Mary Frances Early College of Education
K–12 Assessment Solutions
UNIVERSITY OF GEORGIA

WIN Work Ready Assessments

Pre-Operational Assessment and Equating Processes

UGA Gwinnett Campus:
2530 Sever Road
Suite 149
Lawrenceville, GA 30043

Contact Us:
Toll free: (888) 392-8977
Web: https://coe.uga.edu/directory/k-12-assessment-solutions

Table of Contents

1. Purpose and Scope of Review 3
1.1 Assessment Purpose 3
2. Test Design and Blueprint 4
2.1 Development of the Test Blueprint 4
2.2 Item Types 4
2.3 Universal Design Principles 4
2.4 Evaluation and Recommendations 5
3. Item Development and Review 5
3.1 Item Development 6
3.2 Item Review 6
3.3 Evaluation and Recommendations 6
4. Field Testing and Psychometric Analysis 7
4.1 Field Testing 7
4.2 Item Selection and Form Construction 8
4.3 Alignment Study 8
4.4 Other Testing Activities 9
4.5 Evaluation and Recommendations 9
5. Standard Setting and Cut Scores 10
5.1 Standard Setting Methodology 10
5.2 Panelists and Training 11
5.3 Rounds of Judgment and Use of Data 12
5.4 Evaluation and Recommendations 12
6. Scale Scores, Achievement Levels, and Reporting 13
6.1 Scoring and Scale Scores 13
6.2 Achievement Levels 13
6.3 Reporting and Intended Uses of Results 14
6.4 Evaluation and Recommendations 15
7. Conclusion 15
References 17

1. Purpose and Scope of Review

This review evaluates the technical and procedural quality of the pre-operational processes used for the WIN Work Ready Assessments from item and test development to operational use. The purpose of this review is to determine whether assessment development activities were conducted in accordance with generally accepted assessment best practices and whether sufficient evidence exists to support valid and defensible interpretations of scores.

Subject matter experts from the University of Georgia reviewed available documentation and evidence related to assessment design, item development, field testing, psychometric evaluation, form construction, alignment, standard setting, and readiness for operational implementation.

For the purposes of this evaluation, the University of Georgia reviewed the following primary documents provided by the WIN Team:

- WIN Work Ready Assessments Technical Report
- WIN ASA_Alignment Report 2025
- WIN ASA_Field Testing Summary
- WIN ASA_Standard Setting Report 2025

In addition to these reports, the WIN Team also provided ancillary materials and artifacts supporting test development activities, such as administration materials, panel training materials, and rating sheets.

1.1 Assessment Purpose

The WIN Work Ready Assessments were designed to measure foundational academic skills required for success in job-specific training programs and workforce preparation activities. Assessment development was grounded in the U.S. Department of Labor Occupational Information Network (O*NET, released in 1997), an online database that provides profiles for hundreds of occupations including required skills, daily job duties, experience requirements, wages and employment trends, and job zones. Assessment items are aligned to learning objectives linked to O*NET Job Zones and associated workplace skill expectations. The assessment framework was revised in 2024-2025 following a re-review of O*NET data that identified additional workplace readiness skills requiring additional learning objectives (WIN, 2026a, 2026b, 2026d).

The intended interpretation of assessment results is that examinees are credentialed and possess the foundational math, reading, and data literacy skills necessary to successfully engage in training at a particular job level. The results are not intended to indicate readiness for a specific job, but rather readiness for job specific training and positioned for successful progress.

The development of the current assessment edition involved a multi-year process starting with research and framework development, followed by a 2024 update that incorporated new high-demand workplace skills. Key activities to support these updates included refreshing the item pool, developing and field testing new items, conducting alignment studies to ensure the match between learning objectives and operational forms, and establishing new performance standards

and cut scores using a Modified Angoff methodology (Cizek & Bunch, 2007; WIN, 2026b, 2026c, 2026d).

