Journal of Pediatric Psychology, 2019, 1–10
doi: 10.1093/jpepsy/jsz081
Systematic Review

PROMIS® Pediatric Depressive Symptoms as a Harmonized Score Metric

Aaron J. Kaat,[1] PhD, Michael A. Kallen,[1] PhD, MPH, Cindy J. Nowinski,[1] MD, PhD, Stacy A. Sterling,[2] DrPH, MSW, Sherrilyn R. Westbrook,[2] PhD, and John T. Peters, [2]PhD

[1]Northwestern University and [2]Kaiser Permanente Northern California

All correspondence concerning this article should be addressed to Aaron J. Kaat, PhD, Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 625 N. Michigan Avenue, Suite 2700, Chicago, IL 60611, USA. E-mail: aaron.kaat@northwestern.edu

Received April 30, 2019; revisions received July 29, 2019; accepted September 20, 2019

Abstract

Objective To conduct an evidence-based review of adolescent self-report depression measures and to demonstrate how various measures can be rescored onto a harmonized metric. **Method** Six widely used person-reported outcome measures (PROMs) were reviewed. Psychometric properties were evaluated using previously published guidance for PROMs. Next, two secondary data sources (from an outpatient behavioral health clinic and from the general population) were evaluated to harmonize scores across three of the measures. Both item response theory and equipercentile linking methods were used and compared. **Results** All six PROMs demonstrated a high evidence base for widespread use depending on the purpose of the assessments. Adolescent involvement when developing the PROM for content validity and floor or ceiling effects were the least frequent available evidence. Three of the PROMs were linked to the PROMIS® Pediatric Depressive Symptoms v2.0 (PROMIS-PedDepSx) metric. The scales were highly correlated and essentially unidimensional when aggregated. All linking methods were broadly comparable. Group-level score conversions are recommended to minimize linking bias. **Conclusions** There are a number of strong, widely used PROMs for the evidence-based assessment (EBD) of adolescent depression. However, score comparability is a concern whenever there is a proliferation of measures. Harmonized score metrics support data aggregation and re-analysis. Using four PROMs, one of which served as the scoring metric, we demonstrated the possibility of harmonized depression scores. Future directions for EBD should evaluate whether harmonized PROMs for other pediatric health domains would be useful.

Key words: adolescence; depression; linking; evidence-based assessment; PROMIS pediatric; self-report.

Introduction

Across the lifespan, major depressive disorder (MDD) and other depressive disorders are some of the most common mental health diagnoses. In preschool and middle childhood, MDD is recognizable and diagnosable, but fairly rare, with estimates varying between 1 and 3% (Egger & Angold, 2006; Klein, Torpey, Bufferd, & Dyson, 2008). Prevalence sharply rises in adolescence, with current prevalence rates between 5 and 8% and lifetime (at the end of adolescence) near 15–20% (Avenevoli, Swendsen, He, Burstein, & Merikangas, 2015; Costello, Mustillo, Erkanli, Keeler, & Angold, 2003; Klein et al., 2008). This is commensurate with rates in adulthood (Kessler et al., 2003). Compared to their peers, adolescents with chronic illnesses show higher rates of depression

(Compas et al., 2014; Ferro & Boyle, 2015; Pinquart & Shen, 2010), and among those with chronic illness, more severe depression is related to recent hospitalization (DeWalt et al., 2015).

Evidence-based assessment (EBA) of depression, then, is necessary throughout childhood, but especially so in adolescence and among individuals with chronic illness. EBA involves determining what, where, when, and how to measure important psychological constructs, one of which is depression. Where to measure depression will vary: While behavioral health clinics may be the most common place to treat depression, it is important to assess it across pediatric settings. Depression screening of adolescents aged 12–18 is recommended by the U.S. Preventive Services Task Force on an "opportunistic basis" though the American Academy of Pediatrics recommends annual screening (Siu & on behalf of the US Preventative Services Task Force, 2016). Thus primary care is a common setting for EBA of depression. Additionally, there are several pay-for-performance initiatives involving depression, including screening, monitoring, and evaluating response and remission among adolescents. These are aimed at improving healthcare (NCQA, 2015; Unützer et al., 2012), but they also provide an opportunity for integrated mental health services in a primary care setting. Behavioral health screening is often a part of integrated pediatric care, though questionnaires often implemented therein have an inadequate measurement base for that assessment purpose (Feldman, Lavigne, & Meyers, 2016). Depression assessment is also necessary beyond primary care. Depressed mood in adolescence is associated with increased risk-taking behavior and poorer compliance with medical advice (Bender, 2007; Dobbels, Decorte, Roskams, & Van Damme-Lombaerts, 2010; Gray, Denson, Baldassano, & Hommel, 2011; Katon et al., 2010). Given its high prevalence and impact on other health domains, depression screening, on an annual or more often opportunistic frequency, should be common regardless of the clinical setting.

Regarding "how" to conduct an EBA, the means will vary depending on the purpose of assessment. The gold standard for determining diagnostic status is a structured diagnostic interview. There are several such interviews, with excellent reviews on their measurement properties for depression and other behavioral health disorders (Grills & Ollendick, 2002; Klein, Dougherty, & Olino, 2005). However, diagnostic interviews have poor feasibility outside of psychiatric settings. Pediatric psychologists are often in other settings. Thus, EBA may focus on screening or routine outcome monitoring instead. In these cases, person-reported outcome measures (PROMs, also called self-report measures or rating scales) are most commonly utilized, forming a significant component of the EBA. Symptom reports about internalizing disorders, such as MDD, are best obtained from the person experiencing the symptom. While this is challenging for young children, it is the primary mode of assessment for adolescents.

