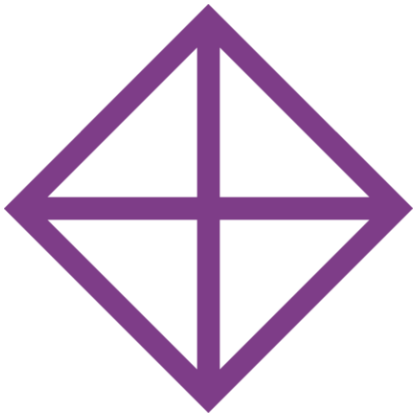


IMPACT Measures Tool: Scoring Guidebook

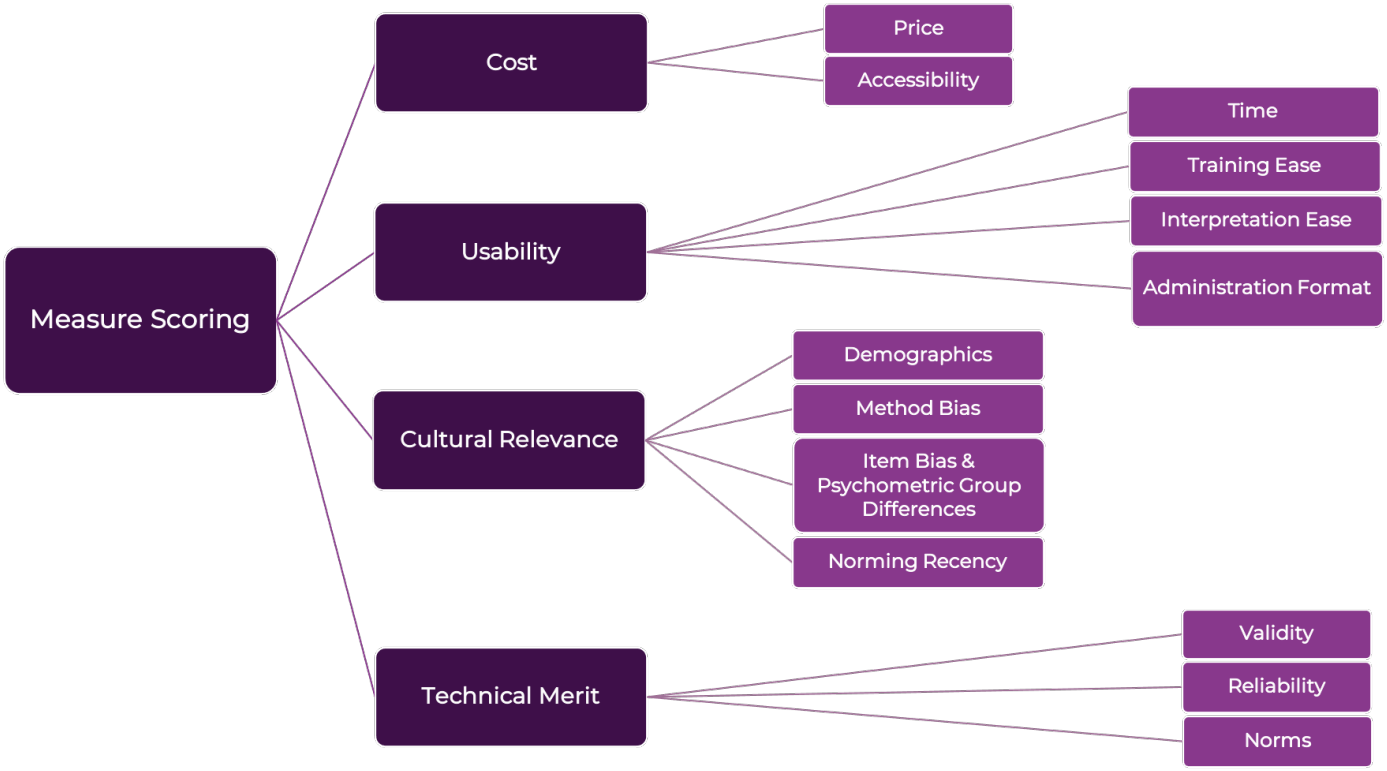
Last updated September 22, 2021

Table of Contents

Introduction

This document outlines how the four categories– cost, usability, cultural relevance (CR), and technical merit (TM)— assessment scores were calculated for the IMPACT Measures Tool. In our ongoing refinement of the IMPACT scoring system, we are continuously researching the information provided by measure developers so as to incorporate these details into how a measure is scored by our system.

Scoring Diagram



Development of Scoring System

Prior to calculating the scores for each measure, existing literature was selected to use in assessing each measure. In many cases, this was a technical manual created by measure developers. In other cases, we used a peer-reviewed research article that collected data to validate the measure and introduce it to the field (validation study). Experts in the field, such as Drs. Stephanie De Anda and Aaron Kaat, were also consulted in the development of the scoring system. All statistics that contributed to our measure scores came from developer websites alongside these references, which are listed on our website.

I. Cost (10 points overall)

Overview: The cost scoring category refers to the actual price of the measure as well as how accessible the measure is. Specifically, information collected for each measure (e.g., the cost of a starter kit) are converted to low, medium, or high scores in each subcategory depending on how the measure data meet criteria. Subscores within each category are then converted to scores for each of the cost categories, which then makeup the total score for cost overall.

A. Price (6 points)

This category accounts for the actual monetary cost 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. The higher the price, the lower the measure scores on cost.

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.

B. Accessibility (4 points)

This category of cost applies specifically to free measures, accounting for the measure's level of accessibility (i.e., measures at cost do not receive any points in this category). Specifically, 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.

For example, some measures are free and downloadable for immediate approved use, whereas other measures may only be available to as a review only copy. Other free measures may require the creation of an account to access the measure.

II. Usability (10 points overall)

