Technical Adequacy of Acadience Reading 7–8

This document is reprinted from the
Acadience Reading 7–8 Assessment Manual
Chapter 8.

Mary Abbott

Roland H. Good,

Jacob S. Gray

Amy N. Warnock

Kelly A. Powell-Smith

acadience® reading 7-8

Suggested Citation:

Abbott, M., Good, R. H., III, Gray, J. S., Warnock, A. N., & Powell-Smith, K. A. (2024). *Technical Adequacy of Acadience Reading 7–8.* Acadience Learning LLC. www.acadiencelearning.org (Preprinted from *Acadience Reading 7–8 assessment manual,* pp. 75–82 by M. Abbott, R. H. Good, J. S. Gray, A. N. Warnock, & K. A. Powell-Smith, 2020, Acadience Learning.

It is important that assessments used for educational decision making be reliable, valid, and adhere to accepted professional standards of measurement (American Educational Research Association [AERA] et al., 2014). This chapter summarizes the evidence gathered that supports the reliability, validity, and decision utility of Acadience Reading 7–8 in assessing middle school content-area reading skills.

Reliability of Acadience Reading 7–8

The reliability of a test denotes the degree to which a test produces stable and consistent results across different time points, test forms, and assessors. It is generally recommended that reliability coefficients be at least .60 if the scores are used for administrative purposes, at least .70 for progress monitoring, at least .80 for screening, and at least .90 if the score is to be used for important individual educational decisions (Salvia et al., 2017). Four types of reliability were evaluated for Acadience Reading 7–8: alternate form, test-retest, internal consistency, and inter-rater.

Alternate-Form, Test-Retest, and Internal Consistency Reliability. Alternate-form reliability refers to the extent that different forms of the same test are correlated. For example, different passages of the Maze assessment would be considered alternate forms of the same test. To estimate alternate-form reliability, a sample of students are assessed with multiple forms and the scores of the forms are correlated. The more highly correlated the alternate forms, the more reliable are the measures. Test-retest reliability refers to the temporal correlation of two measures. The more highly correlated a measure is at two time points, the more reliable the measure.

To assess alternate-form and test-retest reliability for Acadience Reading 7–8, data were collected with approximately 75 students in schools in the Pacific Northwest, Midwest, and Northeastern United States. During this study, students completed their regularly scheduled fall assessment. Previously trained Acadience Learning assessors then administered a repeat of the fall assessment for test-retest reliability. In addition, in order to establish alternate-form reliability, within the same two week window, assessors collected the winter and spring benchmark measures. Alternate-form and test-retest reliability are reported in Table 8.1. Alternate-form reliability ranges from .50 to .94, with most correlations exceeding .70 and many above .80. Test-retest reliability ranges from .62 to .91, with most above .80.

Cronbach's alpha (α) is a measure of internal consistency. This estimate is calculated using both the average correlation among different items, and the number of items. This number represents the expected correlation between the observed test and a hypothetical test with exactly the same properties. An additional interpretation of Cronbach's alpha is the ratio of true score variance to observed score variance, meaning that if $\alpha = .80$, this means 80% of the observed variance in a test is due to variance in the underlying construct, while 20% is due to error. Because of the assumption of tau-equivalence, α actually represents an underestimation of a test's reliability. Traditional recommendations label an α exceeding .70 as having acceptable reliability, and over .90 as indicating excellent reliability. Internal consistency reliability is reported in Table 8.1 and ranges from .75 to .98.

Standard Error of Measurement. The Standard Error of Measurement (SEM) represents the average expected deviance of a student's observed score from their true score. The SEM is a function of both the reliability of a measure, and the observed standard deviation of the measure. The SEM becomes smaller as a test becomes more reliable and has a lower standard deviation. A smaller SEM reflects greater precision in estimating a given student's score. Standard Error of Measurement (SEM) is reported in Table 8.1.