However, as with other health-related quality of life domains, clinicians and researchers are faced with a myriad of depression PROMs, which could be included within an EBA (c.f., Kazdin, 2005). The Patient-Reported Outcomes Measurement Information System® (PROMIS®) initiative attempted to standardize outcome measurement for both children and adults (Irwin et al., 2010). In addition to creating new measures based on item response theory (IRT), PROMIS sought to link existing measures to a harmonized score metric—that is, a common metric which may have been derived through linking or equating studies, item-by-item harmonization, or various other statistical techniques (c.f., Bauer & Hussong, 2009; Kolen & Brennan, 2014). Regardless of the methods chosen for creating a harmonized metric, the emphasis is on the comparability of scores as opposed to the PROM by which one obtained the score. Among adults, numerous PROMs have been linked to PROMIS Depression, and the pediatric and adult versions have been linked to each other (Choi, Schalet, Cook, & Cella, 2014; Reeve et al., 2016). This article reviews common self-reported PROMs appropriate for adolescents, then demonstrates how harmonized scoring can address the proliferation of assessments.

Method

Measure Selection

Rather than a systematic review of all available EBA methods (interview, PROM, clinician-, parent-, or other informant-report) for assessing depression, we chose to conduct a conceptual review of the most widely used PROMs. They are often the first line of assessing depression, given their high feasibility and low respondent burden, though they are only one component of an EBA. We included PROMs that were (a) recommended for physician quality improvement initiatives; (b) named in the cross-cutting or disease-specific emerging measures of the *Diagnostic and Statistic Manual 5th Edition* (DSM-5; APA, 2013); (c) or were frequently used in clinical and research settings. We excluded measures which had a depression subscale but also assessed a broad range of other domains (e.g., externalizing symptoms). PROMs chosen for inclusion were the adolescent-adaptation of the Patient Health Questionnaire 9-item depression scale (PHQ-A; Johnson, Harris, Spitzer, & Williams, 2002; Kroenke, Spitzer, & Williams, 2001), PROMIS Pediatric Bank v2.0 Depressive Symptoms

(PROMIS-PedDepSx; Irwin et al., 2010), the Center for Epidemiological Studies—Depression Child version (CES-DC; Fendrich, Weissman, & Warner, 1990; Radloff, 1991), the Short Mood and Feelings Questionnaire (SMFQ; Messer, Angold, Costello, & Loeber, 1995), the Child Depression Inventory (CDI-2; Kovacs & Beck, 1977), and the Reynolds Adolescent Depression Scale (RADS; Reynolds, 1986). We recognize that the "adult" versions of the CES-D and the Beck Depression Inventory-2 overlap with the CES-DC and CDI-2 during adolescence (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961; Radloff, 1977), but have chosen to emphasize the measures adapted—and thus potentially more appropriate—for adolescents. Then, as a demonstration of harmonization, we link the PHQ-A, CES-DC, and SMFQ to the PROMIS-PedDepSx score metric.

Assessment Criteria

Unlike previous reviews, which were tasked with specific evaluative criteria, this conceptual review selected key quality indicators for adolescent depression and used established guidelines specifically developed for PROMs (Terwee et al., 2007). More comprehensive reviews of additional PROMs, semistructured interviews, and other modalities for EBA are available elsewhere (Klein et al., 2005; Siu & on behalf of the US Preventative Services Task Force, 2016; Stockings et al., 2015; Williams, O'Connor, Eder, & Whitlock, 2009).

See Terwee et al. (2007) for definitions of the reviewed psychometric properties. Evaluations of content validity were based on the original development articles (or articles describing reasons for revisions; Fendrich et al., 1990; Irwin et al., 2010; Johnson et al., 2002; Kovacs & Beck, 1977; Messer et al., 1995; Osman, Gutierrez, Bagge, Fang, & Emmerich, 2010; Reynolds, 1986). For criterion validity, the "gold" standard we used was predictive validity to any diagnostic interview or structured clinician rating.

As this is not intended as a comprehensive review, we collapsed what Terwee et al. (2007) refer to as "reproducibility," "responsiveness," and "interpretability" into "longitudinal validity," as all three refer to reliability and validity over time.

Harmonization Measures

To harmonize the chosen measures, we chose to utilize linking methods (Kolen & Brennan, 2014). We identified extant datasets which coadministered multiple high-quality PROMs: specifically, PROMIS-PedDepSx with the PHQ-A, CES-DC, or the SMFQ.

The PHQ-A is a 9-item instrument designed for use in primary care to screen for depression (Johnson et al., 2002; Kroenke et al., 2001). It is based on older diagnostic criteria for MDD. Between the PHQ-9 and the PHQ-A, the primary difference is a minor rewording of items to align with diagnostic criteria for children (e.g., adding irritability) or for developmental appropriateness (e.g., concentration on school work as opposed to employment). Scores range from 0 to 27. Although the PHQ-9 has severity bands and screening cutoffs that are well-established (Kroenke et al., 2001), more debate exists about appropriate cutoffs with the adolescent version (Richardson et al., 2010). The CES-DC is a 20-item measure designed to assess depressive symptoms in the general population (Fendrich et al., 1990; Radloff, 1991). Scores range from 0 to 60, and a score of 15 has been suggested as a possible cutoff for significant levels of depressive symptoms (Fendrich et al., 1990). The Short Mood and Feelings Questionnaire (SMFQ) is a 13-item version of a longer PROM (Messer et al., 1995). A cutoff of 8 has been proposed for a positive depression screen (Messer et al., 1995).

