EdInstruments

Considerations for Use of Early Grades Reading Assessments[1]

As part of the larger library of instruments, EdInstruments catalogues early grades reading assessments for use in research and practice. These instruments are organized within Academic Knowledge and Skills > ELA > Reading: Foundational Skills on our website. To guide potential users, we briefly describe early grades reading content, highlight general psychometric properties to consider, and underscore specific considerations for use by researchers and practitioners. While we do not endorse individual instruments, we provide examples of instruments for various uses. Users can navigate our website to explore additional options.

Early Grades Reading Content

Reading is a complex skill that comprises several interconnected sub-skills that develop in a somewhat predictable sequence that begins with pre-language skills in infancy and extends into complex text interpretation and meaning-making skills, which continue to develop through adulthood. For K-3 students, educators focus on teaching the following from the National Reading Panel (2000): phonemic awareness, phonics, fluency, vocabulary, and comprehension. Assessments that measure the ability to fluently and accurately connect letters and phonemes to sounds (e.g., letter sound fluency, phoneme segmentation fluency), decode words (e.g., decoding accuracy), read pseudowords (e.g., nonsense word fluency), read words (e.g., word identification), and read connected text (e.g., oral reading fluency), will be the most important and efficient tools for supporting early reading instruction and for research at these grade levels.

Early grades reading assessments can serve a range of purposes for researchers and practitioners. These uses can include individual diagnostics, universal screening, progress monitoring, accountability, academic research, and evaluation, among others. As such, content coverage of a given assessment will vary depending on its intended purpose and use. Users should carefully

[1] We thank Alison Gandhi, Guangming Lin, Nancy Nelson, and Patricia Vadasy for their helpful feedback and expertise. Their participation in this work does not signify endorsement of any individual instrument/assessment. Any errors or omissions are our own.

consider their needs and ensure this aligns with the intended purpose of the assessment and the appropriate grade-level learning objectives.

General Psychometric Considerations

As with use of any instrument, users need to consider the technical adequacy and psychometric properties of early grades reading assessments. Of particular utility to practitioners, users can consult the National Center on Intensive Intervention screening and progress monitoring tools charts for specific information on technical adequacy of specific assessments at specific grade levels for specific uses. Furthermore, the National Center on Improving Literacy released a series of briefs, several of which are focused on psychometrics of literacy screening tools. While written for practitioner audience, these briefs are also applicable to researchers. We highlight these points below.

Validity

Validity is the extent to which theory and evidence support the intended interpretations of scores for proposed uses. Petscher et al. (2019a) highlight aspects of validity that should be considered with literacy assessments. Of particular importance are content validity (i.e., the extent an instrument represents all facets of a given construct), convergent validity (i.e., whether sets of test scores that should be correlated are correlated), discriminant validity (i.e., whether sets of test scores that should not be correlated are not actually correlated), and predictive validity (i.e., whether test scores predict some later measure). Furthermore, consequential validity is paramount for school practitioners. Practitioners should consider how will they will use the measures exactly, the implications for students, who scores should be shared with, and the positive and negative consequences of these decisions.

Reliability

Reliability has to do with the consistency of sets of test scores. As Petscher et al. (2019b) describe:

> "Internal consistency broadly refers to how well a set of item scores correlate with each other. Alternate form [reliability] describes how well two different sets of items within an assessment correlate with each other. Test-retest [reliability] is concerned with how stable two sets of scores are over a fixed period of time. Inter-rater [reliability] is associated with how two different observers of a behavior rate the behavior in the same way. Each of the forms of reliability are distinct and useful for their own purposes but should not be used interchangeably."

Sample Representativeness

Sample representativeness is also an important consideration. When determining whether an instrument is valid for your desired use, the instrument should have been validated using a sample that is representative of your population of interest (Pentimonti et al., 2019a). For

example, when validating assessments that will be used to measure ELs' early reading skills in English, the assessment development process should have involved a reasonable sample of ELs in the validation process so that the scores derived from the assessment and the comparisons that will be made can be generalized to the EL population being assessed. Similarly, it would be inappropriate to use an assessment that was validated solely in socioeconomically advantaged communities to assess students in Title I schools that are eligible for federal funding on the basis of socioeconomic disadvantage.

Bias

A high-quality early reading assessment should also have evidence against bias. That is, students should not receive higher or lower scores for reasons beyond the knowledge or skills being assessed (Pentimonti et al., 2019b). Users should examine whether information is available on differential item functioning or measurement invariance.