Table 8.1
Summary of Acadience Reading 7–8 Reliability

	Benchmark 1		Benchmark 2		Benchmark 3	
	Grade 7	**Grade 8**	**Grade 7**	**Grade 8**	**Grade 7**	**Grade 8**
Alternate-Form						
Maze	.85	.93	.81	.86	.91	.86
Silent Reading	.50	.72	--	--	--	--
Oral Reading	.94	.74	--	--	--	--
Gate 3 Composite	.72	.76	.72	.67	.69	.73
Test-Retest						
Maze	.83	.91	--	--	--	--
Silent Reading	.69	.62	--	--	--	--
Oral Reading	.89	.91	--	--	--	--
Internal Consistency (α)						
Maze	.94	.98	.93	.95	.97	.95
Silent Reading	.75	.88	--	--	--	--
Oral Reading	.98	.90	--	--	--	--
Gate 3 Composite	.89	.91	.90	.86	.87	.89
SEM						
Maze	7.63	4.93	8.43	8.04	5.87	8.05
Silent Reading	3.81	4.34	--	--	--	--
Oral Reading	1.75	3.06	--	--	--	--

Note. Alternate Form reliability calculated as the average correlation of scores across the three passages of a test. Internal Consistency is Cronbach's Alpha. SEM = Standard Error of Measurement

Inter-Rater and Procedural Reliabilities. Inter-rater reliability indicates the extent of agreement among assessors who administer and score the test. Procedural reliability indicates the extent to which an assessor followed administration procedures when collecting assessment data. Inter-rater reliability of Acadience Reading 7–8 was examined within the initial preliminary benchmark study for the Acadience Reading 7–8 Oral Reading (OR) measure on two indicators: Correct Words Read and Comprehension Total Score. Prior to the study, site coordinators completed an Acadience-created OR training module and Acadience staff worked with site coordinators to become reliable on the OR measure. During a single test administration of an assessment, an Acadience staff member scored student performance at the same time as a site coordinator (i.e., "shadow-scoring"). Site coordinators were required to have a 90% or better scoring agreement with Acadience staff. Any scores below 90%

agreement required a re-administration with a different student. Once a site coordinator was reliable, they trained assessor staff and conducted the same inter-rater reliabilities with them. Assessors were required to meet an 85% level of agreement in order to collect OR data with additional students. Inter-rater reliability data were collected with 45 assessors with an average agreement rate of 99% for Correct Words Read and 92% agreement for the Comprehension Total Score.

For the Maze and Silent Reading (SR) measures, scoring takes place after the student has completed the assessment. Therefore, an administration checklist was created to ensure that the assessor appropriately administered the measure (i.e., procedural reliability). After Acadience training, site coordinators first became reliable with an Acadience staff person. Then site coordinators trained assessors and watched an administration. Assessors were required to meet a 90% or better score on the administration checklist. Administration checklist data were collected with 111 assessors for the Maze measure with an average of 98% administration accuracy. For the SR measure, the checklist was collected on 120 assessors with an average of 98% accuracy.

Validity of Acadience Reading 7–8

The validity of a test refers to "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests." (AERA et al., 2014, p. 9). Thus, test validity indicates the extent to which the actual interpretation of test scores corresponds to the theoretical interpretations of a test (Salvia et al., 2017). Evidence of validity includes evidence related to test content, internal structure, relationships between the test and other performances, convergent and discriminative relationships, and consequences of testing (AERA et al., 2014). While different experts use different terminology to describe these concepts, we chose to use Salvia et al.'s (2017) terms: (a) content validity, which includes evidence related to test content; (b) criterion-related validity, which includes evidence of the relationships between the test and other performances; (c) construct validity, which includes evidence related to internal structure; and (d) expert validity, which includes evidence related to how experts in the field view the test's content validity. Each of these concepts is discussed below. Finally, we address the design specifications for Acadience Reading 7–8 in the section on content validity.

Content Validity and Design Specifications. Content validity is the extent to which a test's items represent the domain or area of skills that are to be measured. Evidence of content validity for Acadience Reading 7–8 is provided by a detailed description of the underlying rationale and the research base for the selection measures and content. Establishment of content validity for Acadience Reading 7–8 is based on the rationale underlying its research and development. That research rationale is based on the premise that an assessment of middle school content-area reading skills represent science, social studies, and prose content areas. In addition, such an assessment should (a) measure comprehension skills needed when reading middle school content-area materials, (b) be relatively brief and efficient, (c) be formatted within a process that mirrors the literacy experiences of middle school students, and (d) support teachers in their efforts to make instructional decisions based on the assessment.