The harmonized metric is based on PROMIS-PedDepSx, a PROM developed using both qualitative and quantitative methodology (Irwin, Varni, Yeatts, & DeWalt, 2009; Irwin et al., 2010). For this study, only adolescent self-report was used. This item bank consists of 14 items which can be administered as a computer adaptive test or short form. PROMIS-PedDepSx items reflect the core cognitive and emotional symptoms as opposed to somatic and self-harm symptoms—a benefit when assessing individuals with chronic illness, who may exhibit somatic symptoms for reasons other than depression. Although no official cutoffs have been proposed by the PROMIS group, the American Psychiatric Association has suggested PROMIS-PedDepSx as an emerging measure, with a T-score between 55 and 60 to indicate mild, 60–70 for moderate, and over 70 to indicate severe depression (APA, 2013).

Participants

For both samples described below, demographic information is provided in Table I.

PHQ-A Linking

Linking the PHQ-A to PROMIS-PedDepSx was part of a quality improvement project comparing how the two PROMs compared in four outpatient behavioral health clinics. This study was deemed by the relevant review office to be a quality improvement project and exempt from full review by the Institutional Review Board. Potential participants were all patients between the ages of 12–17 at participating clinics being seen in outpatient Psychiatry during a 5-month period in 2016. All participants completed both measures as part of standard of care on paper-and-pencil forms, with clinical staff manually entering item-level data into the regional data repository. Deidentified data

Table I. *Sample Demographics*

		Pediatric clinical sample; $N = 674$		NIH toolbox validation sample; $N = 1015$	
Characteristic		Mean	*SD*	Mean	*SD*
Age in years		15.4	1.7	12.5	2.7
Characteristic		*N*	%	*N*	%
Gender	Male	249	37.0%	501	49.4%
	Female	425	63.0%	514	50.6%
Race[a]	White	263	39.0%	837	82.5%
	African-American	50	7.4%	118	11.6%
	Asian-American	72	10.7%	21	2.1%
	Native American	3	0.4%	21	2.1%
	Other or multiracial	192	28.5%	49	4.8%
	Missing or prefer not to report	94	13.9%		
Ethnicity	Non-Hispanic	152	22.6%	916	90.2%
	Hispanic	175	26.0%	99	9.8%
	Missing or prefer not to report	347	51.5%	0	0%
Diagnosis	Primary depression	254	37.7%		
	Other mental health concern	420	62.3%		

Note. The PHQ-A was linked using the Pediatric Clinical Sample. Some individuals completed the PHQ-A and PROMIS-PedDepSx on multiple occasions, and thus one randomly selected completion was used for linking calibration, and all available data were used for cross-validating the link. The NIH-TB Validation Study data were used for linking the CES-D Children and the SMFQ to the PROMIS-PedDepSx. Diagnostic information was not collected as part of the Toolbox Validation study.

[a]In the Clinical sample, individuals were able to choose a "multi-racial" category, resulting in a sum of 100% across racial categories, whereas in the Toolbox Validation sample, individuals were allowed to choose more than one race, allowing for a higher than 100% classifications.

were used for these analyses. There were 1,104 assessment occasions for 674 unique participants drawn from three outpatient treatment centers. Most completed only one assessment, but 242 (36%) completed multiple assessments, with the additional assessments useful for cross-validation of the link.

CES-DC and SMFQ Linking

Data for linking the CES-DC and SMFQ to PROMIS-PedDepSx were collected as part of the NIH toolbox calibration and validation study. There were 1,015 assessments completed for children and adolescents, ages 8–17 years. More details regarding this sample, including how it was collected and greater details about its demographics are in a previous publication (Pilkonis et al., 2013).

Harmonization Study Design

There are multiple options for linking scales (Dorans, 2007; Kolen & Brennan, 2014). A hybrid nonequivalent anchor test (NEAT) design was used for linking the PHQ-A and PROMIS-PedDepSx. This allows responses from two samples (the original PROMIS development sample and this sample) to be linked by a set of common "anchor" items—in this case, a custom 8-item short form. NEAT designs like this one, where anchor items are drawn from an IRT-calibrated item bank, have also been referred to as *linking to a calibrated item pool* (Kolen & Brennan, 2014). A full-bank single-group design was utilized for linking PROMIS-PedDepSx to CES-DC and SMFQ. In order to maximize comparability with the current item bank, established item calibrations were utilized where possible, thereby providing another example of linking to a calibrated item pool.

Analyses

Linking is one method for creating a harmonized metric, which is well-described elsewhere (Choi et al., 2014; Dorans, 2007; Kolen & Brennan, 2014). Briefly, in order to link separate scales, the tests should be unidimensional, and the strong assumptions for IRT should hold. Consistent with previous linking studies (Choi et al., 2014), we assessed the dimensionality of the aggregated item sets using correlations and unidimensional confirmatory factor analyses (CFA). We did not assess all IRT assumptions, as modifications to the "parent" instrument were viewed as untenable. We used several methods to evaluate whether the aggregated items were "unidimensional enough" for linking, including statistical rules of thumb and published examples (Reise, Cook, & Moore, 2014).

Then, linking was conducted using Stocking-Lord coefficients following separate IRT calibration, fixed anchor IRT calibration, and equipercentile linking with or without post-smoothing of the score distribution (for more information on these methods, see Kolen & Brennan, 2014). For individuals with multiple assessments in the PHQ-A linking, we used a randomly-selected occasion for calibration purposes. We then cross-validated the links using the superset of all available assessment occasions.

