Evaluating the Measurement Quality of Social and Emotional Learning (SEL) Assessments

A four-part process for evaluating the technical quality of the SEL assessment

BUROS
CENTER FOR TESTING

with support of the Spencer Foundation

www.buros.org/sel

What is the purpose of this guide? The purpose of this guide is to prepare assessment users to know what questions to ask regarding the availability of information and empirical evidence that may support the intended interpretation and use of an SEL assessment for their student population. The extent to which the technical evidence for an SEL assessment addresses these questions will assist informed selection decisions that will translate into more valid interpretations and uses. Share suggestions and comments about their usability and guidance with Dr. Jessica L. Jonson at jjonson@buros.org. A full version of the guidebook can be found at http://buros.org/sel.

Who are the intended users of this guide? This guide is for educators tasked with the selection and use of an SEL assessment. Ideally, a group of educators and experts with relevant insights about the content to be assessed and the use of the resulting information to guide action will be involved in the evaluation. Although specialized psychometric expertise can be helpful, this guide was written for assessment users who may not be experts in the technical details of assessment development.

What type of interpretations and uses does this guide address? This guide is applicable to situations where SEL is measured to provide feedback and improve instruction and programs. The guide is not intended for situations where SEL assessments are being used for accountability or in consequential decision making at a group or individual level. Consequential decisions at an individual student level would involve measuring student learning to screen or diagnose students in need of additional services or intervention or to identify students with a mental health concern. If an SEL assessment will be used for these types of consequential decisions, educators should consult with school district professionals who have received appropriate training and who hold the licenses or certifications necessary to conduct clinical evaluations of children for mental health or special education intervention, mental health diagnosis or special education classification. These professionals should also be well-versed in the tenets of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014). If an SEL assessment will be used to make high-stakes decisions about a school or program, it is highly recommended that evidence be carefully scrutinized with the assistance of someone with psychometric expertise and an understanding of the context in which the decisions will be made.

What do you need to know to use this guide? The importance and relevance of different types of technical evidence are based on what SEL competencies you are wanting to measure; how you intend on using assessment results; your local setting and student population; and the format of the assessment, including how it is administered and scored. For each consideration, the guide outlines what documentation the assessment developer should provide, what the test user should do with that documentation, some explanation and examples, and what to do if an SEL assessment does not meet the consideration. Information needed to answer these questions can be found on assessment developers' websites, technical manuals, administration and scoring manuals or even in published technical evaluations of SEL assessments. It is likely that most SEL assessments will not meet all considerations in this guide. The guide was written to address best practices rather than common practices so users should weigh the pros and cons of using an SEL assessment that may not meet one or more considerations. If an SEL assessment does not meet one or more considerations listed in this guide, it still might be appropriate to use that assessment if caution is used in interpretation and use unless the assessment will be used for high stakes decisions about individuals, schools or programs.

This guide for evaluating the measurement quality of an SEL assessment has four parts and all four parts should be applied when evaluating the technical quality of the SEL assessment.

1. Does the assessment effectively measure the SEL competencies of interest?
2. Does the SEL assessment provide credible evidence for your intended uses?
3. Is the SEL assessment relevant for your students and your setting?
4. Does the SEL assessment address issues related to administration, scoring, and the assessment format?

Evaluating the Measurement Quality of Social and Emotional Learning (SEL) Assessments

1. *Does the assessment effectively measure the SEL competencies of interest?*

Does the assessment effectively measure the SEL competencies of interest?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Clearly identify and define which SEL competencies the assessment measures.	Determine if SEL competencies of interest align with the SEL competencies measured by the assessment.	In order to determine whether an assessment will measure the SEL competencies of interest, those competencies must be stated in measureable terms that not only identify the competency of interest but also what a student will know, do, and/or understand as a result of achieving the SEL competency. SEL assessment should measure not only the competency of interest but also how students are expected to express that competency. • SEL competencies of interest could be more general (e.g. intrapersonal or interpersonal skills) or more specific (e.g. growth mindset, self-efficacy, collaborative problem solving). • Expression of SEL competencies might also differ such as demonstrating awareness (e.g. mindsets, knowledge, beliefs) or applying skills (e.g. learned abilities). For example, if students should demonstrate problem-solving skills, the assessment should measure how students use and apply those skills not whether they are aware of the importance of those skills.	If a measure addresses none of the specific or general SEL competencies of interest or very few, find another assessment. If the measure addresses some but not all SEL competencies of interest, look for a more comprehensive measure or a second measure to supplement information gathered. If an SEL assessment does not provide a clear description of SEL competencies measured, do a formal review of items/tasks to make your own determination or find another assessment that does measure the SEL competencies of interest.

