



IMPACT MEASURES TOOL: SCORING SYSTEM EVIDENCE GUIDE

Last Updated August 24, 2022

Copyright 2020-2021 University of Oregon. All rights reserved. The text of and illustrations on the EC PRISM® and IMPACT Measures Tool® websites (“UO Materials”) are licensed by the University of Oregon under a Creative Commons Attribution–Share Alike 3.0 Unported license (“CC-BY-SA”) except for any third-party materials and measures used by permission or under fair use. The University of Oregon, Center for Translational Neuroscience, EC PRISM®, IMPACT Measures Tool®, and their respective logos are trademarks of the University of Oregon (“UO Names & Trademarks”) registered in the United States and other countries. If you modify or create derivatives of the UO Materials, remove any UO Names & Trademarks from the UO Materials, except for your use of the UO Names & Trademarks to identify the changes you have made to the UO Materials as required to fairly attribute the authors of the original work under the CC-BY-SA license, by stating the following: These works have been modified from works created and owned by the University of Oregon, Center for Translational Neuroscience, <https://ctn.uoregon.edu> & <https://ecmeasures.instituteforchildsuccess.org>.

Table of Contents

<u>ABOUT EC PRISM.....</u>	<u>3</u>
<u>ABOUT THE IMPACT MEASURES TOOL®.....</u>	<u>3</u>
<u>ABOUT THIS EVIDENCE GUIDE</u>	<u>4</u>
<u>IMPACT SCORING CATEGORIES.....</u>	<u>4</u>
<u>COST (10 POINTS MAXIMUM).....</u>	<u>5</u>
<u>USABILITY (10 POINTS MAXIMUM)</u>	<u>8</u>
<u>TECHNICAL MERIT (10 POINTS MAXIMUM).....</u>	<u>16</u>
<u>CULTURAL RELEVANCE (10 POINTS MAXIMUM)</u>	<u>22</u>
<u>REFERENCES.....</u>	<u>28</u>

About EC PRISM

The EC PRISM team at the Institute for Child Success (ICS) provides tools and resources to early childhood organizations such as the IMPACT Measures Tool®, as well as our individualized measurement and evaluation services to support organizations in the areas of early childhood program design, implementation, evaluation, and scale. We aim to increase acceptance, understanding, and capacity of measurement and evaluation within the field.

Our IMPACT suite of tools and resources focus on programs, organizations, and system-level innovations which improve service delivery for all children and families. Due to historical oppression and institutional racism, children and families of color disproportionately experience social and economic adversity. We approach our work through a historical lens that acknowledges the root causes underlying the conditions and inequities experienced by families. We are dedicated to understanding the historical, systemic nature of racism. We aim to center this historical perspective as well as the voices of the communities which we are invited to serve.

About the IMPACT Measures Tool®

The IMPACT Measures Tool® is an online database of early childhood and parenting measures, designed to meet the needs of community-based organizations, academic institutions, health care systems, philanthropic and policy-based organizations, child care systems, and more. **The IMPACT Measures Tool® is not a publisher or distributor of measures.** It was developed to provide innovative solutions to three key challenges:

1. Assist the early childhood field in unifying their approach to measuring early childhood development.
2. Support organizations and programs in evaluating their own impact.
3. Help organizations in determining which measures are best suited to their specific contexts and needs.

Our free tool allows anyone interested in early childhood measurement and evaluation to search, compare, and access measures scored on four key categories of usability, cost, cultural relevance, and technical merit.

About this Evidence Guide

This document provides a detailed outline of the scoring system of the IMPACT Measures Tool®. It includes a description of scoring categories and the research basis behind these categories. It also outlines scoring criteria and points by subcategory, which may be useful for measure developers and publishers, and others interested in a detailed breakdown of how points are assigned. For a more general overview of the IMPACT scoring system and the process for approaching each scoring category, please refer to our [Scoring Guidebook](#).

