*(1), 10–11. https://doi.org/10.1001/archgenpsychiatry.2010.188

Luby, J. L., Heffelfinger, A. K., Mrakotsky, C., Hessler, M. J., Brown, K. M., & Hildebrand, T. (2002). Preschool major depressive disorder: Preliminary validation for developmentally modified DSM-IV criteria. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(8), 928–937. https://doi.org/10.1097/00004583-200208000-00011

Mansolf, M., Vreeker, A., Reise, S. P., Freimer, N. B., Glahn, D. C., Gur, R. E., Moore, T. M., Pato, C. N., Pato, M. T., Palotie, A., Holm, M., Suvisaari, J., Partonen, T., Kieseppä, T., Paunio, T., Boks, M., Kahn, R., Ophoff, R. A., Bearden, C. E., … WGSPD Consortium. (2020). Extensions of multiple-group item response theory alignment: Application to psychiatric phenotypes in an international genomics consortium. *Educational and Psychological Measurement*, *80*(5), 870–909. https://doi.org/10.1177/0013164419897307

Mathiesen, K. S., & Sanson, A. (2000). Dimensions of early childhood behavior problems: Stability and predictors of change from 18 to 30 months. *Journal of Abnormal Child Psychology*, *28*(1), 15–31. https://doi.org/10.1023/a:1005165916906

May, T., Adesina, I., McGillivray, J., & Rinehart, N. J. (2019). Sex differences in neurodevelopmental disorders. *Current Opinion in Neurology*, *32*(4), 622–626. https://doi.org/10.1097/wco.0000000000000714

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. Routledge. https://doi.org/10.1080/00273171.2014.911075

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728–743. https://doi.org/10.1037/a0018966

Medeiros, K., Mazurek, M. O., & Kanne, S. (2017). Investigating the factor structure of the Child Behavior Checklist in a large sample of children with autism spectrum disorder. *Research in Autism Spectrum Disorders*, *40*, 24–40. https://doi.org/10.1016/j.rasd.2017.06.001

Merrell, K. W. (2001). Assessment of children's social skills: Recent developments, best practices, and new directions. *Exceptionality*, *9*(1–2), 3–18. https://doi.org/10.1080/09362835.2001.9666988

Müller, J. M., Achtergarde, S., & Furniss, T. (2011). The influence of maternal psychopathology on ratings of child psychiatric symptoms: An SEM analysis on cross-informant agreement. *European Child & Adolescent Psychiatry*, *20*(5), 241–252. https://doi.org/10.1007/s00787-011-0168-2

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00978

Olson, S. L., Lansford, J. E., Evans, E. M., Blumstein, K. P., & Ip, K. I. (2019). Parents' ethnotheories of maladaptive behavior in young children. *Child Development Perspectives*, *13*(3), 153–158. https://doi.org/10.1111/cdep.12330

Prior, M., Smart, D., Sanson, A., Pedlow, R., & Oberkiaid, F. (1992). Transient versus stable behavior problems in a normative sample: Infancy to school age. *Journal of Pediatric Psychology*, *17*(4), 423–443. https://doi.org/10.1093/jpepsy/17.4.423

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. [American Sociological Association, Wiley, Sage Publications, Inc.]. https://doi.org/10.2307/271063

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rescorla, L. A., Achenbach, T. M., Ivanova, M. Y., Harder, V. S., Otten, L., Bilenberg, N., Bjarnadottir, G., Capron, C., De Pauw, S. S. W., Dias, P., Dobrean, A., Döpfner, M., Duyme, M., Eapen, V., Erol, N., Esmaeili, E. M., Ezpeleta, L., Frigerio, A., Fung, D. S. S., & Verhulst, F. C. (2011). International comparisons of behavioral and emotional problems in preschool children: Parents' reports from 24 societies. *Journal of Clinical Child and Adolescent Psychology: The official journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *40*(3), 456–467. https://doi.org/10.1080/15374416.2011.563472

Rescorla, L. A., Ghassabian, A., Ivanova, M. Y., Jaddoe, V. W., Verhulst, F. C., & Tiemeier, H. (2019). Structure, longitudinal invariance, and stability of the child behavior checklist 1½–5's diagnostic and statistical manual of mental disorders–autism spectrum disorder scale: Findings from generation R (rotterdam). *Autism*, *23*(1), 223–235. SAGE Publications Ltd. https://doi.org/10.1177/1362361317736201