Does the assessment effectively measure the SEL competencies of interest?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Identify why intended respondents for the assessment are in the best position to assess students' SEL competencies.	Consider whether the respondent for the assessment is the best source for assessing the SEL competencies of students in the local population.	If SEL competencies of interest involve attitudes, beliefs, or growth mindsets, respondents could be students reporting on their own SEL competencies. If the SEL competencies are behaviors, respondents should be individuals who know the students well enough to assess their SEL competencies. If the SEL competencies are knowledge or mental processes, responses should involve students demonstrating those SEL competencies through a direct assessment or performance task.	If the intended respondents for the assessment are unfamiliar or unable to assess accurately SEL competencies in the local student population, find another assessment.

This space was intentionally left blank.

The table continues on the next page.

Does the assessment effectively measure the SEL competencies of interest?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Use a representative panel of content experts to develop and/or review items/tasks and scoring protocols to ensure that the assessment addresses SEL competencies sufficiently and appropriately.	Conduct a local review of assessment items/tasks and scoring protocols to determine if those items sufficiently and appropriately address the competencies and outcomes for the local SEL program.	Clear and detailed specifications of the SEL competencies measured is important when developing not only tasks but also scoring protocols to ensure alignment between those defined specifications and the items/tasks and scoring protocols. Assessment developers can use expert review, an assessment blueprint, and/or mapping of items/tasks onto scores, to demonstrate that items/tasks represent a cross-section of competencies measured. For example, • Asking individuals with emotional regulation expertise to review items from an emotional regulation scale and indicate the extent to which each item aligns with the SEL competency and if the set of items overlook important aspects of the SEL competency. • Having a group of content experts review the number and content of items/tasks to determine if the assessment coverall all measured SEL competencies sufficiently. As a general guideline, • Selected-response assessments should have at least three to five items for each competency measured. • Performance assessments typically involve a smaller number of tasks but that could hinder the generalizability of the scores if there is too broad of a set of SEL competencies measured.	If the developer does not document that the assessment sufficiently and appropriately addresses SEL competencies, conduct a local review with relevant expertise or find another assessment that provides this type of documentation and evidence.

Does the assessment effectively measure the SEL competencies of interest?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide empirical evidence that items/tasks used to measure each competency are more highly related to each other than to items that measure other competencies (internal structure).	Determine if evidence supports that items/tasks used to measure SEL competencies are more highly related to each other than they are to items that measure other competencies.	If an assessment claims to measure three competencies, there should be higher correlation among items/tasks that measure the same competency than among items/tasks that measure the other two competencies. Statistical analyses are used to support the assumption that unique rather than redundant information about each competency exists and these analyses typically require large sample sizes. For example, • For selected-responses assessments, confirmatory factor analysis can provide evidence that items load significantly on to factors that represent the different SEL competencies measured by the assessment. • For performance assessments, generalizability may be used to demonstrate that variability exists across different tasks.	If evidence of internal structure does not support that items/tasks measuring a SEL competencies are unique rather than redundant of items/tasks measuring other SEL competencies, use caution when reporting, interpreting, and/or using scores for individual competencies.

Evaluating the Measurement Quality of Social and Emotional Learning (SEL) Assessments

2. *Does the SEL assessment provide credible evidence for your intended uses?*

Does the assessment provide credible evidence for intended uses?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Clearly state the intended interpretation and uses for the assessment score(s) and highlight evidence that justifies using the assessment for those interpretations and uses.	Ensure that the assessment developer's stated interpretations and uses align with local plans for using assessment results and determine if evidence supports those interpretations and uses.	Measures might be developed for screening, formative, interim, and/or summative purposes, and this intent should be specified by the assessment developer and align with local plans for using the data. For example, • If teachers will use the information to guide instruction, then use a formative assessment measure that provides classroom-level data to guide those instructional decisions. • If a school plans to use an assessment in an improvement process, then use an interim or summative measure that provides school-level data to assess progress and determine how to move forward.	If the assessment developers' intended interpretations and uses for an SEL assessment do not align with local plans or are unsupported, find another assessment that does align with plans for use.