Additional Considerations for Use by Researchers

Researchers who study early reading skills select reading assessments based on their specific research questions. Some research studies address empirical questions about reading theory, the development of specific reading skills (e.g., decoding accuracy, or word reading fluency), or the influences on reading skill development. For example, the new revised Wechsler Individual Achievement Test (WIAT–4) appears to be useful in research on orthographic mapping development.

A study will often include measures of proximal outcomes; sometimes this requires measures of subskills. If transfer to distal outcomes is of interest, this might recommend measurement of a broader range of reading skills.

Researchers who work in close or formal partnership with districts or schools often seek out measures that teachers in the research sites will be familiar with, be able to interpret, and can inform instruction. For example, DIBELS tests are widely used in grades K-8 and are useful in research designs that track growth on DIBELS subskills. DIBELS information on risk indicators can be used to confirm student risk status for eligibility in research. Many research reports have been published on DIBELS, which provides researchers comparison samples and interpretations.

For those considering research employing secondary/extant data analysis, the Early Childhood Longitudinal Study (ECLS) assessments have also been widely used in early reading/literacy research. This study also includes a wide-range of survey items as well as additional measures of student achievement.

Several other considerations influence the choice of early reading assessments for research use. It is important that test instructions for assessments are clearly outlined, and that assessors can be trained to administer the tests reliably to research subjects. Some research designs require assessments with alternate forms. Most research studies include multiple assessments, and test administration time is a factor in choosing measures, as well as whether the assessment is

administered individually or in groups. EdInstruments provides guidance with these choices and refers to external resources for further detail.

Additional Considerations for Use by School Practitioners

Early reading practitioners use assessments for three purposes:

- *Screening* all students to identify those who may need additional support;
- *Diagnosing* specific skill strengths and deficits to inform instructional approach, either for a full class or for students needing supplemental and more intensive support; and
- *Progress monitoring* to assess rate of progress and make decisions about when a change in instructional approach is needed, for students receiving supplemental and intensive support.

Screening Assessments Are Not Diagnostic

As schools seek to reduce the amount of time students spend being assessed in favor of instruction, there is often a desire to select and use efficiently administered assessments for multiple purposes. Most commonly this occurs when screening assessments are adopted because of their purported alignment with curriculum and use for diagnostic purposes. However, screening and diagnostic assessments, by definition are constructed in opposition to one another. It is important for practitioners to use assessments aligned with their intended purpose. Common early reading screeners include the NWEA Measures of Academic Progress (MAP) assessments as well as DIBELS. Common diagnostic assessments include the Boehm Test of Basic Concepts Third Edition (Boehm-3), the Group Reading Assessment and Diagnostic Evaluation, and the HMH reading inventory.

Monitor Progress with Instructionally Relevant Measures to Support Adaptations to Instruction

In the early grades, it is critical to monitor progress by assessing skills that are instructionally relevant to support adjustments to instruction based on the data. In the area of reading, expressive measures – instead of receptive measures – should be used to assess the skills students are expected to demonstrate (e.g., we expressively assess a student's letter-sound correspondence knowledge by asking the student to produce the sound the letter "n" makes; we receptively assess the same skill by asking the student to point to the letter that makes the sound /n/). Ideally, progress monitoring measures would be linked to screening assessments to allow for comparison of scores within and across tiers of support over time; however, to support the efficiency and utility of progress monitoring, it is more important that the data from progress monitoring assessments are immediately instructionally relevant and can be used to inform adjustments to intervention received. Several test developers state that their assessments can be used for both screening and progress monitoring, which include aimswebPlus, DIBELS, and STAR reading assessments, for example.

Usability and Cost

Early reading assessments should be easy to use, and ideally come with automated score reports that are easy to understand for both teachers and parents. While most assessments have software packages that support data entry, scoring and reporting, the cost for these additional supports varies, which is also of interest to practitioners.

Context for Use

The context in which assessments will be administered and used is important to consider. For instance, when assessing ELs, it is necessary to consider whether the goal of assessment is to evaluate a child's performance in English (e.g., in determining instructional placement in English-only reading instruction) or their native language (e.g., to assess their reading skills more generally). A major challenge when assessing ELs is the availability of assessments and qualified examiners for accurately measuring the skills of ELs in their native language. Moreover, assessments that have been translated from English into other languages may not have been adapted culturally. Themes in assessments developed in the US primarily for English-only speakers may not translate well to assessments intending to assess the skills of diverse learners who have just arrived in the US from other countries. The language and procedures for testing English learners are also problematic, and test developers/publishers rarely include recommendations for explaining the directions for tests, for example, whether the directions can be provided in the student's language.

