Journal of Pediatric Psychology, 47(5), 2022, 510–522
https://doi.org/10.1093/jpepsy/jsac025
Advance Access Publication Date: 9 April 2022
Original Research Article

Psychometric Considerations in Developing PROMIS® Measures for Early Childhood

Jin-Shei Lai, PhD, OTR, Michael A. Kallen, PhD, MPH, Courtney K. Blackwell, PhD, Lauren S. Wakschlag, PhD, and David Cella, PhD

Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine and Institute for Innovations in Developmental Sciences (DevSci), USA

All correspondence concerning this article should be addressed to Jin-Shei Lai, PhD, OTR, Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 625 N. Michigan Ave., 21st Floor, Chicago, IL 60611, USA. E-mail: js-lai@northwestern.edu

Received June 23, 2021; revisions received March 4, 2022; accepted March 4, 2022

Abstract

Objective The early expression of lifespan health and disease states can often be detected in early childhood. Currently, the Patient-Reported Outcome Measurement Information System (PROMIS®) includes over 300 measures of health for individuals ages 5 years and older. We extended PROMIS to early childhood by creating developmentally appropriate, lifespan coherent parent-report measures for 1–5-year-olds. This paper describes the psychometric approaches used for these efforts; **Methods** 2 waves of data from parents of children ages 1–5 were collected via 2 internet panel companies. Wave 1 data ($n = 1{,}400$) were used to evaluate item pool unidimensionality, model fit, and initial item parameters. Combined data from wave 1 and wave 2 (reference sample; $n = 1{,}057$) were used to estimate final item parameters. Using item response theory methods, we developed and tested 12 item pools: Global Health, Physical Activity, Sleep Disturbance, Sleep-related Impairment, Anger/Irritability, Anxiety, Depressive Symptoms, Positive Affect, Self-Regulation, Engagement, Family Relationships, and Peer Relationships; **Results** Wave 1 analyses supported the unidimensionality of Physical Activity, Positive Affect, Anger/Irritability, Anxiety, Depressive Symptoms, and Global Health. Family Relationships and Peer Relationships were combined to form "Social Relationships"; Sleep Disturbance and Sleep-related Impairment were combined to form "Sleep Problems." Self-Regulation was divided into "Flexibility" and "Frustration Tolerance"; Engagement was divided into "Curiosity" and "Persistence." Short forms were developed for item banks with more than 10 items; and **Conclusions** Using rigorous mixed-methods, we successfully extended PROMIS to early childhood (1–5-year-olds). Measures are now publicly available in English and Spanish (www.healthmeasures.net).

Key words: infancy and early childhood, measure validation, preschool children, quality of life, research design and methodology, statistical approach.

Child health affects developmental functioning, which can lead to long-term negative outcomes such as problem behaviors, lower educational attainment, and psychopathology (Diener et al., 2010; Keyes, 2002, 2010; Smees et al., 2020). National and international efforts to understand children's health often rely on large-scale population-based surveys. For example, the National Survey of Children's Health (Blumberg et al., 2012) provides snapshots of U.S. children's health across a range of conditions and contextual factors (e.g., social determinants of health), and aggregated metrics based on vital statistics, clinical records, and administrative data offer additional indicators (e.g., infant mortality, percent of children living in poverty) often used to describe the broader health and well-being of children.

While such efforts enable comparisons across time and with other developed countries, they are often limited in their ability to evaluate more than just chronic health diagnoses or more than a handful of individual items that stand as proxies for physical, mental, or social health. Conversely, there are many validated—but lengthy—developmentally based assessment tools to conduct deeper and dimensional phenotyping of specific health conditions, precursors, and underlying processes (Halle & Darling-Churchill, 2016). However, these rarely include systematic consideration of quality of life and related health constructs in a unified measurement system (Coghill et al., 2009).

Given our experiences in the Patient-Reported Outcome Measurement Information System (PROMIS®; Cella et al., 2008, 2010) and other similar measurement systems (e.g., Quality of Life in Neurological Disorders; Neuro-QoL^{TM}; Lai et al., 2012), we set out to create parent-report measures to capture quality of life and related health constructs of children aged 1–5 years, as described in Cella et al. (in this issue). These developmentally sensitive measures were created with item response theory (IRT) models for calibration and norming to overcome measurement challenges of monitoring children's quality of life and related health constructs across developmental phases in a psychometrically sound manner. These measure applications include computer adaptive tests (CATs), in which the most informative items of respondents' trait levels (e.g., function, well-being) are selected and administered based on participants' responses to the previously administered item from a "bank" of items. Also, fixed-length short forms (SFs) (Lai, Butt, et al., 2011; Lai, Cella, et al., 2011; Pilkonis et al., 2011) can be customized from these calibrated item banks to consist the most relevant questions for a given context. Both CATs and SFs can produce brief-yet-precise score estimations that are comparable to scores produced by the corresponding full-length item bank. Subsequently, investigators can administer measures tapping multiple key domains of interests without adding significant response burden. The PROMIS provides a well-validated framework for advancing this goal. However, until now, PROMIS only extends down to age 5.

The overarching aim of the PROMIS Early Childhood (EC) initiative was to fill this assessment void by generating and validating psychometrically sound and developmentally appropriate PROMIS parent report measures that can be used in large-scale research studies and clinical practice to evaluate physical, mental, and social health outcomes of 1–5-year-olds. See Blackwell et al. (2020) for conceptual considerations of this initiative. See Cella et al. (this issue) for a discussion on the general qualitative approaches used to develop the PROMIS EC measures. Here, we describe the quantitative psychometric approaches taken to achieve our aims.

Methods

This study was approved by the Northwestern University Institutional Review Board. Data are available upon request.

Recruitment and Participants

The sampling strategy was designed to (a) ensure variability across participants whose scores are spread across the measurement continuum of the targeted domains, such that resulting item parameters are more stable across different sample groups (i.e., wave 1 data collection) and (b) establish population-based reference values (i.e., wave 2 data collection).

To accomplish these goals, we enlisted two internet survey panel companies, Op4G and Ipsos, to partner with us in the collection of two waves of data. The primary goal of the wave 1 testing was to determine items to be included in item banks or calibrated scales. Op4G, an internet panel company, was chosen because of previously successful collaborations for PROMIS pediatric measurement development initiatives (e.g., Paller et al., 2021) in which they were able to recruit sufficient diverse samples within a reasonable timeline. The purpose of wave 2 testing was to finalize item calibrations and establish norms. Ipsos, an internal panel company, was chosen because they have the largest nationally representative probability-based panel. Inclusion criteria for both wave 1 and wave 2 were: (a) 18 years or older; (b) parent (i.e., biological, foster, adoptive, or step-parent) of at least one child aged 12 months to 5 years, 11 months, 30 days at the time of survey collection; and (c) able to read and respond in English. All participants provided informed consent via the panel companies prior to completing surveys (as described below), and Northwestern received deidentified data.