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer. Retrieved August 24, 2022. https://doi.org/10.1007/978-1-4757-2691-6_5

Schroeder, J., Hood, M., & Hughes, H. (2010). (Vol., pp.). Inter-parent agreement on the syndrome scales of the child behavior checklist (CBCL): Correspondence and discrepancies. *Journal of Child and Family Studies*, *19*(5), 646–653. Springer Nature

Stülb, K., Messerli-Bürgy, N., Kakebeeke, T. H., Arhab, A., Zysset, A. E., Leeger-Aschmann, C. S., Schmutz, E. A., Meyer, A. H., Kriemler, S., Jenni, O. G., Puder, J. J., & Munsch, S. (2019). Prevalence and predictors of behavioral problems in healthy Swiss preschool children over a one year period. *Child Psychiatry and Human Development*, *50*(3), 439–448. https://doi.org/10.1007/s10578-018-0849-x

Sun, H., Yussof, N. T. B., Mohamed, M. B. B. H., Rahim, A. B., Bull, R., Cheung, M. W. L., & Cheong, S. A. (2021). Bilingual language experience and children's social-emotional and behavioral skills: A cross-sectional study of Singapore preschoolers. *International Journal of Bilingual Education and Bilingualism*, *24*(3), 324–339. Routledge. https://doi.org/10.1080/13670050.2018.1461802

Tan, T. X., Dedrick, R. F., & Marfo, K. (2007). Factor structure and clinical implications of child behavior checklist/1.5–5 ratings in a sample of girls adopted from China. *Journal of Pediatric Psychology*, *32*(7), 807–818. https://doi.org/10.1093/jpepsy/jsm025

Tandon, M., Cardeli, E., & Luby, J. (2009). Internalizing disorders in early childhood: A review of depressive and anxiety disorders. *Child and Adolescent Psychiatric Clinics*, *18*(3), 593–610. Elsevier. https://doi.org/10.1016/j.chc.2009.03.004

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. SAGE Publications Inc. https://doi.org/10.1177/109442810031002

Vaughn-Coaxum, R. A., Mair, P., & Weisz, J. R. (2016). Racial/ethnic differences in youth depression indicators: An item response theory analysis of symptoms reported by White, Black, Asian, and Latino youths. *Clinical Psychological Science: A journal of the Association for Psychological Science*, *4*(2), 239–253. https://doi.org/10.1177/2167702615591768

Wakschlag, L. S., Briggs-Gowan, M. J., Carter, A. S., Hill, C., Danis, B., Keenan, K., McCarthy, K. J., & Leventhal, B. L. (2007). A developmental framework for distinguishing disruptive behavior from normative misbehavior in preschool children. *Journal of Child Psychology and Psychiatry*, *48*(10), 976–987. https://doi.org/10.1111/j.1469-7610.2007.01786.x

Wakschlag, L. S., Perlman, S. B., Blair, R. J., Leibenluft, E., Briggs-Gowan, M. J., & Pine, D. S. (2017). The neurodevelopmental basis of early childhood disruptive behavior: Irritable and callous phenotypes as exemplars. *American Journal of Psychiatry*, *175*(2), 114–130. American Psychiatric Publishing. https://doi.org/10.1176/appi.ajp.2017.17010045

Wakschlag, L. S., Tolan, P. H., & Leventhal, B. L. (2010). Research review: 'Ain't misbehavin': Towards a developmentally-specified nosology for preschool disruptive behavior. *Journal of Child Psychology and Psychiatry*, *51*(1), 3–22. https://doi.org/10.1111/j.1469-7610.2009.02184.x

Yuan, K.-H., & Chan, W. (2005). On nonequivalence of several procedures of structural equation modeling. *Psychometrika*, *70*(4), 791–798. https://doi.org/10.1007/s11336-001-0930-9

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zheng, S., Mansolf, M., McGrath, M., Churchill, M. L., Bekelman, T. A., Brennan, P. A., Margolis, A. E., Nozadi, S. S., Bastain, T. M., Elliott, A. J., LeWinn, K. Z., Hofheimer, J. A., Leve, L. D., Rennie, B., Zimmerman, E., Marable, C. A., McEvoy, C. T., Liu, C., Sullivan, A., Woodruff, T. J., … Bishop, S. (2024). Measurement bias in caregiver-report of early childhood behavior problems across demographic factors in an ECHO-wide diverse sample. *JCPP Advances*, *4*(1), e12198. https://doi.org/10.1002/jcv2.12198