Overview. In the IMPACT scoring system, the usability category refers to how easy it is to administer the measure, as well as how easy it is to analyze and interpret the results. Specifically, information about each measure (e.g., the number of minutes it takes to complete the measure) are converted to low, medium, or high scores in each subcategory depending on how the measure data meet the criteria. Subscores within each category are then converted to scores for each of the four usability categories, which then makeup the total score for usability overall. The four usability categories described in more detail below are: **time, training ease, interpretation ease, and electronic administration.**

A. Time (4 points)

The time category within usability accounts for three elements of measure administration: the actual time it takes to complete the measure itself, the materials needed to setup and use during administration (e.g., toys), and the process required to calculate the results following administration of the measure (e.g., conversion to standardized scores). Measures score high in the time category if they are short, do not require standardized use of materials, and calculated results are provided. This category is worth the most

points because we have found that, when considering usability, individuals are frequently concerned first with how long the measure will take.

For example, a measure could receive a high score in time to complete if it only takes five minutes to administer, but that same measure could receive a low score in both the materials and scoring calculation categories if it requires audio equipment and a coding process in order to calculate results.

In a small number of cases, developers did not directly report the time to complete the measure. In these cases, we calculated the time to complete the measure through the number of items reported by the developer based on an average of 10-20 seconds per question (Couper & Peterson, 2017). It is important to note that this approach does not yet take reading level or question type into account.

B. Training Ease (3 points)

This category 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 and interact with during measure administration, the lower a measure scores in this category.

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 category – 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. We made this decision because it is often very difficult to determine whether training is actually required to administer a measure; often training is available for purchase from a developer or publisher but is not actually required to administer. In addition, variability in administrator training makes it difficult to standardize across measures.

C. Interpretation Ease (2 points)

This usability category 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). Specifically, some measures are not norm-referenced and do not provide any supplemental materials to aid in the interpretation of the measure results. In contrast, some measures provide downloadable files that include guidance on analysis and/or interpretation of results.