The wave 1 sample consisted of 1,400 parents of children ages 1–5 years (ages 1–3 years: $n = 700$; ages 4–5: $n = 700$) recruited by Op4G through its partnership with nonprofit organizations. When a member joins the Op4G panel, they complete a demographic questionnaire, which is used for targeting individuals for participation in specific research projects. For this study, Op4G identified English-speaking parents in the U.S. with children in the desired age range to complete the wave 1 survey. Those parents then received an e-mail notifying them of a new survey opportunity. Parents were screened to ensure they had a child who met the age inclusion criteria, and, if they agreed to participate, they were randomly assigned to either survey form A or form B. This strategy was taken to lessen respondent burden. Form A ($n = 85$ items) covered the domains of *Family Relationships*, *Peer Relationships*, *Physical Activity*, *Sleep Disturbance*, *Sleep-Related Impairment*, and *Global Health*; form B ($n = 109$ items) covered *Anger/Irritability*, *Anxiety*, *Depressive Symptoms*, *Positive Affect*, *Engagement*, *Self-Regulation*, and *Global Health*. Both forms also included two checklists for parents to identify if their child had (a) any physical health conditions or (b) any emotional or behavioral conditions or developmental disorders (EBD) such as anxiety, attention deficit disorder (ADD) or attention deficit hyperactivity disorder (ADHD), and autism spectrum disorder. Figure 1 shows the number of items included in each item pool in wave 1 testing. Data collection ended when sample composition matched the ratio of race, sex, and education of the 2010 U.S. Census. No response rate was estimated by Op4G.

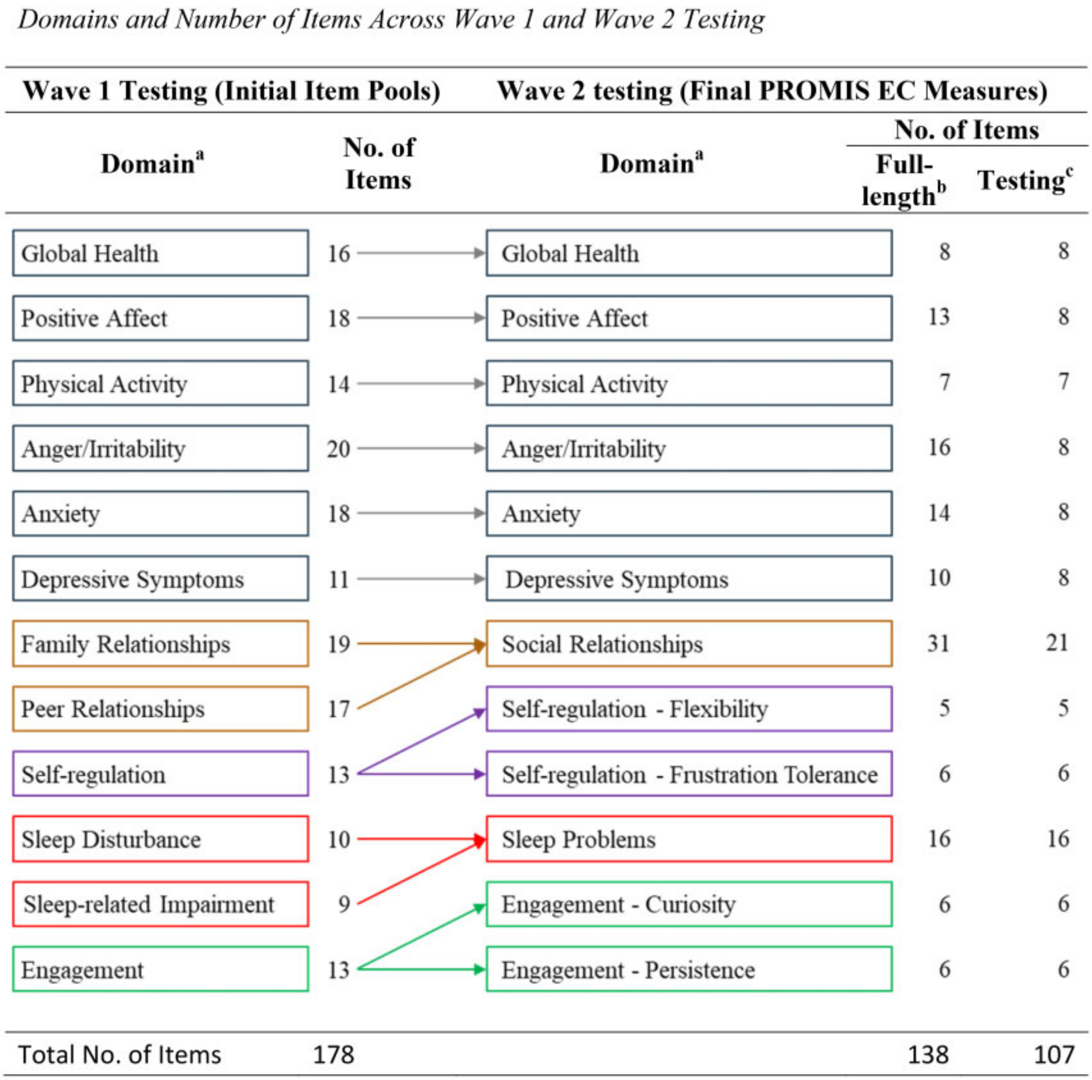
The wave 2 sample included 1,057 parents of children ages 1–5 years who were recruited by Ipsos. Ipsos hosts the longest-standing all-online research panel that is representative of the entire U.S. population. The Ipsos “KnowledgePanel” includes more than 55,000 members who are randomly recruited through probability sampling. If needed, households are provided with Internet access and a computer. Ipsos recruits panel members by using address-based sampling methods. Once household members are recruited for the panel and assigned to a study sample, they are notified by email for survey participation; panelists can also visit their online member page for survey opportunity updates. This allows surveys to be fielded quickly, economically, and with less burden on participants. Final data were weighted using the Ipsos standard categories: sex (male, female), age (18–29, 30–34, 35–39, 40+ years), race/ethnicity (White, Black, other, Hispanic, 2+ races), census region (Northeast, Midwest, South, West), metropolitan status (metro, nonmetro), education (less than high school/high school, some college, bachelor or higher), and household income (under $25,000, $25–$49,999, $50–$74,999, $75–$99,999, $100–$149,999, $150,000 and over). Weights-matched parents who were at least 18-year-old with one or more children of ages 1–5 from the U.S. Census Bureau’s March 2018 Current Population Survey. More information regarding weighting and how wave 2 reflected the general U.S. population is available upon request. Unlike the wave 1 sample who completed all items (including those subsequently excluded) on some domains, the wave 2 sample completed items on all domains resulting from the end of wave 1 testing (see Figure 1). We took this approach to evaluate statistical relationships across domains to further refine measures. With the consideration of response burden, for measures consisting of more than eight items, an 8-item SF was created and administered in wave 2 testing. More detailed information about the SF item section process can be found in the domain-specific papers in this special section. The wave 2 sample also completed the same health conditions checklists as in wave 1. The reported response rate was 43%.

Analysis Plans

Two stages of data analyses were conducted (see Figure 2). Using data from wave 1 testing, we evaluated the dimensionality of item pools to determine the number of measures to develop. We then evaluated measurement stability between subsamples and estimated measure item parameters to construct the SFs to be used in wave 2 testing. Wave 2 data were then used, in combination with wave 1 data, to finalize item parameters and establish reference values. Some specific analytic approaches varied across domains, but underlying measurement principles and analytic goals were the same. We describe here the general analytical approaches taken, with other details provided in the domain-specific manuscripts. For the final measures, we refer to calibrated item sets of 10 or more items as “item banks”; those with fewer than 10 items are referred to as “calibrated scales.” This is the PROMIS arbitrary semantic distinction; item banks and calibrated scales were developed using the same psychometric standards and can be administered using the same approaches (e.g., SFs and CAT).