This space was intentionally left blank.

The table continues on the next page.

Does the assessment provide credible evidence for intended uses?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Identify score(s) provided (e.g. overall score, subscores, performance levels) and items/tasks used to generate each score. Clearly state recommendations and limitations for reporting and interpreting those scores.	Determine if scores provided will guide intended uses or assist in reaching conclusions about students' achievement of SEL competencies. Ensure that local plans for reporting and interpreting assessment results follow developer's recommendations and limitations. Be alert to possible misinterpretation of scores and take steps to minimize inappropriate interpretation and use.	Do not interpret assessment results for purposes unless recommended by the developer with the support of evidence. Examples include: • Most SEL competency assessments are appropriate for assessing students' strengths and do not have enough evidence to support using the assessment for screening or diagnosing mental health issues. • If the assessment reports multiple scores, do not aggregate those into a single score unless the developer provides evidence that doing so is appropriate. • If the assessment reports a single composite score, do not disaggregate the score unless the developer provides evidence that doing so is appropriate. • If the assessment will guide instruction or practice, reported scores should provide enough specificity to inform these intended uses such as by providing subscores on specific domains or competencies. • If an assessment will determine whether SEL has occurred, an SEL program is effective, or whether SEL learning goals are met, reported scores could be more general. Holistic and analytical scoring are typical for many performance assessments. • For holistic scoring, results are a single, holistic judgement about a students' SEL. • In analytical scoring, decisions result in judgements about one or more SEL competencies. Analytical scoring potentially can provide more information about strengths and weaknesses but requires evidence that those scores are able to differentiate between different SEL competencies.	If scores provided by the assessment will not guide intended uses or inform conclusions at the local level, find another assessment. Do not attempt to combine or calculate scores from an assessment without proper psychometric evidence. If assessment developer's recommendations and cautions for reporting or interpreting SEL assessment results do not align with local plans for reporting and interpretation, find another assessment that aligns with local plans.

Does the assessment provide credible evidence for intended uses?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Cite theory, research, or empirical evidence that students/observers/ interviewers interpret and respond to items/tasks as intended.	Review rationale or evidence provided by the assessment developer that respondents respond as intended to determine if it supports the use of the assessment with the local population and setting.	Assessments should find a way to document that respondents are answering items/tasks using the processes and behaviors the developer intended. For example, • Interviewing respondents about their response choices as they complete items. • Collecting feedback from raters about the factors they considered when assigning their ratings.	If there is insufficient rationale or evidence that respondents are interpreting and responding as intended, use other evidence of SEL competencies to confirm interpretations.

This space was intentionally left blank.

The table continues on the next page.

If the assessment will be used to determine students' strengths and needs…

Assessment developer should…	**Test user should…**	**Explanation**	**What to do if an assessment does not meet this criterion?**
Provide empirical evidence of consistency of item results (internal reliability) for all assessment scores reported.	Determine if assessment scores have an acceptable reliability coefficient (.80 or above for coefficient alpha).	Consider reliability evidence for each score to be reported understanding that aggregating scores at a class, group, grade, or school level will be more reliable than scores for individual students. If validity evidence appears to support assessment at the individual student level, a measure of internal consistency will indicate the extent to which a respondent responds similarly across items. Internal reliability typically takes the form of a coefficient alpha. • Coefficient alpha ranges between 0 and 1 with a value closer to 1 indicating better consistency (reliability). • The stakes of an intended use is a basis for determining the degree of reliability required, with higher reliability needed when stakes are higher. • A minimum threshold for reliability is .80. Reliability slightly below .80 is undesirable but may not be problematic. Reliability significantly below .80 is problematic for interpretation and use. **NOTE**: Sufficient reliability evidence is not enough to support the use of scores to make consequential decisions about individual students, such as for diagnosis or program placement.	If the internal reliability of any score reported is below .80, even slightly use caution when interpreting and using those scores for decisions about individual students. If the internal reliability of any score is not reported or considerably below .80, do not report, interpret, and/or use any scores/subscores that do not meet this minimum or find an assessment where all scores reported are sufficiently reliable.