Measure Selection and Review Process

When the IMPACT Measures Tool® was launched, it featured a wide variety of early childhood and parenting measures to provide information to meet the measurement needs of individuals and organizations in various sectors across the field, such as early childhood educators, health care professionals, policymakers, and more. We provide information on many key measures that are widely used in the field, as well as newer and lesser-known measures which may suit the needs of early childhood organizations. Over time, our team of measurement experts has steadily incorporated additional measures onto our website, with a particular focus on equity and cultural relevance. For more information, or to submit a request for a measure to be scored and added to our website, please email us at ecmeasures@instituteformchildsuccess.org.

Once measures are selected, they are reviewed and scored on four categories of usability, cost, cultural relevance, and technical merit based on their user manuals or validation study. Measures that have these resources are scored according to the criteria in this guide.

IMPACT Scoring Categories

Each scoring category has a maximum of 10 total points, and a minimum of 0 points. Each category is composed of subcategories, with scores that add up to the overall total score for the category. The IMPACT Measures Tool® does

not delineate a specific cutoff for each category to indicate a “poor” measure, since organizations have different priorities and needs for measurement and should select a measure based on their unique context. The one exception is the minimum score for technical merit which describes the basic psychometric properties of a measure. If a measure does not meet this minimum score of 3, this indicates that a measure either does not have sufficient information to score this category or the evidence available is minimal and unsatisfactory.

Cost (10 points maximum)

Description	The cost scoring category refers to the affordability and accessibility of the measure.
Background	<p>The cost of a measure varies widely depending on what type of measure it is, how it is administered, what materials are required, the scale at which it will be used, what training is required, and more (Fernald et al., 2017). Measure cost influences overall feasibility in context as well as buy-in from stakeholders regarding the relative costs and benefits of use (Glover & Albers, 2007).</p> <p>Despite a consensus that ideal measures are inexpensive (Fernald et al., 2017; Hunsley & Mash, 2007), a comprehensive picture of the range of costs for different types of assessment tools is not available. Our team therefore took a bottom-up, data-driven approach to determining cost scoring levels by examining the price distribution for 100 pilot measures that represented a wide range of measure types and domains. The initial scoring breakdown was later revised based on further analysis to capture greater variation.</p>

Subcategory	Price	
Description	<p>This subcategory accounts for the actual monetary cost of the measure. Note: the higher the price of a measure, the lower the measure scores (this subcategory aims to capture the affordability of the measure.) Free measures score highest in this category. For at-cost measures, the primary focus includes required materials to learn and administer the measure, such as the starter kit or user manual. Additional costs, such as those related to licensure, subscription, and training, are also considered. There are a wide variety of factors affecting the cost of a measure, with no standard approach across measures. For example, some app-based measures require a one-time purchase of an app whereas others require purchase of a user license and administrator training alongside the app download. Measures requiring purchase of material kits or user manuals to administer often score lowest in this category.</p>	
Scoring criteria	<p>Does the measure require the following items? If so, how much does each cost?</p> <ul style="list-style-type: none"> • Starter kit • Electronic version • Subscription & subscription renewal • License • Training/Certification • Scoring reports <p>All components <i>required</i> for measure administration are summed for Cost score. Note that many measures have materials which are optional but not required. These are not included when determining the Price score.</p>	
	Subcategory: Price	Score

Scoring calculation	\$400	0
	\$200 - \$399	2
	\$1 - \$199	4
	Free	6

Subcategory	Access
Description	<p>This subcategory accounts for the measure’s level of accessibility. At-cost measures do not receive any points in this category. Measures score high if they are immediately downloadable and approved for use by the developer. Measures score lower in this category if they are only available to review, if they require an account login, or are only accessible through a published research article.</p> <p>For example, some measures are free and downloadable for immediate approved use, whereas other measures may only be available as a review-only copy. Other free measures may require the creation of an account to access the measure.</p>
Scoring criteria	<p>How is the measure accessed?</p> <ul style="list-style-type: none"> • At-cost • Free: Only a research article exists; must contact the author/developer for access • Free: Review copy only • Free: Available via a free online account • Free: Available by filling out a contact form • Free: Can be directly downloaded from measure website • Free: Can be directly downloaded from a third-party site