Wave 1 Analyses

The first part of the wave 1 analyses focused on evaluation of the unidimensionality of item pools using exploratory factor analysis (EFA) or confirmatory factor analysis (CFA). We applied EFA to obtain a quantitative sense of the dimensionality of proposed item sets. We explored the number of potential factors using the following criteria: (a) number of factors with eigenvalue > 1; (b) review of the scree plot (i.e., number of factors before the break in scree plot); (c) percentage of variance explained by eigenvalue 1 (criterion: ≥ 40); (d) ratio of eigenvalue 1-to-2 (criterion: ≥ 4); and (e) number of factors that explain $>5\%$ of variance (Hu & Bentler, 1999; Kline, 2008; Lai et al., 2006; Lai, Zelko, et al., 2011). A promax rotation with polychoric correlations was then used to examine the association among factors and items and to examine item loadings (criterion: >0.4). Promax rotation was chosen to allow factors to be correlated so we could explore potential multidimensionality among items. Polychoric correlations were implemented to adjust for the ordinal nature of the item response data. Either SAS 9.4 (Gary, NC: SAS Institute) or



Domains and Number of Items Across Wave 1 and Wave 2 Testing

Wave 1 Testing (Initial Item Pools)		Wave 2 testing (Final PROMIS EC Measures)		
			No. of Items	
Domain[a]	No. of Items	Domain[a]	Full-length[b]	Testing[c]
Global Health	16	Global Health	8	8
Positive Affect	18	Positive Affect	13	8
Physical Activity	14	Physical Activity	7	7
Anger/Irritability	20	Anger/Irritability	16	8
Anxiety	18	Anxiety	14	8
Depressive Symptoms	11	Depressive Symptoms	10	8
Family Relationships	19	Social Relationships	31	21
Peer Relationships	17	Self-regulation - Flexibility	5	5
Self-regulation	13	Self-regulation - Frustration Tolerance	6	6
Sleep Disturbance	10	Sleep Problems	16	16
Sleep-related Impairment	9	Engagement - Curiosity	6	6
Engagement	13	Engagement - Persistence	6	6
Total No. of Items	178		138	107

Figure 1. Domains and number of items across wave 1 and wave 2 testing. [a]Family Relationships and Peer Relationships were combined into one Social Relationships item bank. Sleep Disturbance and Sleep-related Impairment were combined into one Sleep Problems item bank. Self-regulation was separated into Flexibility and Frustration Tolerance calibrated scales. Engagement was separated into Curiosity and Persistence calibrated scales. [b]Full-length: number of items included in the final PROMIS EC measures. Measures with 10 or more items are named "item banks." Measures with less than 10 items are named "calibrated scales." [c]Testing: number of items used for wave 2 testing.

Mplus v7.4 (Los Angeles, CA: Muthen & Muthen) was used to conduct EFA modeling analyses.

If items within the item pool had hypothesized factor structures available, CFAs were conducted to test the hypothesized structure using the weighted least square-mean and variance adjusted estimator with polychoric correlations, as is recommended for ordinal response data (Muthen et al., 1997). Item sets were considered unidimensional, with their items retained in the item pool, when the following criteria were met: comparative fit index (CFI) > 0.9; root mean square error of approximation (RMSEA) < 0.10; $R^2 > 0.30$; residual correlations < 0.20 (Hu & Bentler, 1999; Kline, 2008; Lai et al., 2006; Lai, Zelko, et al., 2011). We employed Mplus v7.4 to conduct CFA modeling analyses.

Bi-factor analysis (BFA; McDonald, 1999), in the CFA family, was used to evaluate sufficient unidimensionality when we took subdomains into account (i.e., for *Social Relationships* and *Sleep Problems*). The BFA specifies a general factor (defined by loadings from all included items) and local factors (or subdomains, defined by loadings from prespecified groups of items related to their subdomains). The general factor and local factors are modeled as orthogonal; therefore, the relationship between items and the general factor is not constrained to be proportional to the relationship between first- and second-order factors, as demonstrated in other hierarchical CFAs. When the general factor explains the covariance between items, uniformly high standardized item loadings on the general factor are observed; it is then appropriate to

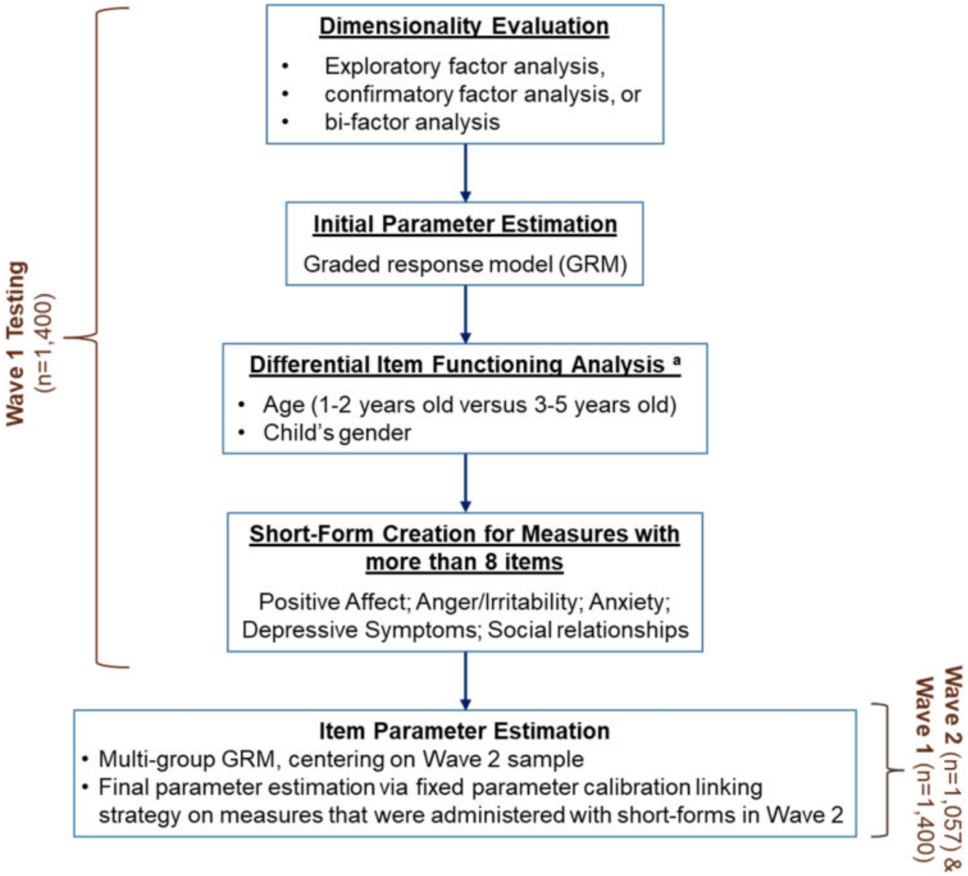

Dimensionality Evaluation
• Exploratory factor analysis,
• confirmatory factor analysis, or
• bi-factor analysis
Initial Parameter Estimation
Graded response model (GRM)
Differential Item Functioning Analysis a
• Age (1-2 years old versus 3-5 years old)
• Child's gender
Short-Form Creation for Measures with more than 8 items
Positive Affect; Anger/Irritability; Anxiety; Depressive Symptoms; Social relationships
Item Parameter Estimation
• Multi-group GRM, centering on Wave 2 sample
• Final parameter estimation via fixed parameter calibration linking strategy on measures that were administered with short-forms in Wave 2
Wave 1 Testing (n=1,400)
Wave 2 (n=1,057) & Wave 1 (n=1,400)

Figure 2. Analysis flowchart. ^{a}DIF on parent's sex was evaluated. However, because there were significantly fewer fathers versus mothers who completed the testing, evaluation of DIF on parent sex was considered exploratory and not included in the flowchart.

report a single score representing the domain of interest. If the subdomains represent demonstrably separate concepts, loadings on the general factor will not be uniformly high, and we reject the conclusion that the item set was sufficiently unidimensional, making it more appropriate to report scores of subdomains separately (Lai, Butt, et al., 2011; Lai et al., 2006). Sufficient unidimensionality was confirmed when factor loadings to local factors were less than those to the general factor and the BFA model had acceptable fit indices (i.e., CFI > 0.90; RMSEA < 0.10; $R^2 > 0.30$; residual correlations < 0.20). Additionally, variance explained by the general factor should be greater than that explained by local factors if sufficient unidimensionality is supported.