2. Test Design and Blueprint

This section summarizes how the WIN Team translates the O*NET Content Model into test blueprints that define what content is measured on the WIN Work Ready Assessments and how learning objectives complexity progresses over five achievement levels. Consistent with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), the WIN Team documents what is being measured, the intended examinee population, and the intended uses of test scores.

2.1 Development of the Test Blueprint

The ***test blueprint*** serves as the primary mechanism for ensuring content alignment and supporting valid interpretations of test scores. Specifically, it describes which learning objectives are assessed, how often they appear, and how much each objective contributes to the overall score (Nunally, 1978; Sireci, 1998).

Assessment development for the WIN Work Ready Assessments was guided by test blueprints organized around learning objectives, achievement levels, and content strands. Work Ready Math, Work Ready Reading, and Work Ready Data each include five achievement levels representing increasing complexity and workplace readiness demands. Learning objectives, discrete skills, and strands were used to define the content domain and guide item development activities (WIN, 2026a, 2026b). Item development, review, and selection processes are governed by the test blueprint.

2.2 Item Types

All assessment items are traditional selected-response (multiple-choice) items, including a stem (question) and requiring examinees to select one correct answer from four options. Available documentation describes item development and blueprinting procedures based on selected-response formats and does not indicate that alternative item types were considered during the development of test specifications or test blueprints.

2.3 Universal Design Principles

Universal design principles are applied from the outset to make tests accessible, fair, and appropriate for the widest range of students (Laitusis & Karvonen, 2025). Rather than addressing accessibility only after tests are developed, the guiding design specifications for the WIN Work Ready Assessments consider accessibility across all phases of development to support valid measurement for all learners.

Universal design principles guide decisions related to language demand, visual presentation, response formats, and interaction design. Items are reviewed for clarity, readability, and potential sources of bias, and accessibility features. These practices are intended to reduce barriers to access by addressing accessibility considerations during assessment development.

The technical report indicates that each item is reviewed at the concept stage and again throughout the development process for bias and sensitivity. Checklists were used to flag stereotyping, gender, ethnicity, religion, or class biases, and emotionally charged topics (WIN, 2026a).

2.4 Evaluation and Recommendations

Evaluation

The WIN documentation indicates that the test blueprint guides item development, form construction, and score reporting, supporting representative coverage of the intended content domain. The WIN test blueprint provides a structured framework linking learning objectives, content strands, and achievement levels to the intended interpretations of scores. The assessment design incorporates procedures intended to promote accessibility, fairness, and content alignment through the application of universal design principles and multiple stages of item review. These practices reduce the need for accommodations by addressing access at the design stage rather than retrofitting accessibility after development.

The selected-response item type supports efficient administration and scoring while providing broad coverage of the targeted content domain. However, some learning objectives may involve skills that could also be measured through performance-based or constructed-response tasks such as “Measure the length, width, and height of products and materials” to fully assess the intent of the skill.

Overall, the documented blueprint and design processes are consistent with generally accepted assessment development practices and provide an appropriate foundation for subsequent development activities.

Recommendations

The WIN Team could expand technical documentation describing the development, review, and approval of test blueprints, including the roles of subject matter experts and other stakeholders involved in decision-making. Additional documentation regarding blueprint governance would strengthen evidence supporting content validity and assessment oversight.

Additionally, future reviews could consider evaluating the alignment between selected-response item types and the skills being measured, particularly for learning objectives that emphasize complex performance or application, to ensure that assessment methods continue to provide the strongest evidence of intended knowledge and skills.

3. Item Development and Review

Item development and review processes are critical to assessment quality because test score validity depends on the quality, alignment, fairness, and accessibility of individual test items. Effective assessment programs use well-established item development procedures, multiple

stages of review, and quality control processes to ensure items accurately measure intended knowledge and skills while minimizing irrelevant barriers to performance.