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.

D. Administration Format (1 point)

This category of usability involves the format the measure is administered in as well as equipment requirements. Measures that require an internet connection are scored lower than those that do not. For example, the Early Years Toolbox does not require the internet to use their app-based measure whereas the NIH Toolbox does. Measures also receive a higher score in this category if they are offered in multiple administration formats. For example, the ASQ-3 is offered in both paper and electronic formats whereas the Early Years Toolbox is only offered electronically.

III. Cultural Relevance (10 points total)

Overview. The purpose of our cultural relevance assessment 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. Unfortunately, many measures are only 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 (less than 1 out of 10) measures actually test for bias across different cultural groups. High cultural relevance scores indicate that the developers of the measure provide some evidence that their measure was applicable to diverse populations or that their measure was specifically tailored for a precisely defined demographic population.

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. We are currently weighing input from outside experts to continue to develop the components of the cultural relevance score. Each component is listed below:

A. Demographics (2 points)

Our demographics subscore reflects both the diversity of the norming sample and the detail to which the characteristics of the norming sample that were reported. We looked for demographic information along the dimensions of age, gender, race/ethnicity, socioeconomic status, linguistic diversity, geographic region, and urbanicity. We also look for the sample size to ensure the measure was tested on a sufficiently large group of cases.

B. Method Bias (2 points)

Our method bias subscore assesses whether the developers of a measure included community members or had diverse populations in mind during the development of the measure. This subscore requires evidence that the developers went above and beyond simply testing the measure on a sample with diverse demographic characteristics, but actually considered these populations while developing the items and consulted their communities during the development process. This process of community involvement is important for preventing bias and ensuring that measures intended for a given community reflect the values and culture of that community. Measures that clearly state their strategy for this inclusion get credit here.

C. Item Bias & Psychometric Group Differences (5 points)

If a measure is tested on a diverse sample, it is possible to show statistically whether or not the measure is biased with respect to

demographic dimensions, such as age, gender, race/ethnicity, socioeconomic status, geographic region, urbanicity, and language. Measures got higher scores if they were able to statistically demonstrate that the measure was unbiased with respect to these dimensions. Most measures do not report these statistics, or only report them for a small number of these dimensions. Many measures do not have a large or diverse enough norming sample to generate these statistics.

D. Norming Recency (1 point)

Our norming recency subscore gives a measure credit if it was originally validated in the past twenty years. Measures that were originally developed before the year 2000 need to have an updated validation study in order to get credit here.

IV. Technical Merit (10 points total)

Overview. The purpose of our technical merit score is to assess the technical quality of the measurement. The technical quality of a measurement tool is reflected in the quality of the data it produces. The data produced by a measurement tool should have certain statistical properties, which show us whether the measure is "**valid**" and/or "**reliable**". (1) Validity is the degree to which a measure is **valid** or "accurate", and (2) reliability is the degree to which a measure is **reliable** or "consistent", and these are the two main components of our technical merit assessment. A third component we consider is the presence/absence of norming data, or (3) **norms**. These scoring components are described in detail below.

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. Consequently, our team triangulated the perspectives of many developers and experts to design a custom approach to determining measure quality (see References).

A. Validity (4 points)

What is validity? Validity refers to whether or not a measurement tool is measuring what it is intended to measure. A measure that is intended to measure language abilities will look very different from a measure intended to measure math abilities. Understanding exactly what a measurement tool is intending to measure, and identifying exactly what a measure is actually

measuring, ultimately involves conducting statistical comparisons. These comparisons can demonstrate whether the data produced by the measurement tool are how they are expected to look.

Expectations about measurements will inherently rely on existing theories about the topic we are measuring. If results look the way we expect, this evidence supports the validity of our theory and measurement tool. In this sense, the validity of a measurement tool relies both on whether the topic of focus (e.g., parent stress) is theoretically valid, and, if it is theoretically valid, whether the tool actually does a good job of measuring it. Our validity calculation aims to capture these higher-level questions.