Table II. *Summary of Measurement Properties of Reviewed Adolescent Self-Report Depression PROMs*

PROM	Age range	Proposed cutoff	Content validity	Internal consistency	Criterion validity	Construct validity—convergent	Construct validity—divergent	Longitudinal validity	Floor/ceiling effects
PROMIS-PedDepSx	8–17	T-score 60 (Moderate)	+	+	0	+	?	+	−
PHQ-A	13–18	Varies	?	0	+	+	0	0	−
CES-DC	12–18	15	?	?	+	+	0	0	−
SMFQ	6–17	8	?	+	+	+	0	+	−
CDI/CDI-2	7–17	Varies	?	+	+	+	+	+	?
RADS/RADS-2	13–18	Raw 76T-score 61	?	+	+	+	0	0	?

Note. CDI = Child Depression Inventory; CES-DC = Center for Epidemiological Studies—Depression Child Version; PHQ-A = Patient Health Questionnaire—Adolescent; RADS = Reynolds Adolescent Depression Scale; SMFQ = Short Mood and Feelings Questionnaire; + = Positive; − = negative; ? = questionable design or insufficient evidence; 0 = no evidence available.

In order to determine whether the IRT or equipercentile method provided a superior harmonized metric, we used a variety of graphical and statistical techniques. We calculated the consistency/linear relationship between scores (i.e., Pearson correlations), and Krippendorff's alpha, which is an indicator of absolute agreement (Hayes & Krippendorff, 2007). Descriptive statistics were calculated for the scores on the individual measure, the differences between them, and the root mean square difference (RMSD). Then, Bland-Altman plots were produced to examine the limits of agreement between the linked and actual scores (Bland & Altman, 1986).

Results

Psychometric Properties of Reviewed PROMs

Table II summarizes the level of evidence for the six PROMs included in the review. Most PROMs had a strong evidence base. Content validity was a consistent area of poor performance; however, this may be due to the criteria used to judge it as opposed to actual poor content validity. Only the development or revision articles were considered, and the standards we chose to apply (Terwee et al., 2007) require involvement of the target population in the development of PROMs—not just testing the items in an appropriate population. While this is a current standard for person-centered assessment, precision medicine, and patient-focused drug development (Collins & Varmus, 2015; Patrick et al., 2007), it has not been a historical emphasis. Indeed, content validity for the purpose of diagnosing (as opposed to screening or quantifying) MDD, may be higher for the PROMs rated as having questionable evidence, insofar as most were developed to reflect diagnostic criteria regardless of target population perspectives.

Ratings for internal consistency reliability were generally high. All reviewed PROMs have an acceptable Cronbach's alpha or comparable statistic (Fendrich et al., 1990; Osman et al., 2010; Shemesh et al., 2005; Thompson et al., 2012; Varni et al., 2014), with the exception of the PHQ-A, for which an internal consistency coefficient could not be identified. However, the standard PHQ-9 adult version has an acceptable level (Lee, Schulberg, Raue, & Kroenke, 2007), and when the PHQ-A was aggregated with PROMIS-PedDepSx items—as described in the harmonized metric below—the combined items have a high internal consistency. Likewise, all measures except the PHQ-A and CES-DC underwent factor analyses, a necessary step for adequate rating (Terwee et al., 2007), at some point of development (Irwin et al., 2010; Messer et al., 1995; Osman et al., 2010; Thompson et al., 2012).

Criterion validity against a gold standard showed acceptable sensitivity and specificity (Fendrich et al., 1990; Johnson et al., 2002; Messer et al., 1995; Reynolds & Mazza, 1998; Shemesh et al., 2005), but as far as we are aware, sensitivity and specificity against a gold standard diagnostic interview has heretofore not been established for PROMIS-PedDepSx.

Construct validity involves evidence for both convergent and divergent validity. It is well-known that anxiety and depression are highly correlated (Lonigan, Carey, & Finch, 1994; Tortella-Feliu, Balle, & Sesé, 2010), especially by self-report, and therefore, divergence between anxiety and depression was not required in the ratings (contrary with other reviews; Klein et al., 2005). All measures showed high convergence with other measures of depression. PROMIS-PedDepSx, PHQ-A, CES-DC, and SMFQ converged with each other (see Table III and Pilkonis et al., 2013). Other studies found convergence with these and other depression PROMs, clinician ratings, or measures of other internalizing symptoms (Fendrich et al., 1990; Hughes, Gullone, & Watson, 2011; Krefetz, Steer, Gulab, & Beck, 2002; Shemesh et al., 2005; Weinberg & Klonsky, 2009). Divergent validity with domains outside of negative affectivity was less

Table III. *Evaluation of the Linking Relationships*

	PHQ-A calibration sample ($n = 674$)		PHQ-A full sample ($n = 1,104$)		CES-D children ($n = 1,015$)		SMFQ ($n = 1,015$)	
Linking evaluation	Pattern scoring IRT	Equipercentile	Pattern scoring IRT	Equipercentile	Pattern scoring IRT	Equipercentile	Pattern scoring IRT	Equipercentile
Pearson correlation	0.83	0.79	0.83	0.78	0.80	0.77	0.76	0.74
Krippendorff's alpha	0.81	0.77	0.80	0.76	0.80	0.77	0.75	0.74
Mean difference	0.12	0.21	0.39	0.52	−0.67	0.53	−0.82	0.49
SD of Differences	6.55	7.43	6.60	7.44	6.51	7.09	7.02	7.53
RMSD	6.54	7.43	6.61	7.45	6.54	7.11	7.06	7.54

Note. RMSD = Root mean square difference.

frequently assessed. PROMIS-PedDepSx and the CDI-2 were exceptions. PROMIS-PedDepSx diverged from PROMs related to physical health, but it remained highly correlated with fatigue and anger (Varni et al., 2014); the CDI-2 diverged from measures of externalizing behaviors (Kuhn, Ahles, Aldrich, Wielgus, & Mezulis, 2018).