The WIN documentation describes a structured item development process involving subject matter experts with relevant domain knowledge and experience and well versed in academic and workforce standards, assessment design principles, accessibility, and bias and sensitivity review process. Items were developed using established specifications and underwent multiple review stages before field testing, including content review, editorial review, bias and sensitivity review, accessibility review, and quality assurance review (WIN, 2026a). Item development procedures emphasized alignment with learning objectives, clarity of language, fairness, and adherence to universal design principles. This is in accordance with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

3.1 Item Development

Item development is a multi-step process designed to ensure that test items are accurate and aligned to the learning objectives. This process is typically guided by item specifications and style guidelines that define expectations for content, format, and rigor. The technical report describes the criteria for developing the parts of an item (WIN, 2026a).

Every item is reviewed at the "concept stage" and throughout the development process, particularly for bias and sensitivity issues. Checklists are used to determine no stereotyping is present and there is no bias based on gender, ethnicity, religion, or socioeconomic status. Following stereotyping and bias reviews, items were reviewed for potentially emotionally charged topics.

The item development procedures emphasize alignment to learning objectives, multiple review activities, and consideration of fairness and accessibility throughout development.

3.2 Item Review

Professional editors review items for length, text density, word usage, sentence structure, and grammar to improve readability and clarity. Following this, the executive content development team conducts a final quality assurance review to verify that items align with learning objectives, contain a single correct answer, feature parallel options, remain free of bias or sensitivity issues, and utilize clear language and graphics (WIN, 2026a).

3.3 Evaluation and Recommendations

Evaluation

The documentation indicates that WIN implemented a structured and multi-stage item development and review process designed to support item quality, alignment, fairness, and accessibility. Item development activities were guided by learning objectives and included multiple review points, checking for content accuracy and alignment, bias and sensitivity, editorial quality, and overall quality assurance. This is evidence that multiple safeguards were incorporated to support the quality of items before field testing.

Overall, the documented procedures reflect generally accepted assessment development practices by emphasizing alignment to learning objectives, systematic review, and attention to fairness and accessibility throughout the development process. There is reasonable evidence that item development and review activities were conducted in an industry-accepted manner.

Recommendations

No significant concerns were identified regarding item development and review procedures. Future technical documentation could provide additional information regarding any training that is provided to item developers and reviewer guidelines and process for agreement. Such information would provide additional transparency regarding item quality control and development governance.

4. Field Testing and Psychometric Analysis

Before new test items are used to calculate student scores, they are first administered through a structured field testing process. After test items have been reviewed and approved by content experts for accuracy and by reviewers for potential bias, the items are embedded in operational test forms as field test items. Responses to field test items do not affect scores; rather, they provide data to evaluate how items function.

4.1 Field Testing

Newly developed items were field tested between January and April 2025 using a matrix-sampling design in which unscored field-test items were embedded within operational assessment administrations (WIN, 2026c). Examinees were presented with five to eight embedded field-test items under operational testing conditions. This approach allowed field-test responses to be collected under conditions that closely resembled operational administration while minimizing burden on individual examinees.

Field-test sample sizes were substantial, including approximately 58,427 Work Ready Math examinees, 56,016 Work Ready Reading examinees, and 56,128 Work Ready Data examinees (WIN, 2026c). Approximately half of the sample consisted of South Carolina high school students participating with prior approval from the South Carolina Department of Education.

After field testing is complete, a series of statistical analyses are conducted to evaluate how each item performed. The analyses examine multiple aspects of item quality, including:

- **Difficulty** – whether the item is appropriately challenging for students.
- **Discrimination** – how well the item distinguishes between students with stronger and weaker overall performance.
- **Distractor performance (for selected-response items)** – how students respond to the incorrect answer choices and whether those options function as intended.

These analyses were used to evaluate item performance and identify items that warranted additional review or revision. Items violating one or more criteria were flagged for review and possible revision. Modified items were not immediately transitioned into operational use but instead remained eligible for future field testing to establish new psychometric evidence.

The field testing process generated empirical evidence regarding item functioning under operational testing conditions and informed subsequent psychometric review activities. The large sample size provides a strong basis for evaluating item functioning and supporting psychometric decision-making.