As a final note, it is useful to consider what validity is *not*. The validity of a measure does not speak to the correct use of a measure. For example, a researcher might have a perfectly valid thermometer, but if they try to use it to measure hormones, they will always fail. This is not a problem that has to do with the thermometer, but rather the fact that it is being used to measure something it was not designed to measure. The validity of a measure only extends to the description of the measure's intended use, and only applies to populations that resemble the population the measure was tested on.

How do we calculate validity? The validity portion of our technical merit assessment score is divided into 5 subtypes of validity (with a 6th subtype for measures that are screening tools): (a) content validity, (b) convergent/divergent validity, (c) internal structure, (d) concurrent validity, (e) predictive validity, and (f) specificity/sensitivity (if a screening tool). Together, content, convergent and divergent validity are considered construct validity. These subtypes were each given a score, and these scores were combined into an overall validity score.

a) *Content validity.* Content validity refers to the opinions of experts about a measurement quality. We assessed content validity by (i) considering how the measure developers describe what their tool is measuring and (ii) considering how experts in the field view the measurement tool. For (i), the clarity of the description contributes to a higher score, whereas if the purpose of the measure is unknown or not described clearly, this contributes to a lower score. For (ii), measures that have been peer-reviewed and/or widely cited get a higher score. These parameters are based on peer-reviewed criteria for measurement assessment (Youngstrom, Meter, Frazier, Hunsley, Prinstein, Ong, & Youngstrom, 2017).

b) *Convergent/divergent validity.* Convergent and divergent validity refers to whether a measurement compares to established measurements in a way that is expected. Our convergent/divergent validity score is calculated by considering statistical results that compare data from the measure at hand with data from other known measurement tools (Youngstrom et al., 2017). If the results are aligned with what is expected by theory, then they are taken into account when computing the convergent/divergent validity subscore.

As an example, if a developer is designing a measurement of anxiety, they might compare their anxiety scores with depression scores in the same population. Because anxiety and depression are known to sometimes occur together in the same patient, one would expect a correlation greater than 0 between these scores, but because anxiety and depression also occur independently without the other, the correlation will be substantially less than 1. In other words, our theoretical framework for understanding what anxiety and depression are informs what we expect the comparison of the scores to look like. In this example, if the actual statistical comparison of the scores aligned with expectations, this would contribute to a higher convergent/divergent validity subscore. The nature of the comparison - whether it is between measures that are similar, different, or somewhere in between - will influence how the statistics are translated into a subscore.

In practice, these types of comparisons are reported using various statistics, including p-values, correlation coefficients, beta weights from linear regression models, t-values, F-values, etc. We established thresholds to convert the statistical values into a subscore.

c) *Internal structure validity.* Internal structure refers to whether items in a measure relate to each other as expected by theory. Our assessment of internal structure validity depends on the results of a particular kind of analysis, called factor analysis (Matsunaga, 2010; Rios & Wells, 2014). Internal structure is only relevant for measurements that have multiple items (such as a survey) and does not apply to most observational measures. Measures that have items that relate to the other items according to the developer's theoretical predictions about this structure will have higher scores.

d) *Concurrent characteristic validity.* Concurrent characteristics refer to how a measurement relates to characteristics of the measured population, such as age or gender (e.g., Holly, Fenley, Kritikos, Merson, Abidin, & Langer, 2019). In many cases, a measure score will vary with age. For example, language ability changes drastically between ages one and three. Therefore, showing that a language abilities measure is statistically related to age as expected would speak to the validity of that measure. A measure might be expected to relate to characteristics other than age, such as gender, race, geographic region, or clinical condition. In order to be considered a concurrent characteristic, the characteristic must be measured at the same time as the measure score.