If the assessment will be used to determine students' strengths and needs (continued)…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide a standard error of measurement and recommended confidence intervals/bands for all reported assessment scores.	When reporting and interpreting scores, include some reference to the true range of those scores based on standard error of measurement and confidence intervals or bands.	If an assessment provides evidence that supports reporting individual scores, also report confidence intervals to capture the true potential range of the students' performance. Confidence intervals are particularly important when comparing two different scores. For example, • Comparing an individual student's score against a criterion score such as proficiency level or norms. • Comparing changes in an individual's score over time. • Comparing the scores of two different individuals.	If standard error of measurement and/or confidence intervals or bands are not available, • contact the developer for this information, • use caution when determining students' strengths and needs, and/or • double check with other information about students' SEL competencies to see if the two sources agree.
See also expectations for "Is the assessment relevant for the students and the setting?"			

If the assessment will be used to compare scores over time…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide empirical evidence that scores are sensitive to changes in SEL over time.	Determine if evidence is applicable to the local setting and program and provides supportive evidence that the assessment will capture changes in SEL that occur over time.	Typically, cross sectional and longitudinal studies provide evidence that the scores of an assessment given at two different points in time would reflect a change in SEL if such a change did occur. • For example, comparing SEL skills at the beginning and end of the school year after students completed the SEL program.	If sensitivity to change over time is unsupported, do not use the assessment to determine if change over time has occurred.

If the assessment will be used to evaluate an SEL Program…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide evidence that assessment score(s) demonstrate change after implementing an SEL program that has been shown to be effective at improving the competencies measured by the assessment.	Determine if evidence provides information that is applicable to the local setting and program.	Evidence of how sensitive an assessment is to change could involve a field testing study. • For example, students who received instruction or maybe even higher quality instruction would score significantly higher on the assessment than students who did not.	If there is insufficient evidence that assessment scores can demonstrate change, be cautious about using scores to evaluate the effectiveness of the SEL program and/or instruction.

If the assessment will be used to improve school/program quality…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide evidence that assessment score(s) are moderately related to desirable educational outcomes (e.g. graduation, absentee rates, etc.)	Determine if evidence provided is applicable to the local quality improvement goals or outcomes.	Longitudinal, quasi-experimental, or experimental research studies can be used to determine if there is a significant correlation between relevant indicators of quality and the assessment score.	If there is insufficient evidence that score(s) are highly related to quality outcomes of local interest, do not use scores to make decisions about improving school/program quality.

This space was intentionally left blank.

The table continues on the next page.

If the assessment will be used to report separate results for different groups of students…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide rationale or evidence that students from different groups conceptualize, define, and experience the SEL competencies assessed by the assessment.	Review rationale or evidence provided to determine applicability to the local setting, SEL program, and demographics of the local student population.	If using the results of an SEL assessment to report separate results for different groups of students, it is important to ensure that relevant groups of student experience the assessed SEL competencies similarly. • For example, if reporting results separately for different racial/ethnic groups then the competencies measured should be culturally relevant for students in the local student population. If group difference are reported, do so cautiously and only after thorough review.	If there is insufficient rationale or evidence different groups of students conceptualize define, and experience SEL competencies similarly, • ask individuals from representative groups to review the relevance of SEL competencies assessed, or • do not report and compare results for different groups of students.
Provide evidence that assessment score(s) are equally valid, reliable, and fair for different groups of students. If not, clearly caution against the reporting of assessment scores for groups of students separately.	Determine if evidence provided is applicable to the local setting, SEL program, and demographics of the local student population and supports reporting scores separately for different groups of students.	Because of potential issues with relevance of SEL assessments for different groups of students (e.g. cultural, gender, age), if schools have an interest in comparing or reporting separately the results for different groups of students the school should: • Justify the use of those results for solving a specific problem of practice rather than just using it to report how different groups perform. • Ensure validity, reliability, and fairness study samples include students from different groups that will be compared or results reported separately. Preferably, require validity, reliability, and fairness study results are report separately for different groups of students.	If there is insufficient empirical evidence that score(s) are valid, reliable, and fair for different groups of students or the assessment developer cautions against it, do not report and interpret scores for groups of students separately.