4.2 Item Selection and Form Construction

Form construction is the process of selecting items for operational use in a manner that maintains content representation, psychometric quality, and comparability across test forms. Effective form construction helps ensure that score interpretations remain consistent regardless of which operational form an examinee receives.

WIN documentation indicates that forms were developed collaboratively by the content development team and psychometrician following the results from field testing. Item selection was informed by assessment blueprints and psychometric analyses. The assessments use fixed-form designs, and two parallel forms were developed for each Work Ready assessment. Operational forms were equated using common-item psychometric methodologies to support score comparability across administrations (WIN, 2026a).

4.3 Alignment Study

The WIN Team conducted an alignment study in August 2025 as part of the assessment revision process (WIN, 2026b). The study was designed to evaluate whether assessment items appropriately reflected the intended learning objectives and supported valid score interpretations.

The alignment study was conducted following substantial revisions to Work Ready Math, Work Ready Reading, Work Ready Data, including the addition of new learning objectives identified through a review of updated O*NET workforce-readiness data. Eight subject matter experts participated in the standard setting process. Panelists represented across baccalaureate, graduate, and post-graduate levels and had expertise in K-12 and adult education, workforce experience, and corrections experience. They represented four states in the South; one participant was from South Carolina. Panelists were selected through a structured recruitment and interview process intended to ensure relevant expertise and diverse perspectives (WIN, 2026b).

Prior to conducting the review, panelists received an orientation and training session led by the WIN psychometrician. Reviewers were provided with assessment forms, learning objectives, examples of the skills associated with each objective, and standardized scoring materials. Panelists independently reviewed items and evaluated whether each item appropriately measured its intended learning objective.

The review process required panelists to compare the content and cognitive demands of each item to the specified learning objectives. Because the learning objectives incorporated both knowledge and skills and level of complexity, alignment judgments simultaneously evaluated content representation and degree of rigor. Individual judgments were collected, and agreement rates were calculated across reviewers. Figure 1 shows a sample score sheet that documents individual reviewer decisions for each item.

Figure 1. Panelist Score Sheet [Sample]

	ALIGNMENT			ROUND 1					ROUND 2				
Question #	FORM 1	FORM 2		Question #	level 2	level 3	level 4	level 5	Question #	level 2	level 3	level 4	level 5
1	Yes	Yes		1	Yes	Yes	Yes	Yes	1	Yes	Yes	Yes	Yes
2	Yes	Yes		2	Yes	Yes	Yes	Yes	2	Yes	Yes	Yes	Yes
3	Yes	Yes		3	Yes	Yes	Yes	Yes	3	Yes	Yes	Yes	Yes
4	Yes	Yes		4	Yes	Yes	Yes	Yes	4	Yes	Yes	Yes	Yes
5	Yes	Yes		5	Yes	Yes	Yes	Yes	5	Yes	Yes	Yes	Yes
6	Yes	Yes		6	Yes	Yes	Yes	Yes	6	Yes	Yes	Yes	Yes
7	Yes	Yes		7	Yes	Yes	Yes	Yes	7	Yes	Yes	Yes	Yes
8	Yes	Yes		8	Yes	Yes	Yes	Yes	8	Yes	Yes	Yes	Yes
9	Yes	Yes		9	Yes	Yes	Yes	Yes	9	Yes	Yes	Yes	Yes
10	Yes	Yes		10	Yes	Yes	Yes	Yes	10	Yes	Yes	Yes	Yes

Results demonstrated strong alignment for Work Ready Math and Work Ready Reading and moderate-to-strong alignment for Work Ready Data (WIN, 2026b). While the Data assessment showed a higher proportion of moderately aligned items than the other content areas, no Data items were classified as weakly aligned.

4.4 Other Testing Activities

In addition to development activities, assessment programs establish procedures to support secure and standardized administration, accessibility, and score integrity. The assessments are typically delivered online through a web-based proprietary system. Online versions are offered in in English, Spanish, and American Sign Language. Paper versions including standard booklet, large print, and Braille are also available. The technical report itself did not provide details regarding administration procedures, test security requirements, accommodation policies, or administrator training presentations. This may be because WIN provides customized test administration materials for partners like South Carolina districts and schools that are posted in the state's online resource center.