Evidence for longitudinal use of these PROMs was also high. Many have been used in longitudinal studies already. Terwee et al. (2007) highlight a need for establishing cut points on meaningful differences. This has been done less frequently, though cross-sectional minimally-important differences have been established for PROMIS-PedDepSx (Thissen et al., 2016). Surprisingly, only PROMIS-PedDepSx, SMFQ, and CDI have clear evidence for test–retest reliability in peer-reviewed publications (DeWalt et al., 2015; Finch, Saylor, Edwards, & McIntosh, 1987; Messer et al., 1995; Varni et al., 2014). While a technical manual may have some of this information, most studies have used internal consistency as the only evidence for reliability of these PROMs (c.f., Stockings et al., 2015).

The final psychometric property that needs to be evaluated for PROMs is floor or ceiling effects. Most studies have not considered this before, and floor and ceiling rates are not reported on any of these measures in the available literature. However, the linking analyses conducted (see Supplementary Online Figures 1, 2, and 3) clearly demonstrate a floor effect for PROMIS-PedDepSx, PHQ-A, SMFQ, and CES-DC. It is likely that the CDI and RADS would also have floor effects if they were directly evaluated. This should not be surprising, as depression is a skewed trait, with many ways to get a higher depression score and few ways to be less-depressed than not depressed.

Harmonization

Classic item analysis and model fit for the CFA supported the ability to create a harmonized metric with the aggregated item sets. For all three links, the scales were highly correlated ($r = 0.81$, 0.83, and 0.79 for the PHQ-A, CES-DC, and SMFQ, respectively), and when aggregated with PROMIS had good internal consistency (Cronbach's alpha = 0.97, 0.96, and 0.96), and were broadly unidimensional (RMSEA = 0.07, 0.08, and 0.08). All linking methods—both IRT-based and equipercentile with or without smoothing—resulted in broadly comparable results.

Table III summarizes the quality of the linking relationship for pattern-based scoring on the IRT-based harmonized metric and for the harmonized metric linked through the equipercentile method with no smoothing. The IRT-based harmonized metric was superior insofar as it minimized bias (i.e., mean difference) and had a smaller linking error indexed by the RMSD. Similar to previous studies (e.g., Choi et al., 2014), we have developed sum score conversion tables based off of the IRT linking results; these are provided in Supplementary Online Appendices A–C. Supplementary Figures 1–3 graphically represent score recovery beyond the summary values in Table III. These figures are Bland-Altman plots for the agreement between the PHQ-A, CES-DC, and SMFQ scores transformed onto the PROMIS-PedDepSx metric using the sum-score conversion table with data from all assessment occasions. As is evident, agreement is best for individuals with higher depressive symptoms (though wide variability in agreement at the individual level persists), and there were floor effects on all tests.

Conclusions

Evidence Base for PROMs

All reviewed PROMS had a high evidence base. However, we focused on PROMs that did not include adult age ranges. This limited the available evidence. Far more research has evaluated the psychometric properties of the PHQ-9, BDI-2, or CES-D than their adolescent counterparts. Previous reviews have emphasized those measures or have compared them interchangeably with the adolescent version (Klein et al., 2005; Stockings et al., 2015; Williams et al., 2009). Additionally, more psychometric studies have

evaluated cultural and linguistic adaptations for translated versions. This is important but does not reflect the psychometric properties of the original instrument. Had this review considered additional forms or cultural adaptations, additional sources of evidence for their appropriateness would have been available.

A potential limitation of this EBA review is that more studies have utilized a PROM than have reported on its psychometric properties. Some of the empirical basis may have been hidden within the text as opposed to abstracts or keywords of the manuscripts. For example, the main hypotheses for Kuhn et al. (2018) related to externalizing behaviors, but they included depression, allowing for evaluation of discriminant validity. It is likely that similar evidence may have been missed for the reviewed PROMs.

One interesting finding is that very few studies have specifically evaluated the psychometric properties of these PROMs among children and adolescents with chronic illness. An exception is PROMIS-PedDepSx, which included several chronic conditions when it was being developed and validated (DeWalt et al., 2015; Hinds et al., 2013; Irwin et al., 2010; Selewski et al., 2014; Varni et al., 2014). PROMIS-PedDepSx has other benefits in chronic conditions as well. By only emphasizing cognitive symptoms, it prevents threats to construct validity in chronic conditions where somatic symptoms may be due to illness and not depression—a necessary rule-out for MDD (APA, 2013; Klein et al., 2008). PROMIS-PedDepSx will not replace semistructured interviews—indeed, more research is necessary on its sensitivity and specificity against a gold standard—but it is a prime candidate for a harmonized scoring metric.

Harmonized Depression Metrics

Among adults, use of a harmonized metric for depression has already been suggested. Several PROMs already can be scored on the PROMIS Depression harmonized metric (Choi et al., 2014). This study extended the common metric by linking the PHQ-A, CES-DC, and SMFQ to PROMIS-PedDepSx. This study had many strengths, including using a single group design where possible, which is optimal for linking (Dorans, 2007), and a hybrid NEAT design where, for practical clinical purposes, a single group design was not feasible (Kolen & Brennan, 2014). Under both designs, we anchored the calibrations for the other depression tests to the existing PROMIS-PedDepSx metric (Irwin et al., 2010), thereby supporting the broader utility of the linking relationships. Then, we derived the relationship between the scales using multiple linking methods, so that we could determine which method minimized the difference between linked and actual PROMIS-PedDepSx scores. For the PHQ-A, this was done both in the original sample and in an extended sample that included other assessment occasions. This rigorous approach supports the robustness of the results and their broad application to diverse new populations.