Scoring calculation	Subcategory: Access	Score
	At-cost	0
	Free: Only a research article exists; must contact the author/developer for access	
	Free: Review copy only	
	Free: Available via a free online account	2
	Free: Available by filling out a contact form	
	Free: Can be directly downloaded from measure website	4
	Free: Can be directly downloaded from a third-party site	

Scoring Calculation: Overall Cost Score

Price Subscore (max 6) + Access Subscore (max 4) = Cost Score (max 10)

Usability (10 points maximum)

Description	The usability score is designed to reflect practical considerations in measure use, including the ease to administer, score, and interpret a measure and its data. Four subcategories capture this information: training ease, time to complete, interpretation ease, and administration format.
Background	Measures with high usability present minimal burden to administer, score, and interpret (Glasgow & Riley, 2013; Youngstrom et al., 2017).

Time is a key factor for determining the burden placed on measure administrators and respondents (Glasgow & Risley, 2013). Ideal early childhood measures can be administered quickly, but many are time-intensive (Fernald et al., 2017). Short surveys and those with the option of being administered and scored electronically may minimize the time required for administration and increase flexibility of use (Glasgow & Riley, 2013), while many interview formats are likely to be more time-consuming (Youngstrom et al., 2017). Training needed to administer a measure can also vary; observational measures and direct assessments tend to require more extensive training to administer reliably and accurately interpret results, whereas training requirements for surveys tends to be minimal (Fernald et al., 2017). In the IMPACT scoring system, measures score higher in Usability by being brief (based on duration provided by the developer or estimated based on number of items), offering multiple modes of administration (i.e., paper and electronic), and requiring minimal training (based on the materials/processes one would need to use to administer the measure).

Interpretability of results – in essence, being able to derive qualitative meaning from quantitative scores (Terwee et al., 2007) – is also critical for pragmatic, real-world use. Norms or cutoffs provided by measure developers can promote “clinical utility” (Hunsley and Mash, 2018) by facilitating actionable interpretation of results such as decision-making regarding treatment (Holly et al., 2019; Stanick et al., 2019; Youngstrom et al., 2017). Other forms of information can also aid interpretation; for example, measures of classroom quality should “be accompanied by sufficient information about next steps for improving program quality, along with the resources, time and intentional support necessary to implement that improvement” (Brooks et al., 2022). Therefore, the IMPACT system assesses not only whether norms or cutoffs are

	provided but also whether supplemental materials are available to aid score interpretation.
--	---

Subcategory	Training Ease
Description	<p>This subcategory of usability assesses the level of difficulty to learn to administer the measure based on the materials required (e.g., toys or standardized instructions). Specifically, the more materials that are required to learn how to use during measure administration, the lower a measure scores in this subcategory. For example, questionnaires typically do not require any materials other than the survey items provided to the respondent, and therefore these measures score high in this subcategory – there is no training required to administer surveys. In contrast, direct assessments often require the administrator to use specific materials such as a stimulus book or toys during administration of the measure with the child. For this reason, direct assessments score low in this category given the effort required to learn to administer the measure using these materials. At this time, specific training offered by measure developers is not included in the calculation of a measure’s training ease score. Our team reached this decision as it is often very difficult to determine whether training is required to administer a measure; often training is available for purchase from a developer or publisher but is not required to administer and/or this requirement may not be enforced. In addition, variability in administrator training makes it difficult to standardize across measures.</p>
Scoring criteria	<p>Does the measure require the following materials or processes?</p> <ul style="list-style-type: none"> • Stimulus book

	<ul style="list-style-type: none"> Standardized use of materials Detailed coding description Video Audio Standardized instructions None or minimal coding None of the above are required 	
Scoring calculation	Criteria	Score
	Stimulus book	0
	Standardized use of materials	
	Detailed coding description	
	Video	
	Audio	1.5
	Standardized instructions	
	None or minimal coding	
	None of the above are required	3