Timing and Scope of Criterion Measures

As mentioned above, school systems often seek to be more efficient in their administration of assessment measures. As such, there is often drive to adopt a single assessment system that spans as many grade levels as possible. Although this is a respected aim, the unintended consequences of adopting a system-wide assessment tool could lead to weaker technical adequacy evidence in some grades as compared to others. Frequently, assessment systems that span early elementary school and later elementary school or even middle school have not been validated at each grade level. In these cases, validation against an end-of-year outcome often begins in third grade, coinciding with the administration of state assessments that are used as the criterion or outcome measure. As a result, these assessment systems work very well for assessing the literacy skills of students in grade 3 and beyond but have little to no evidence for their utility in grades K-2.

Decision-making "Stakes"

Because of the assessment challenges described above and the limited availability of assessments that have been rigorously validated for all subgroups of the population, it is important to consider the "stakes" of the decisions being derived from assessment data. High stakes decisions, such as determining whether a student is eligible for special education, require assessments that have robust technical adequacy evidence for the target population. Low stakes decisions, such as determining whether a student needs more practice with a particular skill, could potentially be carried out using assessments with somewhat weaker technical adequacy evidence.

Accessibility

Numerous assessments have been validated to support accurate assessment of early literacy skills for students with high-incidence disabilities. However, when assessing students with low-incidence disabilities (e.g., visual impairment, hearing loss, significant cognitive impairment) and moderate to severe display, extensive modifications may be needed to permit established assessments to be accessible to these students, which may render assessment scores of little value. Validated assessments that take into account the accessibility needs of students with low-incidence disabilities are needed to accurately assess these students' early literacy skills.

Concluding Remarks

A common theme in assessment is the lack of readily available information on the psychometric quality of the assessments, most often with tests from the major test developers. While we provide guidance on psychometric considerations and other tools exist, test developers must do more to make their technical reporting available.

Although technology has enhanced the tools we have available to assess students with special needs and other diverse learners, there remain major gaps between widely available, validated assessment tools and best practices in assessment for these students. Computer adaptive tests (CATs), which are highly efficient for screening for early literacy risk among typically developing students and those with mild to moderate disabilities, are limited by a shortage of validation studies that evaluate performance against early elementary school outcomes. As technology improves, efforts have been taken to ensure expressive measures of early literacy skills are embedded in CATs; however, many CATs still do not do this and when they do it is limited to oral reading fluency. For students who are not yet readers, including many children with or at risk for disabilities, this is a major gap that should be addressed in future assessment development efforts.

Similarly, computer-based diagnostic assessments that pinpoint student difficulties and provide instructional tools that teachers can immediately use to deliver explicit, systematic reading instruction to students would be a highly useful contribution to the field. Currently available instructional diagnostic assessments are primarily paper based. Numerous computer adaptive screening assessment systems do have linked instructional resources that allow students to interact with a game-based or other technology-driven system to receive intervention; however, this approach needs to integrate technology with typical classroom practice. The field needs a more seamless, coordinated approach using computer adaptive screening and diagnostic assessment that delivers evidence-based instructional resources to teachers. This includes delivery of supplemental intervention and supports aligned to student needs, and progress monitoring tools that measure student development of instructionally relevant skills. The development of expressive literacy measures would be a major advance in the field.

Finally, the paucity of assessments available for ELs in their native language(s) and cultural translation is a major challenge when assessing diverse learners. This challenge is exacerbated when considering assessment of ELs who may have special needs. In addition, early literacy

assessment tools are lacking for students with mild to moderate forms of low-incidence disabilities. Moreover, vocabulary and oral language assessments are largely unavailable for screening and progress monitoring in grades K-3. Developing and validating such tools for administration with ELs, other diverse learners, and students with special needs would complement currently available assessments in the marketplace.

Additional References

Petscher, Y., Pentimonti, J., & Stanley, C. (2019a). Validity. U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy. https://improvingliteracy.org/brief/understanding-screening-validity

Petscher, Y., Pentimonti, J., & Stanley, C. (2019b). Reliability. U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy. https://improvingliteracy.org/brief/understanding-screening-reliability

Pentimonti, J., Petscher, Y., & Stanley, C. (2019a). Sample representativeness. U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy. https://improvingliteracy.org/brief/understanding-screening-sample-representativeness

Pentimonti, J., Petscher, Y., & Stanley, C. (2019b). Bias. U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy. https://improvingliteracy.org/brief/understanding-screening-bias