The PHQ-A, CES-DC, and SMFQ were broadly unidimensional when aggregated (independently) with the PROMIS-PedDepSx items. Fixed-anchor IRT co-calibration resulted in optimal linking in all cases, by minimizing the difference between linked and actual scores (an index of bias) and the variability in score recovery. The practical results from this study—presented in the Supplementary online Appendices—support a wide range of clinical applications.

Clinical Utility

EBA requires not just utilization of appropriate assessment tools, but also appropriate interpretation of those tools in light of the purpose of assessment. As shown in Table II, there are a wide range of PROMs available as a component within an EBA. For the purposes of within-person assessment (e.g., routine outcome monitoring), a clinician might choose any of these. However, if the emphasis is on between-person (i.e., group) differences, comparability between the multitude of PROMs pose a significant assessment challenge.

We have demonstrated how a harmonized metric could be formed for adolescents, addressing this challenge. This provides psychologists with the ability to choose among several high-quality PROMs within an EBA but maintain a comparable score across groups (e.g., different settings, treating psychologists, or research protocols). A future direction of this effort could be including alternative PROMs in physician quality improvement initiatives and pay-for-performance initiatives (NCQA, 2015; Unützer et al., 2012). Pediatric psychologists integrated into primary care are uniquely situated to support screening, referral, and treatment for MDD. Moving away from one PROM within these initiatives has other downstream benefits as well, such as preventing mono-method bias (Cook & Campbell, 1979; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

This study also allows a comparison of clinical cutoffs. Choi et al. (2014) compared clinical cutoffs for adult depression measures. A similar comparison can be made here. The proposed PHQ-A, CES-DC, and SMFQ cutoffs are 11, 15, and 8, respectively (Fendrich et al., 1990; Messer et al., 1995; Richardson et al., 2010). These correspond to T-scores of 59, 54, and 59 on the harmonized PROMIS-PedDepSx metric, which is approximately equivalent to the proposed values distinguishing no from mild depression (T-score = 55) and mild from moderate depression (T-score = 60). This supports the clinical utility of all of the

PROMs, but showing their coordination in identifying similar individuals on a harmonized reporting metric.

Limitations and Future Directions

This study is not without its limitations. First, a comprehensive review of all PROMs was not conducted, much less a review of all other assessment modalities. When conducting an EBA, PROMs are only one piece of a comprehensive assessment (Klein et al., 2005). Second, when demonstrating the harmonized metric, local independence and differential item functioning were not considered. Ideally, for IRT-based linking to proceed, all of the statistical assumptions for IRT should be met (Dorans, 2007), but modification to an existing scale was not considered reasonable when the original PROMs have been firmly established. Third, we did not evaluate subpopulation invariance for the proposed links. This has been previously suggested and has been done in some studies but continues to be rare in health outcomes research.

A final limitation is more of a reasonable caution: Harmonized scoring metrics derived as part of this study are most appropriate for group-level data. Individual-level scores have a much larger error (i.e., measurement error associated with the IRT-based scoring plus linking error). Converting group-level data will minimize the errors for comparison purposes, but researchers or clinicians converting instruments during ongoing data collection should be aware of the increased error (and thus reduced reliability) of the linked individual scores. However, if the purpose of an EBA is individual screening or monitoring, clinicians have a range of high-quality options available to them.

Future researchers should continue to build upon this and similar linking studies. More research is necessary on the optimal way to link individual-level scores. In order to encourage widespread adoption of a harmonized reporting metric, it will be necessary for future researchers to demonstrate the concordance of scores using linked and actual measure, and also for derived measures, such as pay-for-performance quality improvement initiatives.

Funding

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health (NIH) under award numbers U2CCA186878 and 1RC4CA157236 (PI Cella), by the Office of the Director of the NIH under award number 1U24OD023319 (MPIs Gershon & Cella), and through the Blueprint for Neuroscience Research and the Office of Behavioral and Social Sciences Research within the NIH under contract No. HHS-N-260-2006-00007-C (PI Gershon). Additional data collection was conducted as part of a clinical service receiving no additional funding by Kaiser Permanente Northern California. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplementary Data

Supplementary data can be found at: https://academic.oup.com/jpepsy.

References

APA. (2013). *American Psychiatric Association. Diagnostic and statistical manual (DSM-5)* (5th edn). Washington: American Psychiatric Association.

Avenevoli, S., Swendsen, J., He, J.-P., Burstein, M., & Merikangas, K. R. (2015). Major depression in the National Comorbidity Survey–Adolescent Supplement: Prevalence, correlates, and treatment. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*, 37–44.e32.

Bauer, D. J., & Hussong, A. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*, 101–125.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

Bender, B. G. (2007). Depression symptoms and substance abuse in adolescents with asthma. *Annals of Allergy, Asthma & Immunology*, *99*, 319–324.

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*, 307–310.

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*, 513.

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, *372*, 793–795.

Compas, B. E., Desjardins, L., Vannatta, K., Young-Saleme, T., Rodriguez, E. M., Dunn, M., … Gerhardt, C. A. (2014). Children and adolescents coping with cancer: Self- and parent reports of coping and anxiety/depression. *Healthy Psychology*, *33*, 853.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry*, *60*, 837–844.