Skills involved in reading comprehension at the middle school level fall into three main categories: (a) vocabulary knowledge (word level), (b) identifying details (sentence level), and (c) inference-making (passage level). The reading research on vocabulary is well documented. Vocabulary research indicates a strong, positive, reciprocal relationship between word knowledge and reading comprehension (Baumann et al., 2003). In terms of detail comprehension, literal questions are the most direct, basic type of question. However, students also need to be able to connect information from the current sentence being read to a previous sentence or to put two pieces of related information together to gain understanding of the text (Landi & Perfetti, 2007). For example, "Maddy threw the bananas and strawberries into the blender. The smoothie felt cool as it slid down her throat." The reader must make a connection or association that bananas and strawberries put into a blender can make a smoothie. At the passage level, students with good comprehension use inferences to help facilitate text understanding and to build an internal representation of text content (Graesser et al., 1994). Finally, an additional component of comprehension is content purpose. Narvaez et al. (1999) reported that variability in text type influences the kinds of inferences that readers generate. That is, students read a prose passage differently than they read a more concept-dense science passage. By design, the three Acadience Reading 7–8 measures (Maze, Silent Reading, and Oral Reading) focus on these comprehension skills across the content areas of prose, science, and social studies.

Acadience Reading 7–8 measures are designed to briefly and efficiently assess content-area reading skills. The Acadience Reading 7–8 multiple-gating system potentially reduces the testing load for students. Each measure assesses comprehension in different ways through a series of indicators. This notion of comprehension being assessed by use of indicators is a critical one. Focusing on measuring indicators allows for a relatively efficient assessment that is reliable and valid for the purposes of identifying students who may need additional instructional support and monitoring.

The Gate 1 Maze comprehension indicator is used as an initial screening tool that assesses general reading skill. During Maze, the student is presented with three passages in which some words are replaced by a multiple-choice box that includes the original word and two distractors. The student reads each passage silently and selects the word in each box that best fits the meaning of the sentence. Each Maze passage has a 3-minute time limit. The Maze triad of passages takes 10–12 minutes to complete. The student's Maze Total Adjusted Score is equated and used as the Gate 1 Score.

For students who score Below or Well Below Benchmark at Gate 1, an additional indicator, Silent Reading is administered. Silent Reading is a group-administered measure that assesses vocabulary, sentence comprehension (passage details), and inference. The student is presented with three passages and 30 multiple-choice questions (10 per passage) and is given 36 minutes to read the three passages silently and answer the questions. The multiple-choice questions cover passage vocabulary, details, and inference. The students' Maze and SR scores are equated and averaged to create the Gate 2 Score.

Students whose Gate 2 Scores are Well Below Benchmark enter Gate 3 and are assessed individually with the OR measure. Oral Reading is an individually administered measure that assesses accurate and fluent reading of connected text and reading comprehension. Oral Reading is composed

of three indicators: Correct Words Read, Accuracy, and Comprehension. During Oral Reading, the student is presented with three passages and given 90 seconds to read each passage out loud. After each passage, the student is asked to provide a brief recall of everything they can remember about the passage. Following the recall, the student is asked to define vocabulary from within the passage and answer two inference questions about the passage. An OR triad takes approximately 15–20 minutes to complete. The student's OR scores (Correct Words Read, Accuracy, Comprehension) are equated and averaged with the equated Maze and SR scores to calculate the Gate 3 Score. By using a gating system, student testing time is tailored to student need and potentially reduced.

Acadience Reading 7–8 content represents the complex subject matter that students encounter during their middle school years. Determining passage topics began with an in-house team reviewing current middle school science, social studies, and language arts textbooks. Members of the team gathered samples from each content area and completed an analysis of the main features found in the text. Features reviewed included sentence and word length, text complexity in terms of the complexity of the vocabulary, as well as inferential features of the text including cause and effect and figurative language. For each content area, a set of topics, passage titles, and potential vocabulary were created and approved. Passage authors followed strict protocols that included a rigorous process of readability analysis, passage appropriateness, and also a repeated review for factual accuracy. The final step was collection of expert validity data (for details, see Expert Validity section).

The final area of content validity is the need to support teachers in making instructional decisions. The goal is to gather enough information to know whether the student sufficiently comprehends grade-level content-area reading material. According to Cain and Oakhill (2012), comprehension assessment has greater potential to provide useful instructional information if the assessment covers a range of skills. As previously noted, Acadience Reading 7–8 measures different levels of comprehension, and through the gating procedures, assesses comprehension in different test formats. Each test format provides additional insight into the instructional needs of the individual student. For example, in a Trifactor Item Analysis of Silent Reading (Gray et al., in review) distinct constructs across content areas (prose, science, social studies) and skills (vocabulary, details, inference) were found. These separate constructs increase the power and flexibility of school staff to refine intervention by content area and comprehension skill with potentially small targeted modifications to instruction. By focusing on support within general education settings as well as established supplemental and intervention support systems such as reading labs and special education, student academic support can be boosted across the instructional day (e.g., Korinth & Fiebach, 2018).