Subcategory	Time to Complete: Time
Description	The time section of the time to complete subcategory refers to the time needed to complete or administer the measure. Typically, this is reported by the developer as a range (i.e. 5-10 minutes) or average (i.e., 10 minutes). Note, this does not include additional coding or scoring time.
Scoring criteria	<ul style="list-style-type: none"> Minutes required to complete the measure*. <p>*If the developer reports a range, the upper end of the range informs score. If this information is missing, and</p>

	measure is a survey, the range is estimated based on the number of items (10-20 sec/item) (Couper & Peterson, 2017). If this information is missing and the measure is not a survey, the score is either estimated based on available information if feasible, or our team emails the developer to inquire.	
Scoring calculation	Criteria	Score
	16+ min	0
	6-15 min	1
	5 min or less	2

Subcategory	Time to Complete: Materials	
Description	The materials section of the time to complete subcategory refers to the materials that must be set up and used for measure administration.	
Scoring criteria	<ul style="list-style-type: none"> • Stimulus book • Standardized use of materials • Video • Audio • Standardized instructions • None of the above are required 	
Scoring calculation	Criteria	Score
	Stimulus Book	0
	Standardized use of materials	
	Video	
	Audio	1

	Standardized instructions	
	None of the above are required	2

Subcategory	Time to Complete: Scoring	
Description	The scoring section of the time to complete subcategory refers to how long it takes to score the measure. This is often not stated explicitly by developers and therefore is inferred based upon characteristics of the measure that influence scoring processes.	
Scoring criteria	<ul style="list-style-type: none"> • Stimulus book • Standardized use of materials • Coding required • Manual scoring • Subscales • Free automatic scoring 	
Scoring calculation	Criteria	Score
	Stimulus book	0
	Standardized use of materials	
	Coding required	
	Manual scoring	1
	Subscales	
	Free automatic scoring	2

Subcategory	Scoring Interpretation
-------------	------------------------

<p>Description</p>	<p>This usability subcategory accounts for how easy it is to interpret the results of a measure following administration. Specifically, direct assessments, surveys, and screening tools score high in this category if they are norm- or criterion-referenced and include free supplemental materials to guide the interpretation process (e.g., an infographic). Some measures are not norm-referenced and do not provide any supplemental materials to aid in the interpretation of the measure results. In contrast, other measures provide downloadable files that include guidance on analysis and/or interpretation of results.</p> <p>Observations and interview score high in this category when they provide a detailed description of the coding process, which facilitates a clearer understanding of the codes and measure results. Specifically, some measures provide a very minimal description for their coding procedures, whereas other measures lay out in detail coding domains, dimensions, and/or indicators of observed behavior.</p>	
<p>Scoring criteria</p>	<ul style="list-style-type: none"> • Coding required: code definitions are not provided or provide limited detail • Coding required: code definitions are detailed • Norms and/or cutoffs are provided to aid scoring interpretation • Free supplemental materials are available to aid scoring interpretation 	
<p>Scoring calculation for observations and interviews</p>	<p>Criteria</p>	<p>Score</p>
	<p>None or minimal code descriptions</p>	<p>0</p>
	<p>Detailed code descriptions</p>	<p>2</p>
	<p>None or minimal code descriptions AND norms/cutoffs are provided</p>	<p>2</p>
	<p>No norms/cutoffs</p>	<p>0</p>

Scoring calculation for surveys, screening tools, and direct assessments	Norms/cutoffs are provided, without free supplemental materials	1
	Free supplemental materials are available, but norms/cutoffs are not provided	1
	Norms/cutoffs are provided, and free supplemental materials are available	2

Subcategory	Administration Format	
Description	<p>This subcategory of usability considers the number of options for administration of a measure (in person, electronically, or both) and whether administration of the measure requires an internet connection.</p> <p>Note: Many publishers offer electronic data systems for <i>scoring and tracking results</i>. This does not influence their administration format score. For a measure to score points for having an electronic version, it must be possible for the measure to be <i>administered</i> electronically (for example, an app-based direct assessment or an online survey).</p>	
Scoring criteria	<ul style="list-style-type: none"> • In person/paper administration available • Electronic administration available • Internet required/not required 	
Scoring calculation	Criteria	Score
	One format option (In-person/paper or electronic)	0