Because concurrent validity is focused on potential associations or differences between groups with regards to the measure, many types of statistical analyses can be performed to establish concurrent validity. For example, researchers can perform t-tests for the above-mentioned example if they only separated the sample into two groups (those with and without dyslexia diagnosis). However, if they decide to divide the sample into more than two groups, they might decide to perform an analysis of variance (ANOVA) with post-hoc tests to examine potential subgroup differences. In another example, if the developers want to examine age-related differences with the reading score (presumably older children can do better), they can run a correlation or regression analysis if they treat "age" as a continuous variable. We employed thresholds for statistics to convert the statistical values into a subscore.

e) *Predictive validity.* Predictive validity refers to how well a measure score predicts a future outcome measure (Mislevy & Rupp, 2012; Zimmerman, Klusmann, & Hampe, 2017). The future outcome can be any number of things, including future grades, income, or clinical diagnosis. If a measure score can be used to predict a future outcome, we use this as evidence for the validity of the measure. Measures that were better at predicting future outcomes received higher scores.

f) *Sensitivity/specificity.* Sensitivity and specificity statistics are a standard way of describing the effectiveness of a screening tool (Macy, 2012). Screening tools must predict a gold standard with high sensitivity and specificity in order to be considered valid. Sensitivity is the ability of a

measure to correctly identify a gold standard diagnosis (true positive rate). Specificity is the ability of a measure to correctly identify a gold standard non-diagnosis (true-negative rate).

Although not currently included in our scoring system, we recognize that what should be considered as acceptable, good, or excellent for sensitivity or specificity may depend on the circumstances and the conditions. For example, for a screening tool where the consequences of failing to identify a child as at risk far outweigh those of falsely identifying a child as at risk, the bar for "acceptable" for specificity may be lower.

B. Reliability (5 points)

Reliability refers to the consistency of a measure with respect to certain variables. Three types of reliability were considered: (a) how stable a measure is with respect to time (**test-retest reliability**), (b) how stable a measure is with respect to observer (**interrater reliability**), and (c) how stable individual components of a measure are with respect to the other components (**internal consistency**). The properties of the measure itself determine which type(s) of reliability are applicable.

a) *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.

The longer the span between the test and the retest, the lower we would expect their correlation to be. Therefore, the magnitude of the correlation needs to be higher for a test-retest that was conducted within a shorter span compared to one with a longer time span. This system is adapted and modified from Hunsley & Mash (2007)'s system where the only considerations for a high, moderate, or low score would be the time frame that passed. Unfortunately, the time frame they devised for the three scores were meant for adult populations and not very suitable for early childhood measures. We therefore modified the scoring system to consider both magnitude and time frame in determining the "strength" of the correlation.

b) *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. If a measure relies on a rater, then there is an assumption that the measure is consistent regardless of who the rater is. Even if raters have different genders, races, education levels, or training, the expectation is that the raters should be reliable. Measures that have higher inter-rater reliability receive higher scores in our assessment.

c) *Internal consistency*. For many measures, especially questionnaires, the entire measure is often divided into various domains. We call these domains "subscales". For example, a language measure can contain items on vocabulary, grammar, and reading comprehension. The overall score would pertain to all of 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. 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 high internal consistency statistics receive a higher score in our scoring system.

C. Norms (1 point)

We define the "norms" subscore according to whether the developers provided the means and standard deviation for the score(s) of the measure. The premise behind including such a section is that, if means and standard deviation are provided, users—if they so choose—can calculate where certain scores fall in respect to the norming sample. Most measure developers provide means and standard deviation for their measures, so almost all measures get full points for this subscore.

Conclusion

The IMPACT Measures Tool is the culmination of a great deal of work across team members in the development of the scoring system, data collection and entry as well as website development. We continue to research measure information provided by developers to account for additional details and complexity into our scoring system. If you have any questions or concerns, please contact ecprism@uoregon.edu.

References

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.

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

Matsunaga, M. (2010). How to Factor-Analyze Your Data Right: Do's, Don'ts, and How-To'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.

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

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.