The second part of wave 1 analyses focused on IRT modeling and item misfit. We modeled responses to candidate item sets using the graded response model (GRM; Samejima, 1997) as implemented in IRTPRO v3.1 (Chapel Hill, NC: Vector Psychometric Group). We defined item misfit as occurring when the item fit test ratio of chi-squared to degrees of freedom was >3.0 (Crişan et al., 2017; Stark et al., 2006).

Item parameters were used to estimate item information functions at both the level of individual items and the level of the entire item bank, in order to characterize the precision of items and the overall scale across the measurement continuum. Items with higher information function values are more likely to be chosen by computerized adaptive testing (CAT), given their high precision and reduced error rate at measuring the domain of interest. For a measure with more than eight items, we constructed an 8-item SF to be included in the wave 2 testing with consideration of response burden. Items were chosen by the study team of clinicians and measurement experts based on the item's clinical-representative and information function value that provided the best attainable score-level reliabilities across our targeted *T*-score range. We also simulated CAT administration for item banks with more than eight items using the following settings: (a) minimum number of items to administer = 4; (b) maximum number of items to administer = 8; and (c) SE stopping criterion: < 0.4.

Differential item functioning (DIF). We investigated stability of measurement properties of items by evaluating DIF on the factors of developmental differences for toddlers versus preschoolers (1–2-year-olds vs. 3–5-year-olds), parent's sex (female vs. male), and child's sex (female vs. male) using data from the full-length measures administered in wave 1. However, because there were significantly fewer fathers compared to mothers, DIF on parent sex was investigated for exploratory purposes only. Differential item functioning was conducted by using lordif (Choi et al., 2011), a hybrid logistic ordinal regression/IRT ability score framework. We flagged items for potential DIF if they had a Nagelkerke *pseudo-R-squared* value $\geq$0.20. We evaluated DIF "score impact" by comparing scores with and without DIF items by computing the standard deviation (SD) of score differences and their root mean square difference (RMSD). We also calculated the percent of individual cases whose "score impact" (absolute value) exceeded their scores' standard error (SE) estimate when all items were administered. We considered "score impacts" exceeding score SEs as nontrivial differences.

Analyses Using a Combined Wave 1 and Wave 2 Sample

Step 1. This phase focused on estimating final item parameters, creating SFs, and estimating general population reference values. We first conducted multigroup item calibration analyses, using wave 2 item response data as our reference sample. Because the wave 2 sample was collected using a random probability-based sampling strategy and weighted on key variables as described above to match U.S. parents with children ages 1–5 years, we centered the final item parameters on the wave 2 sample.

Step 2. In consideration of response burden, the wave 2 sample did not complete all items included in final measures (see Figure 1). Therefore, after completing Step 1 analyses, we created measurement links to items that were not administered in wave 2 so that all item parameters were placed on the same metric defined by Step 1. This was done by using "fixed parameter calibration" (FPC), as implemented in the PROsetta Stone linking methodology (Cella et al., 2016; Lai et al., 2014); in which, items that were not administered to wave 2 participants were calibrated by fixing item parameters obtained in Step 1. As a result, all items were on the same metric. We employed IRTPRO v3.1 to conduct the multigroup IRT centering and item calibration analyses. Measures were reported using the PROMIS T-score metric, where mean (of wave 2 sample) = 50 and standard deviation = 10. Higher scores represent *better Global Health*, *Physical Activity*, *Social Relationships*, *Positive Affect*, *Engagement—Curiosity*, *Engagement—Persistence*, *Self-Regulation—Flexibility*, and *Self-Regulation—Frustration Tolerance*. In contrast, higher scores represent *worse Sleep Problems*, *Anger/Irritability*, *Anxiety*, and *Depressive Symptoms*.

Reliability Analyses

We estimated classical test theory (CTT) based Cronbach's alpha as well as IRT-based reliabilities at both measure and individual *T*-score levels. The IRT-based overall reliability is a distribution-informed estimate, calculated from score SD and median SE. Individual score-level reliability is also derived from a SE estimate (i.e., reciprocal of test information). Reliabilities at individual *T*-score levels varied across the measurement continuum. We focused on the *T*-score range most likely to require clinical attention. A reliability $\geq$0.70 was considered acceptable (Reeve et al., 2007).

We examined the associations between each PROMIS EC domain using both Pearson correlation coefficient and Spearman's rho correlations, as appropriate per measure score distribution. We evaluated the strength of correlations using standard intervals established in the literature ($r = 0$, *no correlation;* $r =$ below $\pm$0.10, *low*; $r = \pm 0.30$, *moderate*; $r \geq \pm 0.50$, *large*; $r = 1$, *perfect correlation*; Cohen, 1988). We evaluated known-group differences based on parent report of (a) whether the child had an EBD and (b) whether the child had any physical condition or disorder (e.g., asthma, blood disorder). Using a median split to create groups, we also established known groups based on PROMIS EC *Global Health T* scores ("low" < 45 vs. "high" $\geq$ 45) when appropriate. We used parametric or nonparametric one-way analysis of variance (ANOVA), as appropriate per measure score distribution, to investigate domain score differences by known groups. We based effect size on using the following interpretation guidelines: for Cohen's *d*: "small" effect = 0.20–0.49; "medium" effect = 0.50–0.79; "large" effect$\geq$ 0.80. Eta-squared in one-way ANOVA was estimated, when appropriate, using the following criteria: small = 0.02–0.06 (exclusive), medium = 0.06 (inclusive)–0.14 (exclusive); and large $\geq$0.14 (Cohen, 1988).

Results

Sample

Sample characteristics are shown in Table I. Most questionnaires were completed by mothers (75.9%, 82.9%, and 68.8% for wave 1 form A, wave 1 form B, and wave 2, respectively). Children of participants were similar in age (mean = 2.58 years [SD = 1.2] and 2.62 years [SD = 1.2] for wave 1 and wave 2, respectively) and samples consisted of a similar percentage of males and females; most were White, of non-Hispanic origin, and did not have physical or EBD health conditions. The wave 2 sample was "healthier" than the wave 1 sample as more parents reported their child having no physical (X^2 [1, $N = 2{,}444$] = 45.61, $p < .001$) and/or no EBD (X^2 [1, $N = 2{,}437$] = 74.03,