	Two format options (In-person/paper AND electronic)	0.5
	Internet required	0
	Internet not required	0.5

Usability Subscore: Time to Complete*

Time + Materials + Scoring/6*4 = Time to Complete Subscore

Note: This calculation converts the time (max 2), materials (max 2), and scoring (max 2) components of the time to complete subscore into a subscore with a maximum of 4 points (see equation below). This weights the time to complete more heavily in the overall usability score, reflecting the importance of time burden for ease of use.

Usability Calculation: Overall Usability Score

Time to Complete Subscore (max 4) + Training Ease (max 3) + Scoring Interpretation (max 2) + Administration Format (max 1)

Technical Merit (10 points maximum)

Description	The purpose of our technical merit score is to assess the technical quality of the measurement, which is reflected in the quality of the data it produces. The two main components of this score are validity (the degree to which a measure is accurate) and reliability (the degree to which a measure is consistent). A third component we
--------------------	---

	<p>consider is the presence/absence of norming data, or norms.</p>
<p>Background</p>	<p>The standards for assessing measurement quality come from the field of psychometrics. This field is nuanced, and many aspects of quality assessment remain ambiguous or disputed. Our team triangulated the perspectives of many developers and experts in order to design a custom approach to determining measure quality with regard to validity, reliability and norms.</p> <p>Validity refers to whether or not a measurement tool is measuring what it is intended to measure. Understanding exactly what a measurement tool is intending to measure, and identifying exactly what a measure is actually measuring, involves conducting statistical comparisons. These comparisons can demonstrate whether the data produced by the measurement tool are aligned how they are expected to appear. In our score, peer-reviewed parameters are set for assessing validity across five subtypes, with a sixth for screening tools: content (Youngstrom, Meter, Frazier, Hunsley, Prinstein, Ong, & Youngstrom, 2017), convergent/divergent (Youngstrom et. Al, 2017), internal structure (Matsunaga, 2010; Rios & Wells, 2014), concurrent characteristic (Holly, Fenley, Kritikos, Merson, Abidin, & Langer, 2019), predictive (Mislevy & Rupp, 2012; Zimmerman, Klusmann, & Hampe, 2017), sensitive/specificity (Macy, 2012). These subtypes were each given a score, and these scores were combined into an overall validity score</p> <p>Reliability refers to the consistency of a measure in relation to certain variables. Three types of reliability were considered: test-retest reliability, interrater reliability, and internal consistency. The properties of the measure itself determine which type(s) of reliability are applicable. Test-retest statistics consider how reliable a measure is over time. We considered the amount of time that passed</p>

	<p>between the first and second test administrations, and the calculated reliability statistic (Hunsley & Mash, 2007). Inter-rater reliability statistics capture the consistency of different raters when using a measure that requires ratings, and these statistics are only applicable for measures that employ a rating system. Internal consistency captures how well items within a domain relate to each other. The most common ways of reporting internal consistency include Cronbach’s alpha, ordinal alpha, and coefficient omega (Henson, 2001). Measures that report higher forms of reliability receive higher scores in our system.</p> <p>We define the norms subscore according to whether the developers provided the means and standard deviation for the score/s of the measure. If means and standard deviation are provided, users can calculate where certain scores fall in respect to the norming sample.</p>
--	---

Subcategory	Norms	
Description	The means and standard deviation for the score/s of the measure are provided by the developers. This section was included so that, if means and standard deviation are provided, users—if they choose—can calculate where certain scores fall in respect to the norming sample.	
Scoring criteria	Developers must provide means and standard deviation for their measures.	
Scoring calculation	Criteria	Score
	None	0
	Mean	.5