Evaluating the Measurement Quality of Social and Emotional Learning (SEL) Assessments

3. *Is the SEL assessment relevant for your students and your setting?*

Is the assessment relevant for the students and the setting?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Identify the intended population for the assessment and clearly articulate if there are any inclusion or exclusion criteria.	Select an assessment that is intended for the key demographics (e.g. age/grade) of the local student population to be assessed.	Use assessments only with individuals who are demographically representative of the intended population. For example, • Do not use an assessment developed for Grades 9 and up if the intended population of the assessment is middle or elementary school students • Do not use an assessment with English Learners (ELs) if a developer indicates that the assessment is not appropriate for those students.	If the intended population for the assessment does not align with the key demographics of the local student population to be assessed, look for another assessment.
Provide a rationale and evidence that what and how SEL competencies are measured is developmentally appropriate for the grades/ages of students in the intended population.	Review the rationale and evidence to determine if the assessment is developmentally appropriate for the grade/ages of students in the local population.	Developmental appropriateness is particularly important if an assessment will be used to track SEL competency development over ages or grades. Student development of SEL competencies can differ not only because: • Different SEL competencies become important at different developmental stages. • Ways in which those SEL competencies are demonstrated or displayed changes over time. An assessment developer should address these developmental considerations when developing and validating the assessment.	If there is an insufficient rationale or evidence that an assessment is developmentally appropriate for the grades/ages of students in the local population, use for the grades/ages for which it would be appropriate or find another assessment.

Is the assessment relevant for the students and the setting?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Indicate the reading level and linguistic competency needed by respondents.	Determine if respondents will have appropriate levels of reading and linguistic competence.	The reading level and linguistic complexity of an assessment is not only important to consider in terms of students but for other respondents as well. • For example, if a parent report would require a sixth grade reading level and English proficiency, ensure that most if not all parents will meet those requirements; if not, determine how to accommodate the participation of parents who do not meet those requirements.	If an assessment developer does not specify reading or linguistic competency needed by respondents, • ask the assessment developer for more information, or • have a reading specialist/ELL coordinator review the assessment to determine if it is appropriate for the local population.
Specify the availability of language and ability accommodations are available. For available accommodations, provide guidance on when to use the accommodation and how to administer and score it.	Determine if accommodations for students or respondents in the local population are available.	If the setting has a linguistically diverse student population or a sizable number of students with identified disabilities, the availability of accommodations would allow these students to participate. Seeking out the availability of multi-language versions or forms for students with disabilities would be another option.	If needed accommodations are not available, ask assessment developer for more information. If adequate accommodations do not exist, • do not use the assessment for relevant students, or • seek out experts who can assist in identifying accommodations that would remove barriers for these students but not change the SEL competencies measured.

Is the assessment relevant for the students and the setting?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Use a culturally representative panel to review SEL competencies measured by the assessment to determine if how those SEL competencies are measured are relevant for different cultures.	Review the demographics and findings of the panel to insure individuals from cultural groups represented in the local setting and student population are included and the assessment will fairly assess SEL competencies for those cultural groups.	How an SEL assessment defines and measures competencies may not be relevant to respondents from different cultural groups because the value of those SEL competencies or how they are represented may vary. Cultural differences are an important consideration when developing SEL programs and assessments along with systematic review and/or empirical studies to ensure they are not culturally biased. For example, • Review panels should include members of each relevant cultural group or people either who work with or are familiar with those groups. • Comments from such individuals should be considered seriously.	If the SEL competencies addressed by the assessment have not been reviewed and approved by a culturally representative panel, • ask a panel that represents cultural groups in local student population about the relevancy of the SEL competencies, or • do not use the assessment for making decisions about unrepresented cultural groups.

This space was intentionally left blank.

The table continues on the next page.

Is the assessment relevant for the students and the setting?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Have a diverse panel familiar with the needs of different students review the content and format of the assessment for bias, sensitivity, and accessibility. Document whether that panel found with a high level of agreement that assessment is unbiased, sensitive, and accessible. If the panel identifies items or format as biased, insensitive, or inaccessible, describe how those issues were addressed.	Review the demographics of the panel and the findings of the panel to determine if the review is applicable to local setting and student population and if any bias issues were raised that might be a concern for the local student population.	Individuals of different backgrounds or individuals who are aware of capability differences among students should be involved in the development and review of SEL assessments. This includes: • Review panels representing different racial/ethnic groups, ages, gender, individuals with disabilities, etc. • Reviewing items for topic and wording that could be unfair or ratings of students by individuals who might have an unconscious bias.	If the developer has not used a panel that is representative of the local student population to review for bias, sensitivity, and accessibility, ask a local group that is familiar with the needs of different students in the local population to review the assessment and its items. If there is insufficient documentation that an assessment will be fair for specific demographic groups, do not use the assessment for those demographic groups.