DeWalt, D. A., Gross, H. E., Gipson, D. S., Selewski, D. T., DeWitt, E. M., Dampier, C. D., … Varni, J. W. (2015). PROMIS® pediatric self-report scales distinguish subgroups of children within and across six common pediatric chronic health conditions. *Quality of Life Research*, *24*, 2195–2208.

Dobbels, F., Decorte, A., Roskams, A., & Van Damme-Lombaerts, R. (2010). Health-related quality of life, treatment adherence, symptom experience and depression in

adolescent renal transplant patients. *Pediatric Transplantation*, *14*, 216–223.

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, *16(1)*, 85–94.

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*, 313–337.

Feldman, M., Lavigne, J. V., & Meyers, K. M. (2016). Systematic Review: Classification accuracy of behavioral screening measures for use in integrated primary care settings. *Journal of Pediatric Psychology*, *41*, 1091–1109.

Fendrich, M., Weissman, M. M., & Warner, V. (1990). Screening for depressive disorder in children and adolescents: Validating the center for epidemiologic studees depression scale for children. *American Journal of Epidemiology*, *131*, 538–551.

Ferro, M. A., & Boyle, M. H. (2015). The impact of chronic physical illness, maternal depressive symptoms, family functioning, and self-esteem on symptoms of anxiety and depression in children. *Journal of Abnormal Child Psychology*, *43*, 177–187.

Finch, A. Jr, Saylor, C. F., Edwards, G. L., & McIntosh, J. A. (1987). Children's Depression Inventory: Reliability over repeated administrations. *Journal of Clinical Child Psychology*, *16*, 339–341.

Gray, W. N., Denson, L. A., Baldassano, R. N., & Hommel, K. A. (2011). Treatment adherence in adolescents with inflammatory bowel disease: The collective impact of barriers to adherence and anxiety/depressive symptoms. *Journal of Pediatric Psychology*, *37*, 282–291.

Grills, A. E., & Ollendick, T. H. (2002). Issues in parent-child agreement: The case of structured diagnostic interviews. *Clinical Child and Family Psychology Review*, *5*, 57–83.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89.

Hinds, P. S., Nuss, S. L., Ruccione, K. S., Withycombe, J. S., Jacobs, S., DeLuca, H., … DeWalt, D. A. (2013). PROMIS pediatric measures in pediatric oncology: Valid and clinically feasible indicators of patient-reported outcomes. *Pediatric Blood & Cancer*, *60*, 402–408.

Hughes, E. K., Gullone, E., & Watson, S. D. (2011). Emotional functioning in children and adolescents with elevated depressive symptoms. *Journal of Psychopathology and Behavioral Assessment*, *33*, 335–345.

Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J.-S., … DeWalt, D. A. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Res*, *19*, 595–607.

Irwin, D. E., Varni, J. W., Yeatts, K., & DeWalt, D. A. (2009). Cognitive interviewing methodology in the development of a pediatric item bank: A patient reported outcomes measurement information system (PROMIS) study. *Health and Quality of Life Outcomes*, *7*, 3.

Johnson, J. G., Harris, E. S., Spitzer, R. L., & Williams, J. B. W. (2002). The patient health questionnaire for adolescents: Validation of an instrument for the assessment of mental disorders among adolescent primary care patients. *Journal of Adolescent Health*, *30*, 196–204.

Katon, W., Richardson, L., Russo, J., McCarty, C. A., Rockhill, C., McCauley, E., … Grossman, D. C. (2010). Depressive symptoms in adolescence: The association with multiple health risk behaviors. *General Hospital Psychiatry*, *32*, 233–239.

Kazdin, A. E. (2005). Evidence-based assessment for children and adolescents: Issues in measurement development and clinical application. *Journal of Clinical Child & Adolescent Psychology*, *34*, 548–558.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., … Wang, P. S. (2003). The epidemiology of major depressive disorder results from the National Comorbidity Survey Replication (NCS-R). *JAMA*, *289*, 3095–3105.

Klein, D. N., Dougherty, L. R., & Olino, T. M. (2005). Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *34*, 412–432.

Klein, D. N., Torpey, D. C., Bufferd, T. J., & Dyson, M. W. (2008). Depressive disorders. In T. P. Beauchaine, & S. P. Hinshaw (Eds.), *Child and adolescent psychopathology* (pp. 477–509). Hoboken, NJ: Wiley.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: methods and practices* (3rd ed.). New York, NY: Springer.

Kovacs, M., & Beck, A.T. (1977). An empirical-clinical approach toward a definition of childhood depression. In J. G. Schultebrant & A. Raskin (Eds.), *Depression in Childhood: Diagnosis, Treatment, and Conceptual Models* (pp. 1–2(5)). New York, NY: Raven Press.

Krefetz, D. G., Steer, R. A., Gulab, N. A., & Beck, A. T. (2002). Convergent validity of the Beck Depression Inventory-II with the Reynolds Adolescent Depression Scale in psychiatric inpatients. *Journal of Personality Assessment*, *78*, 451–460.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*, 606–613.

Kuhn, M. A., Ahles, J. J., Aldrich, J. T., Wielgus, M. D., & Mezulis, A. H. (2018). Physiological self-regulation buffers the relationship between impulsivity and externalizing behaviors among nonclinical adolescents. *Journal of Youth and Adolescence*, *47*, 829–841.

Lee, P. W., Schulberg, H. C., Raue, P. J., & Kroenke, K. (2007). Concordance between the PHQ-9 and the HSCL-20 in depressed primary care patients. *Journal of Affective Disorders*, *99*, 139–145.

Lonigan, C. J., Carey, M. P., & Finch, A. (1994). Anxiety and depression in children and adolescents: Negative affectivity and the utility of self-reports. *Journal of Consulting and Clinical Psychology*, *62*, 1000.