	Mean and Standard Deviation	1
--	-----------------------------	---

Subcategory	Validity				
Description	<p>Validity refers to whether a measurement tool is measuring what it is intended to measure. Understanding exactly what a measurement tool is intending to measure, and identifying exactly what a measure is actually measuring, involve conducting statistical comparisons. These comparisons can demonstrate whether the data produced by the measurement tool align with how they are expected to look.</p>				
Scoring criteria	<ul style="list-style-type: none"> • Content validity: How well the construct is represented in the measure overall • Convergent/divergent validity: Whether a measure relates to another measure of a similar construct and whether a measure does NOT relate to another measure of a different construct • Internal structure validity: How each item of a measure mathematically reflect the theorized overall construct being measured • Concurrent characteristic validity: The extent to which a measure relates to specific characteristics (e.g., age) • Predictive validity: The extent to which a measure predicts specific future outcome(s) • Screener sensitivity/specificity: The ability of the measure to identify the individuals who need further assessment (i.e., is truly at risk) and identify those who are NOT at risk (i.e., no false positives) 				
	<table border="1"> <thead> <tr> <th>Criteria</th> <th>Score</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> </tr> </tbody> </table>	Criteria	Score		
Criteria	Score				

Scoring calculation	Content validity: Definitions (none, named, brief, extensive)	0, 1/3, 2/3, 1
	Content validity: Consulting experts (none, authors are experts, mentions with minimal detail, describes in detail)	0, 1/3, 2/3, 1
	Convergent/divergent validity (varies based on statistic type)	0 – 1
	Internal structure validity (varies based on statistic type)	0 – 1
	Concurrent characteristic validity (varies based on statistic type)	0 – 1
	Predictive validity (varies based on statistic type)	0 – 1
	Screeener sensitivity/specificity (varies based on statistic type)	0 – 1

Subcategory	Reliability
Description	<p>Reliability refers to the consistency of a measure with respect to certain variables. Three types of reliability were considered:</p> <p>Test-retest reliability. Test-retest statistics consider how reliable a measure is over time. We considered two factors: (i) how much time passed between the first and second test administrations, and (ii) the calculated reliability statistic.</p> <p>Inter-rater reliability. Inter-rater reliability statistics capture the consistency of different raters when using a measure that requires ratings, and these statistics are only applicable for measures that employ a rating system.</p>

	Internal consistency captures how well items within a domain relate to each other. For many measures, especially questionnaires, the entire measure is often divided into various domains. We call these domains “subscales”. The overall score would pertain to all the items on the assessment, whereas the subscale scores would be restricted to the items belonging in each of the vocabulary, grammar, or reading comprehension subscales.	
Scoring criteria	<ul style="list-style-type: none"> • Test-retest reliability: How stable a measure is with respect to time • Inter-rater reliability: How stable a measure is with respect to observer • Internal consistency: How stable individual components of a measure are with respect to the other components 	
Scoring calculation	Criteria	Score
	Test-retest reliability (varies based on statistic type)	0 – 1
	Inter-rater reliability (varies based on statistic type)	0 – 1
	Internal consistency (varies based on statistic type)	0 – 1

Technical Merit Calculation: Norms Subscore

Value assigned for data reported = Norms Subscore

Technical Merit Calculation: Validity Subscore

(4 * (8 * (Average of applicable reliability components: convergent/divergent validity, internal structure, concurrent characteristic validity, predictive validity, screening tool sensitivity/specificity) + content validity))/10 = Validity Subscore

Technical Merit Calculation: Reliability Subscore

(Average of applicable reliability components: test-retest value, inter-rater, internal consistency) * 5 = Reliability Subscore

Technical Merit Calculation: Overall Technical Merit Score

Norms Subscore + Validity Subscore + Reliability Subscore = Overall Technical Merit Score

Cultural Relevance (10 points maximum)