Criterion-Related Validity. Criterion-related validity is the extent to which a test relates to other tests that measure the same or similar constructs. Two types of criterion-related validity are commonly described. Concurrent validity refers to how a student's performance on the test relates to a criterion measure of the same construct administered at the same time. Predictive validity refers to how a student's performance at one point in time predicts that student's performance on the criterion measure at a later point in time.

To examine the criterion-related validity of Acadience Reading 7–8, data were collected on 76 seventh- and 74 eighth-grade students in four school districts across the United States. These data were a part of the Acadience Reding 7–8 Benchmarks Study, and were collected during the 2017–2018 school year. The participating schools included students across a range of different ethnic groups and socioeconomic levels, including students who were English language learners. In addition to collecting all three benchmark measures for Acadience Reading 7–8, sites collected an external criterion measure, the Stanford Achievement Test Series, Tenth Edition (SAT10; Pearson, 2003). The SAT10 is a standardized, norm-referenced, timed achievement test. Students in this study completed the Advanced 1 vocabulary and reading comprehension sections which took 70 minutes. During the assessment, the students read passages and answered multiple choice questions.

A summary of the criterion validity (both predictive and concurrent) results is provided in Table 8.2. All correlations are statistically significant at the $p < .05$ level. Correlation coefficients range from moderate to large. The strong concurrent validity coefficients indicate that Acadience Reading 7–8 is strongly related to performance on the SAT10 Reading scale. The predictive validity coefficients indicate that Acadience Reading 7–8 is an effective predictor of future reading performance and may be used to predict which students will have later reading difficulties.

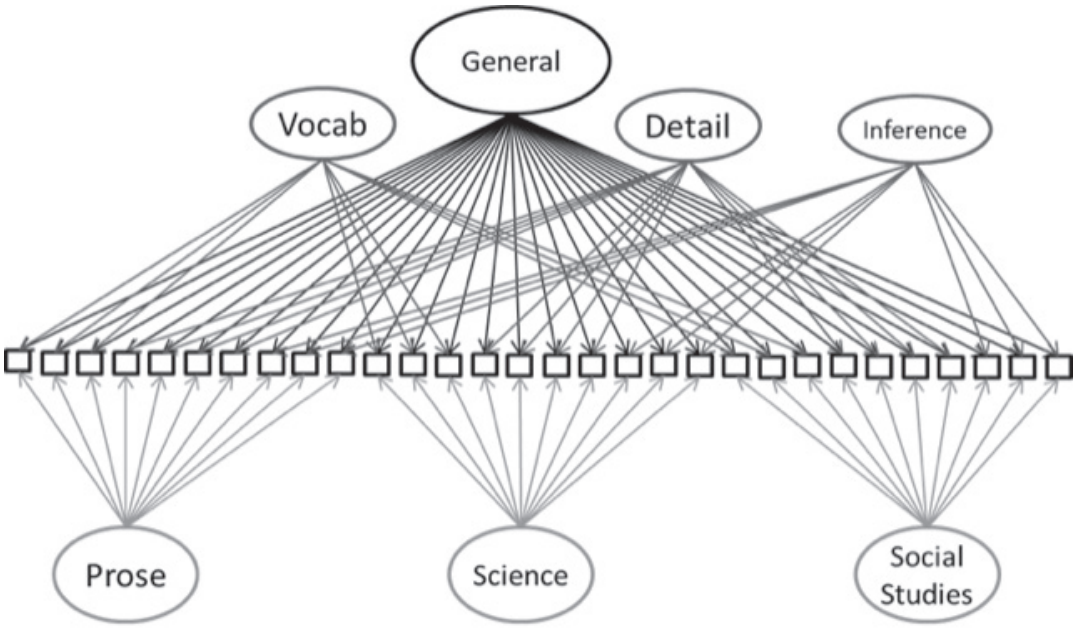
Table 8.2
Summary of Acadience Reading 7–8 Validity

	Predictive Validity				Concurrent Validity	
	Grade 7		Grade 8		Grade 7	Grade 8
	BOY	MOY	BOY	MOY	EOY	EOY
Maze	.82	.80	.69	.72	.79	.73
Silent Reading	.79	.82	.60	.66	.86	.71
Oral Reading						
Words Read	.55	.53	.57	.43	.62	.61
Accuracy	.75	.53	.72	.54	.72	.72
Comprehension	.75	.78	.79	.57	.72	.65

Note. BOY = beginning of year; MOY = middle of year; EOY = end of year.