This space was intentionally left blank.

The table continues on the next page.

Is the assessment relevant for the students and the setting?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide empirical evidence that responses to items/tasks and reported scores are not significantly different for students with similar levels of SEL competency from different demographic groups (e.g. race/ethnicity, language, and gender). If statistical differences exist, indicate actions taken to understand those potential differences better.	Review the evidence provided to determine if the assessment addresses key student groups in the local population and if it raises fairness concerns for individuals from those groups. Demographics to consider include gender, race/ethnicity, socioeconomic status, and language background.	Assessments can consist of items/tasks that do not function the same way for different group of students or differences between relevant subgroups on reported scores. At the item, measurement invariance studies (e.g. differential item functioning or multigroup confirmatory factor analyses) gather evidence that assessment items performs the same way for different groups of students. • If studies find a lack of measurement equivalence, a follow-up study determining whether differences are potentially due to bias should occur before a user can credibly use the assessment for measuring the SEL competencies of those diverse groups. • Those conducting the analyses must also be aware of the assumptions of the procedures and number of individuals needed to conduct those analyses to avoid misinterpretation of results. At the score level, differential prediction is a common method used to determine through regression analysis whether there are differences between relevant subgroups on reported scores.	If there is insufficient evidence that students with similar levels of SEL competency from demographic groups respond at the item/task or score level similarly, • do not report and compare the scores from subgroups, or • find another valid assessment for student populations that are demographically very diverse. If there is evidence of lack of equivalence at the item or score level that was not addressed by the assessment developer, do not interpret and use assessment results for those subgroups especially if they are a key group in the student population.

Evaluating the Measurement Quality of Social and Emotional Learning (SEL) Assessments

4. *Does the SEL assessment address issues related to administration, scoring, and the assessment format?*

Does the assessment address issues related to administration, scoring and the assessment format?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide detailed and clear instructions if test users will administer and score the assessment. If applicable, indicate if there are specific qualifications or training experiences needed to administer and score the assessment.	Ensure that all individuals administering and scoring the assessment receive instructions provided by the assessment developer. If applicable, ensure qualified or trained individuals are available to administer and score.	Logistics and required training time should be considered when making decisions to use a particular assessment. Training of the following individuals might be necessary: • Individuals administering assessments, completing rating scales, or conducting observation may need training on how to complete the assessments. • Individuals compiling and reporting data may need training on developer recommendations for reporting, interpretation, and use. • Individuals who will use and communicate findings might also need training such as how to communicate findings to students and families. Some assessments require that those administering and/or scoring an assessment have certain qualifications such as a degree, graduate coursework, or specific formal training. Even if an assessment does not have requirements for administration and scoring, consider guidance that encourages standardized administration and scoring for comparable scores.	If requirements for administration and scoring are unaddressed in the assessment documentation, ask the assessment developer for more information. Do not use the assessment if qualified individuals are not available or training of individuals to administer and score the assessment would not be possible.

Does the assessment address issues related to administration, scoring and the assessment format?

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
If the test developer administers or scores the assessment, describe the process for conducting the assessment and/or the procedure used for generating scores.	Ensure that the basis for administering items and/or generating scores aligns with definitions for SEL competencies and supports local plans for interpretation and use.	Some test developers will use automated means for administering or scoring assessments that often involve algorithms. Algorithms for scoring assessments or selecting items can be very technical, but developers should be able to explain conceptually how the algorithm works. This conceptual explanation will help indicate whether the assessment's administration and scoring procedures are appropriate for the local setting and SEL program.	If there is insufficient information about how the assessment is administered and scored, ask the developer for more information. If administration and scoring procedures are not appropriate for the local setting, student population, or SEL program, find another assessment.
Indicate if specific technological devices and software to administer and/or score the assessment are required or recommended.	Ensure that the all settings (e.g. schools) administering the assessments have access to required or recommended technological devices and software.	If administering an assessment via a technological device, there likely are requirements for the devices and type of software available on those devices. Differences in mode (e.g. paper and pencil vs. computer-delivered), device (e.g. desktop computer vs. tablet), or operating system (e.g. Windows vs. Macintosh) could differentially affect how assessments are completed by respondents and compromise score comparability.	If the required devices or software are not available, find another assessment. If not all settings administering the assessment have access to recommended technological devices and software, • find another assessment, • do not use the assessment in those settings, or • request evidence from the assessment developer that differences in devices or software used to administer or score the assessments will not affect score comparability.