<p>Definition</p>	<p>The purpose of our cultural relevance category is to identify whether there is evidence that a measure is generalizable to diverse cultural contexts, or if the measure successfully targets a specific cultural context intentionally. There are four main components to our cultural relevance scoring system: (1) demographics, (2) method bias, (3) item bias & psychometric group differences, and (4) norming recency. Each of these subcomponents are weighted with a number of points to contribute to the overall cultural relevance score.</p>
--------------------------	---

<p>Background</p>	<p>High cultural relevance scores indicate that the developers of the measure provide some evidence that their measure was applicable to diverse populations (see below) or that their measure was specifically tailored for a precisely defined demographic population reflecting best practice in current cultural research and assessment (Henrich et al., 2010). Measures also score higher if they provide evidence that they considered diverse populations and/or integrated input from community members in the process of developing the measure. Unfortunately, many measures are tested using homogenous norming samples, and information about norming samples is often not thoroughly reported, and so the cultural relevance scores on average tend to be lower than the technical merit scores. Furthermore, only a fraction of measures actually test for bias across different cultural groups (Rodriguez et al., 2021) or actively integrate community input in measure development processes (Bogart et al., 2021).</p>
--------------------------	--

Subcategory	Demographics
<p>Description</p>	<p>The demographics subscore reflects both the composition of the norming sample and the detail to which the characteristics of the norming sample were reported. We review the sample size to confirm that the measure was tested on a sufficiently large group of individuals.</p>
<p>Scoring criteria</p>	<ul style="list-style-type: none"> • Age • Gender • Race/ethnicity • Socioeconomic status: high, middle, low

	<ul style="list-style-type: none"> • Linguistic diversity: multilingual, bilingual, monolingual • Geographic region • Urbanicity: urban, suburban, rural 	
Scoring calculation	Criteria	Score
	Age (not reported, reported)	0 or 1
	Gender (not reported, not representative, representative, targeted)	0, 1, 2
	Race/ethnicity (not reported, not representative, representative, targeted)	0, 1, 2
	Socioeconomic status (not reported, not representative, representative)	0, 1, 2
	Linguistic diversity (not reported, one language, multiple language groups)	0, 1, 2
	Geographic region (not reported, not representative, representative, targeted)	0, 1, 2
	Urbanicity (not reported, 1 type, 2 or more or targeted)	0, 1, 2

Subcategory	Method Bias
Description	The method bias subscore assesses whether the developers of a measure included community members in the process of developing the measure or took diverse populations into consideration while developing the measure. This subscore requires evidence that the developers consulted with represented communities

	during the development process or considered diverse populations while developing the items.	
Scoring criteria	<ul style="list-style-type: none"> • Community • Diverse Populations 	
Scoring calculation	Criteria	Score
	Community Not Involved, Diverse Population Not Considered	0
	Community Not Involved, Diverse Population Considered	1
	Community Involved, Diverse Population Not Considered	1
	Community Involved, Diverse Population Considered	2

Subcategory	Item Bias and Psychometric Group Differences	
Description	If a measure is tested on a diverse sample, it is possible to show statistically whether the measure is biased with respect to demographic dimensions, such as age, gender, race/ethnicity, socioeconomic status, geographic region, urbanicity, and language. Measures were assigned higher scores if they were able to statistically demonstrate that the measure was unbiased with respect to these dimensions.	
Scoring criteria	<ul style="list-style-type: none"> • Item bias • Statistical result 	
	Criteria	Score

Scoring calculation	Item bias (not reported, conducted and statistical result not reported, conducted and statistical result reported)	0, 1, 2
	Psychometric differences (varies based on statistic type)	0, 0.5, 1, 2

S

Subcategory	History	
Description	Scientific advancements push and change the field of measurement over time. The norming recency subscore gives a measure credit if it was originally validated in the past twenty years, or if the measure was developed before the year 2000 and has an updated validation studies.	
Scoring criteria	<ul style="list-style-type: none"> Originally validated after 2000 Validation study updated after 2000 	
Scoring calculation	Criteria	Score
	Validated before 2000	0
	Validated since 2000	1

Cultural Relevance Calculation: Demographics Subscore

(Average of demographic characteristics: gender, race/ethnicity, socioeconomic status, linguistic diversity, geographic region, urbanicity)/13*2 = Demographics Subscore