Table I. *Sample Characteristics*

		Wave 1				Wave 2	
		Form A ($n = 700$)		Form B ($n = 700$)		$n = 1057$	
Variable	Categories	*n*	(%)	*n*	(%)	*n*	(%)
Sex (parent)	Male	169	(24.1)	120	(17.1)	330	(31.2)
	Female	531	(75.9)	580	(82.9)	727	(68.8)
Hispanic	Yes	133	(19.0)	135	(19.3)	163	(15.4)
Sex (child)	Male	366	(52.3)	348	(49.7)	526	(49.8)
	Female	334	(47.7)	352	(50.3)	528	(50.0)
Race (child)	White	511	(73.0)	499	(71.3)	841	(79.6)
	Black or African American	85	(12.1)	97	(13.9)	70	(6.6)
	Asian American	22	(3.1)	20	(2.9)	36	(3.4)
	Native American or Alaska Native	10	(1.4)	5	(0.7)	4	(0.4)
	Native Hawaiian or Other Pacific Islander	3	(0.4)	4	(0.6)	0	(0)
	Multiracial	58	(8.3)	65	(9.3)	86	(8.1)
	Not reported	11	(1.6)	10	(1.4)	20	(1.9)
Health condition	No physical health condition	428	(61.1)	456	(65.1)	793	(75.0)
	No emotional/behavioral/developmental health conditions	504	(72.0)	556	(79.4)	927	(87.7)
	Allergies	182	(26.0)	149	(21.3)	136	(12.9)
	Asthma	74	(10.6)	70	(10.0)	29	(2.7)
	Blindness or problems with seeing, even when wearing glasses	11	(1.6)	8	(1.1)	19	(1.8)
	Anxiety problems	48	(6.9)	32	(4.6)	12	(1.1)
	Attention deficit disorder	83	(11.9)	53	(7.6)	14	(1.3)
	Behavioral or conduct problems	49	(7.0)	27	(3.9)	9	(0.9)
	Developmental delay	43	(6.1)	36	(5.1)	34	(3.2)
	Speech or other language disorder	48	(6.9)	47	(6.7)	68	(6.4)

$p < .001$) health condition. The most prevalent condition across waves was allergies (including food, medication, insect, or other) followed by ADD/ADHD (wave 1 form A), asthma (wave 1 form B), and speech/language disorder (wave 2).

Wave 1 Analyses Results

Here we provide an overview of analysis results across measures. More detailed analysis results for each measure, including factor analyses, are described in the special section domain specific manuscripts that follow.

In Figure 1, we show the initial and revised domains from wave 1 to wave 2 testing, as well as the items per domain for each wave. In wave 1, 178 items across 12 domains (item pools) were developed and tested. Different factor analysis techniques (EFA, CFA, or BFA) were applied depending on the nature of the item pools being evaluated (details are described in domain specific manuscripts in this issue). Results from factor analyses, along with input from psychometric and developmental experts, led to refinement of some domains; specifically, *Engagement* was separated into *Curiosity* and *Persistence; Self-Regulation* was divided into *Flexibility* and *Frustration Tolerance*; *Sleep Disturbance* and *Sleep-related Impairment* were combined into one *Sleep Problems* item bank; and *Family Relationships* and *Peer Relationships* were combined into one *Social Relationships* item bank. The other six domains (*Global Health*, *Positive Affect*, *Physical Activity*, *Anger/Irritability*, *Anxiety*, and *Depressive Symptoms*) stayed the same.

We further analyzed items, estimating item parameters, and evaluating item fit using the GRM. We did not identify any significant DIF on all comparisons across measures. Therefore, no items were removed due to DIF evaluations for item bias. As a result, 138 items across 12 measures were retained at the end of wave 1 testing.

Short forms were created for measures without subdomains and with more than eight items (*Positive Affect*, *Anger/Irritability*, *Anxiety*, and *Depressive Symptoms)*; they were subsequently used for wave 2 testing. For *Social Relationships*, a 6-item SF with two items from each subdomain was created; meanwhile, 21 items representing subdomains of *Family Relationships* (8 items), *Peer Relationships* (8 items), and *Caregiver–Child Interactions* (5 items) were retained for wave 2 testing. For *Sleep Problems*, an 8-item SF with 4 items from each subdomain was created; meanwhile, all 16 items (8 items for *Disturbance* and 8 items for *Impairment*) were retained for wave 2 testing (see Figure 1).

Analyses Results Using a Combined Wave 1 and Wave 2 Sample

Multigroup item calibration analyses were applied with all 12 measures; of them, FPC linking was also

Table II. *IRT-Based Reliability Function Across the Measurement Continuum*

	# of items	Cronbach's alpha	IRT-based reliability									
			Overall	$T=20$	$T=30$	$T=40$	$T=45$	$T=50$	$T=55$	$T=60$	$T=70$	$T=80$
Higher scores, better well-being												
Global Health	8	0.91	0.88	0.93	0.93	0.92	0.92	0.90	0.80	0.51	0.00	0.00
Relationship	31	0.95	0.96	0.97	0.97	0.97	0.97	0.97	0.95	0.91	0.63	0.00
Physical Activity	5	0.82	0.74	0.89	0.89	0.87	0.88	0.83	0.64	0.00	0.00	0.00
Curiosity	6	0.90	0.89	0.91	0.91	0.90	0.90	0.91	0.88	0.78	0.00	0.00
Persistence	6	0.90	0.86	0.87	0.87	0.85	0.86	0.86	0.85	0.85	0.70	0.00
Positive Affect	13	0.95	0.94	0.97	0.96	0.94	0.94	0.97	0.95	0.86	0.00	0.00
Flexibility	5	0.91	0.87	0.87	0.89	0.84	0.86	0.89	0.85	0.87	0.79	0.00
Frustration Tolerance	6	0.90	0.87	0.82	0.89	0.86	0.84	0.85	0.89	0.86	0.88	0.45
Higher score, worse symptom												
Sleep Problems	16	0.95	0.95	0.00	0.60	0.91	0.94	0.95	0.95	0.95	0.95	0.94
Anger/Irritability	16	0.95	0.94	0.05	0.82	0.94	0.95	0.96	0.95	0.95	0.96	0.95
Anxiety	14	0.96	0.93	0.00	0.00	0.56	0.85	0.94	0.96	0.96	0.96	0.96
Depressive Symptoms	10	0.94	0.87	0.00	0.00	0.00	0.65	0.88	0.94	0.95	0.96	0.96

Note. Expected reliabilities which were estimated using simulation; $T=50$ (SD = 10) is the average score for the general population.

applied for five measures: *Social Relationships*, *Positive Affect*, *Anger/Irritability*, *Anxiety*, and *Depressive Symptoms*. Final parameters were centered on the wave 2 sample, and scores were reported using the PROMIS *T*-score metric. Both CTT- and IRT-based reliabilities were greater than 0.70 (see Table II). Item response theory-based reliabilities at the individual level varied across the measurement continuum. For all measures, higher reliabilities tended to be associated with worse well-being (e.g., *GH T* scores $\leq$ 55 where IRT-based score-level reliabilities are $\geq$0.70) or worse symptoms (e.g., higher *Sleep Problems T* scores $\geq$40 where IRT-based score-level reliabilities are $\geq$0.70); lower reliabilities tended to be associated with better well-being or less symptom burden. These results coincide with the test information function curves for all measures, as depicted in Figure 3a and b. It was not surprising to find that measures with fewer items showed relatively smaller test information function values (i.e., lower reliabilities), as the information function value at the measure level is the accumulation of information values across all items. Regardless of measure item counts, all measures exhibited acceptable reliabilities ($\geq$0.70) across at least a 30-point *T*-score range (e.g., *T* scores = 20–50 for *Physical Activity* and *T* scores = 50–80 for *Depressive Symptoms*).