Does the assessment address issues related to administration, scoring and the assessment format?

If assessment scores are determined using norms…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Report norms should be: • based on a recent, representative sample of sufficient size, • document the demographics of the students included in the sample (e.g. gender, age/grade, race/ethnicity, SES, geographic location), and • describe the setting in which the norm data were gathered.	Ensure the norm study and sample is: • current (gathered in last 5-7 years), • of sufficient size (500 or more total and 100 or more per grade/age group), • gathered from a setting similar to the local setting, and • collected from a student sample that includes representation of the local student population (e.g. gender, race/ethnicity, SES, geographic location).	Norm samples should include and document: • A proportional representation of students from different demographic groups (note number of English Learners in the sample). • The relevant setting in which a norm sample was administered the assessment. For example, norms developed using a predominately students from urban high school would not be relevant for rural middle school students.	If the norm sample is not current, is not of sufficient size, or does not represent students from different demographic groups relevant to the local population, • ask the developer about the availability of updated and relevant norm information, • do not use the norm-referenced scores for reporting or decision-making, or • find another assessment with applicable norms.

If there are multiple forms (different versions) for an assessment (e.g. Forms A & B)…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide evidence of score consistency across the different forms.	Determine if the evidence supports that scores from different forms of the assessment are comparable.	Equating is a commonly used technical process that establishes scores are interchangeable across different versions of a test. Equating samples need to be large and representative of the population under consideration for assessment.	Only use one form of the assessment if there is insufficient evidence that scores from multiple forms would provide consistent results across students.

If the assessment is a completed by a student…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Indicate how development or administration of the SEL assessment addresses common issues such as memory bias, social desirability bias, or reference bias.	Determine if the developer has provided convincing evidence or rationale that the SEL assessment is not susceptible to these biases.	Memory, social desirability, and reference biases are common issues to address in the development or administration of assessments where the student is the respondent. • Memory bias occurs if respondents are not aware or accurate in the assessment of their SEL behaviors or actions. • Social desirability bias involves the respondent providing an answer considered attractive instead of what is true for him/her. • Reference bias are responses affected by whom respondent compares his/her SEL competence. Such as, if an assessment has consequential decisions for students, they also may not be inclined to answer accurately.	If there is insufficient evidence or rationale for how potential biases were addressed or mitigated in development or administration, • ask the assessment developer for more information, or • ask a small group of potential respondents or individuals familiar with respondents to review items and determine if these biases could be problematic.

This space was intentionally left blank.

The table continues on the next page.

If the assessment is a rating or observation scale completed by someone other than the student…

Assessment developer should…	Test user should…	Explanation	What to do if an assessment does not meet this criterion?
Provide evidence that the administration and scoring protocol will lead to consistent decisions across different raters/observers (interrater reliability) and avoid or mitigate potential biased ratings	Use recommended training and protocols to avoid or mitigate biases. Determine if interrater reliability is acceptable (Kappa or Intraclass Correlation Coefficient (ICC) statistic of .70 or higher).	These types of assessment should provide evidence of interrater reliability because some teachers might rate differently than other teachers across items/tasks or students. Common rating issues include: • Inclination to rate students they "like" more positively than other students (halo effect). • Use more leniency or severity in ratings. • Misinterpret/misattribute sources of behavior. • Rating accuracy affected if respondents have a personal or professional stake in the results of the assessment (e.g. evaluate teacher performance). Such disparities would affect the consistency across raters. Therefore, these types of assessments should provide instructions on how to help raters/observers overcome these response biases. • For example, training observers on actual students, vignettes or videos with discussion of differences in ratings may be quite productive for calibrating ratings.	If there is insufficient information about how to avoid or mitigate rater response bias, • ask assessment developer for more information, or • ask a small group of potential respondents to review items and determine if these biases could be an issue for them or others. If there is insufficient evidence of interrater reliability or interrater reliability is considerably below .70, • ask the assessment developer for more information, • consider more training for raters/observers, or • find another assessment.