Cultural Relevance Calculation: Method Bias Subscore

Value assigned for data reported = Method Bias Subscore

Cultural Relevance Calculation: Item Bias and Psychometric Group Differences Subscore

**If no item bias is reported, put all weight in psychometric differences*

**If no psychometric differences are reported, put all weight in item bias*

*$$\left(\left(\frac{\text{sum of applicable item bias}}{8} \right) * 5 + \left(\frac{\text{sum of applicable psychometric group differences}}{8} \right) * 5 \right) / 2 = \text{Item Bias and Psychometric Group Differences Subscore}$$*

Cultural Relevance Calculation: History Subscore

Value assigned for data reported = History Subscore

Cultural Relevance Calculation: Overall Cultural Relevance Score

$$\text{Demographics Subscore} + \text{Method Bias Subscore} + \text{Item Bias and Psychometric Group Differences Subscore} + \text{History Subscore} = \text{Overall Cultural Relevance Score}$$

References

Bogard, K., Ortiz-Cortes, V., Taylor, S., Jackson, R., & Belmonte, R. (2021). An exploratory approach to defining and measuring child health and well-being with parents and grandparents.

Brooks, J.L., Gayl, C.L., & Wernstedt-Lynch, C. (2022.) Measuring the Quality of Early Learning Environments: A guide to evaluating ideal learning environments for young children. Washington, DC: Trust for Learning.

Fernald, L. C., Prado, E., Kariger, P., & Raikes, A. (2017). A toolkit for measuring early childhood development in low and middle-income countries.

Glasgow, R. E., & Riley, W. T. (2013). Pragmatic measures: what they are and why we need them. *American Journal of Preventive Medicine*, 45(2), 237-243.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of school psychology*, 45(2), 117-135.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and evaluation in counseling and development*, 34(3), 177-189.

Holly, L. E., Fenley, A. R., Kritikos, T. K., Merson, R. A., Abidin, R. R., & Langer, D. A. (2019). Evidence-base update for parenting stress measures in clinical samples. *Journal of Clinical Child & Adolescent Psychology*, 48(5), 685-705.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annu. Rev. Clin. Psychol.*, 3, 29-51.

Hunsley, J., & Mash, E. J. (2008). Developing criteria for evidence-based assessment: An introduction to assessments that work. *A guide to assessments that work*, 2008, 3-14.

Macy, M. (2012). The evidence behind developmental screening instruments. *Infants & Young Children*, 25(1), 19-61.

Matsunaga, M. (2010). How to Factor-
28

Version Date: 8/24/22

Analyze Your Data Right: Do's, Don'ts, and HowTo's. *International journal of psychological research*, 3(1), 97-110.

Mislevy, J. L., & Rupp, A. A. (2012). Predictive validity. *Encyclopedia of research design*. Thousand Oaks: SAGE Publications, Inc, 1077-8.

Rodriguez, V. J., La Barrie, D. L., Zegarac, M. C., & Shaffer, A. (2021). A Systematic Review of Parenting Scales Measurement Invariance/Equivalence of by Race and Ethnicity: Recommendations for Inclusive Parenting Research. *Assessment*, 10731911211038630.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, 26(1), 108-116.

Stanick, C. F., Halko, H. M., Nolen, E. A., Powell, B. J., Dorsey, C. N., Mettert, K. D., ... & Lewis, C. C. (2021). Pragmatic measures for implementation research: development of the Psychometric and Pragmatic Evidence Rating Scale. *Translational behavioral medicine*, 11(1), 11-20.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., ... & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology*, 60(1), 34-42.

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein, M. J., Ong, M. L., & Youngstrom, J. K. (2017). Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science and Practice*, 24(4), 331-363.

Zimmermann, S., Klusmann, D., & Hampe, W. (2017). Correcting the predictive validity of a selection test for the effect of indirect range restriction. *BMC Medical Education*, 17(1), 1-10.