Messer, S. C., Angold, A., Costello, E. J., & Loeber, R. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: Factor composition and structure across development. *International Journal of Methods in Psychiatric Research*, *5*, 251–262.

NCQA. (2015). *HEDIS 2016: Healthcare effectiveness data and information set, Volume 2, technical update.*

Washington DC: Nactional Committee for Quality Assurance.

Osman, A., Gutierrez, P. M., Bagge, C. L., Fang, Q., & Emmerich, A. (2010). Reynolds adolescent depression scale-second edition: A reliable and useful instrument. *Journal of Clinical Psychology*, *66*, 1324–1345.

Patrick, D. L., Burke, L. B., Powers, J. H., Scott, J. A., Rock, E.P., Dawisha, S., … Kennedy, D.L. (2007). Patient-reported outcomes to support medical product labeling claims: fDA perspective. *Value in Health*, *10*, S125–S137.

Pilkonis, P. A., Choi, S. W., Salsman, J. M., Butt, Z., Moore, T. L., Lawrence, S. M., … Cella, D. (2013). Assessment of self-reported negative affect in the NIH Toolbox. *Psychiatry Research*, *206*, 88–97.

Pinquart, M., & Shen, Y. (2010). Depressive symptoms in children and adolescents with chronic physical illness: An updated meta-analysis. *Journal of Pediatric Psychology*, *36*, 375–384.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401.

Radloff, L. S. (1991). The use of the Center for Epidemiologic Studies Depression Scale in adolescents and young adults. *Journal of Youth and Adolescence*, *20*, 149–166.

Reeve, B. B., Thissen, D., DeWalt, D. A., Huang, I.-C., Liu, Y., Magnus, B., … Tulsky, D. S. (2016). Linkage between the PROMIS® pediatric and adult emotional distress measures. *Quality of Life Research*, *25*, 823–833.

Reise, S. P., Cook, K. F., & Moore, T. M. (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 13–40). New York: Routledge.

Reynolds, W. M. (1986). A model for the screening and identification of depressed children and adolescents in school settings. *Professional School Psychology*, *1*, 117.

Reynolds, W. M., & Mazza, J. J. (1998). Reliability and validity of the Reynolds Adolescent Depression Scale with young adolescents. *Journal of School Psychology*, *36*, 295–312.

Richardson, L. P., McCauley, E., Grossman, D. C., McCarty, C. A., Richards, J., Russo, J. E., … Katon, W. (2010). Evaluation of the Patient Health Questionnaire-9 Item for detecting major depression among adolescents. *Pediatrics*, *126*, 1117–1123.

Selewski, D. T., Massengill, S. F., Troost, J. P., Wickman, L., Messer, K. L., Herreshoff, E., … Gipson, D. S. (2014). Gaining the Patient Reported Outcomes Measurement Information System (PROMIS) perspective in chronic kidney disease: A Midwest Pediatric Nephrology Consortium study. *Pediatric Nephrology*, *29*, 2347–2356.

Shemesh, E., Yehuda, R., Rockmore, L., Shneider, B. L., Emre, S., Bartell, A. S., … Newcorn, J. H. (2005). Assessment of depression in medically ill children presenting to pediatric specialty clinics. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*, 1249–1257.

Siu, A. L., & on behalf of the US Preventive Services Task Force. (2016). Screening for depression in children and adolescents: U.S. Preventive Services Task Force recommendation statement screening for depression in children and adolescents. *Annals of Internal Medicine*, *164*, 360–366.

Stockings, E., Degenhardt, L., Lee, Y. Y., Mihalopoulos, C., Liu, A., Hobbs, M., & Patton, G. (2015). Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders*, *174*, 447–463.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., … de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*, 34–42.

Thissen, D., Liu, Y., Magnus, B., Quinn, H., Gipson, D. S., Dampier, C., … DeWalt, D. A. (2016). Estimating minimally important difference (MID) in PROMIS pediatric measures using the scale-judgment method. *Quality of Life Research*, *25*, 13–23.

Thompson, R. D., Craig, A. E., Mrakotsky, C., Bousvaros, A., DeMaso, D. R., & Szigethy, E. (2012). Using the Children's Depression Inventory in youth with inflammatory bowel disease: Support for a physical illness-related factor. *Comprehensive Psychiatry*, *53*, 1194–1199.

Tortella-Feliu, M., Balle, M., & Sesé, A. (2010). Relationships between negative affectivity, emotion regulation, anxiety, and depressive symptoms in adolescents as examined through structural equation modeling. *Journal of Anxiety Disorders*, *24*, 686–693.

Unützer, J., Chan, Y.-F., Hafer, E., Knaster, J., Shields, A., Powers, D., & Veith, R. C. (2012). Quality improvement with pay-for-performance incentives in integrated behavioral health care. *American Journal of Public Health*, *102*, e41–e45.

Varni, J. W., Magnus, B., Stucky, B. D., Liu, Y., Quinn, H., Thissen, D., … DeWalt, D. A. (2014). Psychometric properties of the PROMIS® pediatric scales: Precision, stability, and comparison of different scoring and administration options. *Quality of Life Research*, *23*, 1233–1243.

Weinberg, A., & Klonsky, E. D. (2009). Measurement of emotion dysregulation in adolescents. *Psychological Assessment*, *21*, 616.

Williams, S. B., O'Connor, E. A., Eder, M., & Whitlock, E. P. (2009). Screening for child and adolescent depression in primary care settings: A systematic evidence review for the US Preventive Services Task Force. *Pediatrics*, *123*, e716–e735.