Descriptive statistics of all measures are shown in Table III. Score distributions of each wave can be found in manuscripts for each measure in this special issue. Wave 2 means were 50 on all measures by design. For measures that were administered using SFs in wave 2 testing (i.e., *Positive Affect*, *Anger/irritability*, *Anxiety*, *Depressive Symptoms*, and *Social Relationships*), we compared SF scores between the two waves to avoid potential measurement error due to the use of estimated scores with missing data. Compared to the wave 2 sample, the wave 1 sample reported significantly (p < .01) worse scores (i.e., worse well-being or symptom) on *Global Health*, *Sleep Problems*, *Anxiety*, and *Depressive Symptoms*, but better scores (i.e., better well-being or less symptom) on *Physical Activity*, *Engagement—Persistence*, *Self-Regulation—Flexibility*, and *Self-Regulation—Frustration Tolerance*. Most effect sizes were moderate, with the greatest effect size found on *Sleep Problems* (0.68). See Table IV for the PROMIS EC correlation matrix and refer to individual papers for discussion of correlation and known-groups differences results.

Discussion

Evaluating young children's health in a psychometrically robust manner can be challenging. The assessment of young children's health should incorporate a developmental and relational perspective. At this young age, items should reflect perceptions of caregivers given young children's less developed communication skills that might prevent them from expressing their feeling effectively (Matza et al., 2013). Here, we described the overall approaches used to establish supportive psychometric evidence of the successful PROMIS extension of measures to young children ages 1–5 years. The resulting measures from this initiative demonstrated acceptable psychometric properties, from overall measure reliabilities being greater than 0.70 to score-level-specific reliabilities providing the most precision in the "intervention required" or "clinically most significant" *T*-score range. As such, these new measures expand the horizon of the current PROMIS measurement system by extending it to younger children, thus enabling a lifespan coherent measurement system.

Each of the 12 new measures captures key developmental features based on the core concept of each domain. Going forward, these measures can be enhanced with additional item content, as clinicians, parents, and measurement experts recognize or identify new domain-specific content that would importantly supplement existing content. One of the unique advantages of item banks and calibrated scales developed using IRT is that they can quite easily accept new item content without changing the score

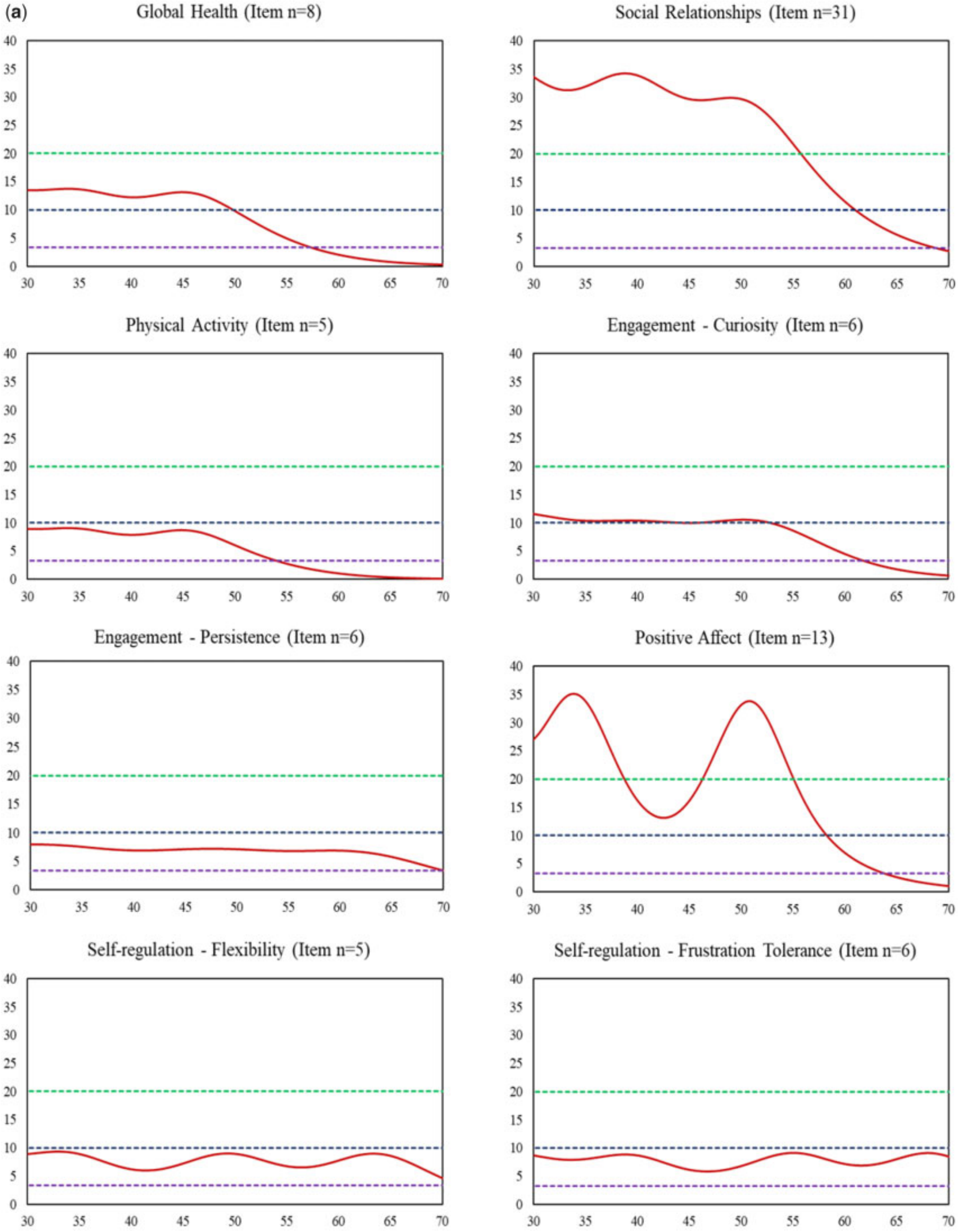


Figure 3. Information function curve of each measure (red solid line). (a) Measures with Positive Direction: Higher Scores Represent Better Health (*X*-axis: *T* score; *Y*-axis: Information function). *Note. X*-axis represents *T* scores with mean = 50 and standard deviation = 10. *Y*-axis represents information function at the measure level. An information function of 3.3, 10, and 20 is corresponding to a reliability of 0.70 (green dashed line), 0.90 (navy blue dashed line), and 0.95 (purple dashed line), respectively. (b) Measures with negative direction: higher scores represent worse health (*X*-axis: *T* scores; *Y*-axis: information function). *Note. X*-axis represents *T* scores with mean = 50 and standard deviation = 10. *Y*-axis represents information function at the measure level. An information function of 3.3, 10, and 20 is corresponding to a reliability of 0.70 (green dashed line; the top line), 0.90 (navy blue dashed line; the middle line), and 0.95 (purple dashed line; the bottom line), respectively.