Construct Validity. Construct validity of a test is the extent to which a test measures a theoretical trait or characteristic and includes evidence of convergent and discriminant power and evidence of the consequences of testing. Acadience Reading 7–8 is designed to be used to make educational decisions about the student's need for instructional support in acquiring critical content-area reading skills. As such, the consequences of the assessment should result in accurate identification of students who need instructional support. Evidence of the construct validity for Acadience Reading 7–8 is demonstrated in the Trifactor Item Analysis of Silent Reading (Gray et al., in review) in which distinct constructs across content areas (prose, science, social studies) and skills (vocabulary, details, inference) were found (see Figure 8.1). Additional support for the construct validity of Acadience Reading 7–8 is provided by the expert validity data collected for each passage.

Figure 8.1
A Trifactor Model of Acadience Reading 7–8 Silent Reading



Expert Validity. As part of the development process for Acadience Reading 7–8, expert validity was collected on each passage. We selected content-area specialists who had extensive experience in teaching middle school/high school within the given content area. All experts had extensive teaching experience in their content area and were recently retired. One was an administrator at the time of her retirement.

Experts were asked to rate each passage for the appropriateness of (a) the topic for the grade, (b) passage content, and (c) level of content detail within the passage. Each expert provided a rating of 1 to indicate a satisfactory or 0 to indicate that the passage needed reexamination. Experts also provided comments to justify their ratings. Experts were encouraged to note any vocabulary, phrasing, or content that they believed might be inappropriate or confusing. Expert comments and their ratings were used to further refine passages prior to any further passage formatting and readability testing with students. Experts were given a gift card for providing feedback on the passages to thank them for their input.

In general, ratings were strong for the 126 passages reviewed by the experts. For Appropriate Topic for Grade, 125 of 126 were given a rating of 1. One expert review gave two passages a 0 rating. For both passages, the reviewer noted that the subject matter of these passages was too complicated to cover in one passage. Therefore, the content became too dense to be grade appropriate.

For the Appropriate Content for Grade, 124 of 126 passage were given a rating of 1. Comments on passages that failed to get a 1 rating revolved around passage accuracy. One reviewer again noted that although facts for two passages were accurate, due to the complexity of the passage content, the facts could be led to a too simplistic view of a complicated subject.

The final category was Level of Detail for Content, in which 122 of 126 passages received a rating of 1. Comments about passages that failed to receive a rating of 1 revolved around "wordsmithing." Two of the experts provided detailed editing about how we might improve wording, phrasing, and sentences within the passages. Several suggestions were made about how the passage might be simplified to further narrow the topic. We reviewed every suggestion and modified some passages based on this valuable feedback.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Baumann, J. F., Kame'enui, E. J., & Ash, G. E. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, D. Lapp, J. R. Squire, & J. M. Jensen (Eds.), *Handbook of research on teaching the English language arts* (2nd ed., pp. 752–785). Erlbaum.

Cain, K., & Oakhill, J. (2012). Reading comprehension development from seven to fourteen years: Implications for assessment. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 59–76). Rowman & Littlefield Education.

Graesser, A. C., Singer, M., & Trabassko, T. (1994). Constructing inferences during narrative text comprehension. *Psychology Review, 101,* 371–395.

Gray, J. S., Abbott, M., Good, R. H., & Powell-Smith, K. A. (in review). A trifactor item response theory model for silent reading ability.

Korinth, S. P., & Fiebach, C. J. (2018). Improving silent reading performance through feedback on eye movements: A feasibility study. *Scientific Studies of Reading, 22*(4), 289–307.

Landi, N., & Perfetti, C. A. (2007). An electrophysiological investigation of semantic and phonological processing in skilled and less-skilled comprehenders. *Brain and Language, 102,* 30–45.

Narvaez, D., van den Broek, P., & Ruiz, A. (1999). Reading purpose, type of text and their influence on thinkalouds and comprehension measures. *Journal of Educational Psychology, 91*(3), 488–496.

Pearson. (2003). Stanford Achievement Test Series, Tenth Edition (SAT10).

Salvia, J., Ysseldyke, J. E., & Witmer, S. (2017). *Assessment in special and inclusive education* (13th ed.). Cengage.