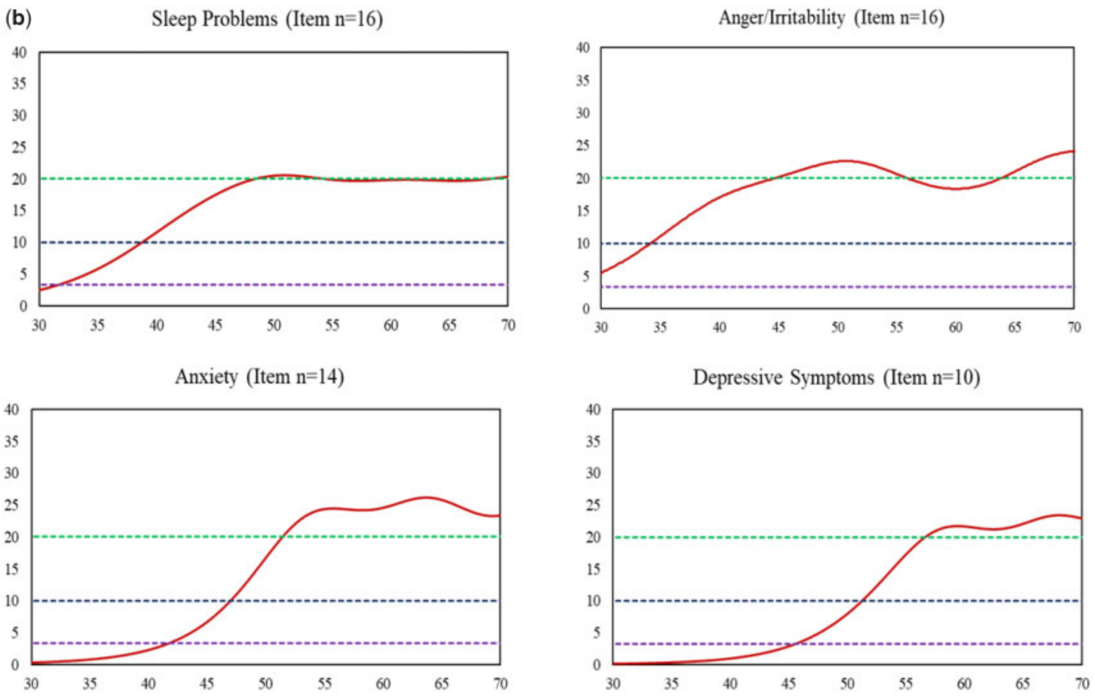


Figure 3. (Continued)

Table III. *Descriptive Statistics*

	Wave 1		Wave 2			
Domain	Mean (SD)	Minimum–maximum	Mean (SD)	Minimum–maximum	*p* value	Effect size
Higher scores, better well-being						
Global Health	47.4 (10.0)	11.4–61.7	50.0 (9.0)	15.4–61.9	<.001	0.27
Social Relationships[b]	50.1 (10.2)	13.7–63.3	49.9 (8.0)	22.6–63.3	.14	−0.02
Physical Activity	52.2 (9.7)	30.7–79.6	50.0 (8.9)	30.9–78.3	<.001	−0.23
Engagement—Curiosity	50.0 (10.2)	16.3–64.7	50.0 (9.2)	16.3–64.7	.91	0.01
Engagement—Persistence	53.2 (11.0)	15.1–70.8	50.0 (9.1)	15.1–70.8	<.001	−0.32
Positive Affect[b]	50.0 (10.3)	12.5–63.6	50.0 (9.2)	12.8–63.4	.87	−0.01
Self-Regulation—Flexibility	51.4 (11.5)	17.7–71.7	50.0 (9.2)	17.7–71.7	.01	−0.14
Self-Regulation—Frustration Tolerance	52.6 (11.8)	18.0–75.9	50.0 (9.2)	18.0–75.9	<.001	−0.25
Higher score, worse symptom						
Sleep Problems	56.5 (10.2)	31.2–88.0	49.9 (9.4)	30.1–87.7	<.001	−0.68
Anger/Irritability[b]	49.6 (10.6)	30.0–83.3	50.0 (9.4)	30.0–85.9	.49	0.03
Anxiety[b]	51.6 (10.4)	39.6–87.9	49.9 (8.9)	39.6–87.9	<.001	−0.18
Depressive Symptoms[b]	52.9 (9.9)	41.0–86.6	50.0 (8.7)	41.0–87.4	<.001	−0.31

Note. SD = standard deviation.
[a] Means comparison between wave 1 and wave 2 (reference). Scores were centered on the wave 2 sample.
[b] The SF scores were used when selected items were administered in wave 2 testing.
[c] Cohen's *d* (absolute value). Small: 0.2 (inclusive)–0.5; moderate: 0.5 (inclusive)–0.8; large: ≥0.8.

metric, after proper testing and analysis to confirm their fit. Validation against developmental gold standard direct assessment, such as task- and observation-based paradigms, will also be important for some domains. Because of the use of state-of-the-science psychometric methods, proposed new items can be linked to the current item banks and calibrated scales using IRT-based linking methods. Thus, these new measures provide a foundation which can be expanded upon, enriching item content and increasing measurement range and precision. A critical next step for real world use in the context of pediatric research will be validation for clinical populations and testing implementation within the healthcare decision making and outcome evaluation settings. The brevity and ease of use of CAT and SFs enable precise estimation of a trait and also multiple traits while simultaneously minimizing response burden to informants. This is particularly important given symptoms can coexist (e.g., a cluster of fatigue, depression, and sleep

Table IV. *Correlation Matrix of PROMIS EC Domains*

Domains	Global	Phys. Act.	Sleep[a]	Anx.[a]	Ang./Irr.[a]	Dep. Symp.[a]	Pos. Aff.[a]	Engage.—Cur.	Engage.—Per.	Self-Reg.—Flex.	Self-Reg.—Frus. Tol.	Soc. Rel.[b]
Global Health	1.00											
Physical Activity	0.08	1.00										
Sleep Problems[a]	−0.27	0.11	1.00									
Anxiety[a]	−0.32	0.10	0.47	1.00								
Anger/Irritability[a]	−0.33	0.07	0.52	0.43	1.00							
Depressive Symptoms[a]	−0.36	0.09	0.42	0.56	0.42	1.00						
Positive Affect[a]	0.41	0.11	−0.27	−0.32	−0.29	−0.41	1.00					
Engagement—Curiosity	0.42	0.18	−0.16	−0.24	−0.18	−0.32	0.50	1.00				
Engagement—Persistence	0.38	0.15	−0.22	−0.22	−0.29	−0.25	0.37	0.53	1.00			
Self-Regulation—Flexibility	0.41	0.12	−0.33	−0.32	−0.36	−0.32	0.41	0.49	0.55	1.00		
Self-Regulation—Frustration Tolerance	0.37	0.16	−0.32	−0.23	−0.45	−0.26	0.37	0.41	0.58	0.63	1.00	
Social Relationships[b]	0.46	0.11	−0.25	−0.31	−0.27	−0.39	0.53	0.53	0.50	0.50	0.43	1.00

Note. Global=Global Health; Phys. Act.=Physical Activity; Sleep=Sleep Problems; Ang./Irr.=Anger/Irritability; Dep. Symp.=Depressive Symptoms; Pos. Aff.=Positive Affect; Engage. - Cur.=Engagement—Curiosity; Engage. - Per.=Engagement—Persistence; Self-Reg.- Flex.=Self-Regulation—Flexibility; Self-Reg.- Frus. Tol.=Self-Regulation—Frustration Tolerance; Soc. Rel.=Social Relationships.

[a] 8-item short form.

[b] 6-item short form, which includes two items from each subdomain.

problem), and healthcare providers can incorporate multiple CATs or SFs into their busy clinical practices to get a better holistic understanding on the child's health to provide better targeting interventions and timely referral.

For score interpretation, U.S.-based reference values are available, due to the use of the random probability-based wave 2 sample as the centering group. Regarding administration, several options exist for SF use; in addition, CAT administration is available—a feature particularly useful for item banks that have more than 10 items (*Anger/Irritability*, *Anxiety*, *Depressive Symptoms*, *Positive Affect*, and *Social Relationships*) as CATs decrease response burden yet produce similarly precise estimates.

Our study had several limitations. First, because we did not collect any short time period time-separated data, during which child health status would be expected to remain stable, we were not able to evaluate test–retest reliability. Second, because we also did not collect any longitudinal data, during which child health status might potentially change, we were not able to evaluate responsiveness or sensitivity to change. Finally, we relied predominantly on parent report data for our measure validity analyses and evidence gathering; thus, validity data from objective sources as well as other nonparent sources were not available or analyzed.

In summary, using rigorous mixed-method approaches as described in Cella et al. (this issue; qualitative methods) and here (quantitative methods), we successfully extended PROMIS to early childhood for young children aged 1–5 years. This extension includes six-item banks ($\geq$10 items) and six calibrated scales ($<$ 10 items) measuring key domains of young children's health that closely align with PROMIS pediatric and adult physical, mental, and social health domains. All PROMIS EC measures have been translated into Spanish and evidence to support their psychometric properties is not yet available. The deliberate development of the measures for pragmatic use, including SFs and CATs and their public availability in English and Spanish (Healthmeasures.net), has them poised for rapid translation to epidemiologic research, clinical research, and clinical care.

Funding

Research reported in this publication was supported by the Environmental influences on Child Health Outcomes (ECHO) program, Office of The Director, National Institutes of Health, under Award Number U24OD023319 with co-funding from the Office of Behavioral and Social Sciences Research (OBSSR; Person Reported Outcomes Core).

Conflicts of interest: None declared.

References

Blackwell, C. K., Wakschlag, L., Krogh-Jespersen, S., Buss, K. A., Luby, J., Bevans, K., Lai, J. S., Forrest, C. B., & Cella, D. (2020). Pragmatic health assessment in early childhood: The PROMIS® of developmentally based measurement for pediatric psychology. *Journal of Pediatric Psychology*, *45*(3), 311–318.

Blumberg, S. J., Foster, E. B., Frasier, A. M., Satorius, J., Skalland, B. J., Nysse-Carris, K. L., Morrison, H. M., Chowdhury, S. R., & O'Connor, K. S. (2012, June). Design and operation of the National Survey of Children's Health, 2007. *Vital and Health Statistics*, *1*(55), 1–149.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D. J., Choi, S., Cook, K., Devellis, R., Dewalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Pilkonis, P., Revicki, D., Rose, M., ... Lai, J.-S; PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2965562/

Cella, D., Schalet, B. D., Kallen, M., Lai, J.-S., Cook, K. F., Rutsohn, J., & Choi, S. W. (2016). *PROsetta Stone® Analysis report: A Rosetta Stone for patient reported outcomes*. (Vol. 2). Department of Medical Social Sciences. Northwestern University.

Cella, D., Yount, S., Gershon, R., & Rothrock, N. (2008, October 22–25). The Patient-Reported Outcomes Measurement Information System (PROMIS): Four years in and four to go. International Society for Quality of Life Research (ISOQOL), Montevideo, Uruguay.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*(8), 1–30. http://www.jstatsoft.org/v39/i08/

Coghill, D., Danckaerts, M., Sonuga-Barke, E., & Sergeant, J, ADHD European Guidelines Group. (2009, May). Practitioner review: Quality of life in child mental health–conceptual challenges and practical choices. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *50*(5), 544–561. https://doi.org/10.1111/j.1469-7610.2009.02008.x

Cohen, S. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, *41*(6), 439–455. https://doi.org/10.1177/0146621617695522

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D-w., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143–156. https://doi.org/10.1007/s11205-009-9493-y

Halle, T. G., & Darling-Churchill, K. E. (2016). Review of measures of social and emotional development. *Journal of Applied Developmental Psychology*, *45*, 8–18. https://doi.org/10.1016/j.appdev.2016.02.003

Hu, L. T., & 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. https://doi.org/10.1080/10705519909540118

Keyes, C. L. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior*, *43*(2), 207–222. https://doi.org/10.2307/3090197

Keyes, C. L. (2010). The next steps in the promotion and protection of positive mental health. *CJNR (Canadian Journal of Nursing Research)*, *42*(3), 17–28.

Kline, R. B. (2008). *Principles and practice of structural equation modeling* (2nd edn). Guilford Press.

Lai, J.-S., Butt, Z., Zelko, F., Cella, D., Krull, K., Kieran, M., & Goldman, S. (2011). Development of a parent-report cognitive function item bank using item response theory and exploration of its clinical utility in computerized adaptive testing. *Journal of Pediatric Psychology*, *36*(7), 766–779. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3146757/

Lai, J.-S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., & Stone, A. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, *92*(10 Suppl), S20–S27. https://doi.org/10.1016/j.apmr.2010.08.033

Lai, J.-S., Crane, P. K., & Cella, D. (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Quality of Life Research*, *15*(7), 1179–1190. https://doi.org/10.1007/s11136-006-0060-6

Lai, J.-S., Nowinski, C., Victorson, D., Bode, R., Podrabsky, T., McKinney, N., Straube, D., Holmes, G. L., McDonald, C. M., Henricson, E., Abresch, R. T., Moy, C. S., & Cella, D. (2012). Quality-of-life measures in children with neurological conditions: Pediatric neuro-QOL. *Neurorehabilitation and Neural Repair*, *26*(1), 36–47. https://doi.org/10.1177/1545968311412054

Lai, J.-S., Zelko, F., Butt, Z., Cella, D., Kieran, M. W., Krull, K. R., Magasi, S., & Goldman, S. (2011). Parent-perceived child cognitive function: Results from a sample drawn from the US general population. *Child's Nervous System*, *27*(2), 285–293. http://www.ncbi.nlm.nih.gov/pubmed/20652814

Lai, J.-S., Zelko, F., Krull, K., Cella, D., Nowinski, C., Manley, P., & Goldman, S. (2014). Parent-reported cognition of children with cancer and its potential clinical usefulness. *Quality of Life Research*, *23*(4), 1049–1058. https://doi.org/10.1007/s11136-013-0548-9

Matza, L. S., Patrick, D. L., Riley, A. W., Alexander, J. J., Rajmil, L., Pleil, A. M., & Bullinger, M. (2013). Pediatric patient-reported outcome instruments for research to support medical product labeling: Report of the ISPOR PRO good research practices for the assessment of children and adolescents task force. *Value in Health*, *16*(4), 461–479.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Earlbaum Associates, Inc.

Muthen, B. O., Du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. https://www.statmodel.com/download/Article_075.pdf. Retrieved 31 July 2020.

Paller, A. S., Lai, J.-S., Jackson, K., Rangel, S. M., Nowinski, C., Silverberg, J. I., Ustsinovich, V., & Cella, D. (2021). Generation and validation of the Patient-Reported Outcome Measurement Information System Itch Questionnaire–Child (PIQ-C) to measure the impact of itch on life quality. *Journal of Investigative Dermatology*. Advance online publication. https://doi.org/10.1016/j.jid.2021.10.015.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D, PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS):depression, anxiety, and anger. *Assessment*, *18*(3), 263–283. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3153635/

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., & Cella, D, PROMIS Cooperative Group. (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*(5 Suppl 1), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Samejima, F. (1997). The graded response model. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer-Verlag.

Smees, R., Rinaldi, L. J., & Simner, J. (2020, February). Well-being measures for younger children. *Psychological Assessment*, *32*(2), 154–169. https://doi.org/10.1037/pas0000768

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *The Journal of Applied Psychology*, *91*(1), 25–39. https://doi.org/10.1037/0021-9010.91.1.25