4.5 Evaluation and Recommendations

Evaluation

The documentation indicates that WIN implemented field-testing and psychometric evaluation processes that support developing quality items and operational forms. Field-test items were administered under conditions closely resembling operational testing. The documented field testing process reflects accepted assessment development practices by collecting item performance data under operational testing conditions. Further, sample sizes exceed those typically required for stable estimation of classical and IRT item statistics. These analyses provide evidence that test items are technically sound, measure student knowledge accurately, and operate fairly across student groups.

The documented form construction procedures demonstrate appropriate consideration of both content and psychometric characteristics. The use of blueprint-driven item selection, psychometric review, parallel forms, and equating procedures provides evidence that WIN took reasonable steps to support consistent score interpretation across forms.

The alignment study provides additional validity evidence supporting the relationship between assessment content and the learning objectives. WIN's documentation describing panelist expertise, training, independent evaluations, agreement rates, and process for flagged items supports the conclusion that operational forms adequately represent the intended workplace-

readiness learning objectives and contribute to defensible interpretations of scores. The documented procedures provide evidence that WIN evaluated and improved content alignment through a systematic review process.

Overall, the field testing, psychometric analysis, form-construction, and alignment-review activities are consistent with generally accepted assessment-development practices and provide substantial evidence supporting the technical quality of the operational assessments.

Recommendations

Future technical documentation could provide additional details regarding item selection criteria and form construction decision rules to increase transparency regarding development of operational forms; however, the information reviewed provides sufficient evidence that accepted form development practices were followed.

Continued review of moderately aligned Work Ready Data items may further strengthen alignment evidence in future assessment revisions and support continued monitoring of content representation. No significant concerns were identified regarding the alignment study methodology.

Future technical documentation could include descriptions of WIN-recommended administration procedures, test security protocols, accommodation policies, administrator training, and test security measures to provide standardized guidance and evidence supporting secure and consistent implementation. Although supporting state administration materials were reviewed, incorporating recommended best practices into technical documentation would provide additional evidence supporting standardized implementation across testing settings.

5. Standard Setting and Cut Scores

Standard setting is the process through which performance standards and associated cut scores are established to support meaningful interpretation of assessment results. Because the WIN Work Ready Assessments underwent substantial revision in August 2025, including the addition of new learning objectives derived from updated O*NET analyses and lengthening the test forms, WIN conducted a new standard setting study prior to the operational release of the revised assessments in December 2025.

The standard setting process was designed to establish the minimum score required for each of the five achievement levels used by the Work Ready Math, Work Ready Reading, and Work Ready Data assessments. These achievement levels are linked to progressively more complex foundational workforce-readiness skills and corresponding O*NET Job Zones. The resulting cut scores provide the basis for classifying examinees into achievement levels and support interpretation of assessment results in relation to workforce preparation and career readiness.

5.1 Standard Setting Methodology

WIN employed a Modified Angoff (Yes/No) method to establish cut scores for each assessment (WIN, 2026d). The study was conducted virtually and facilitated by a WIN psychometrician. The Modified Angoff approach is a widely recognized standard setting methodology used in

educational and credentialing assessments because it relies on subject matter expert judgments regarding the performance expected of minimally competent candidates at each performance level.

Under the Modified Angoff approach, panelists reviewed every item and determined whether a borderline examinee at a specified achievement level would be expected to answer the item correctly. This approach simplifies judgments to a binary decision, thereby reducing cognitive burden and promoting consistency of judgments across panelists. The methodology focuses on the knowledge and skills represented in the performance/achievement level descriptors and the characteristics of examinees functioning at the threshold between adjacent achievement levels. Panelists discussed the skills that distinguish adjacent achievement levels and used those distinctions to guide judgments regarding expected item performance. This process ensures that cut score recommendations are rooted in the intended meaning of the achievement levels rather than item difficulty alone.

5.2 Panelists and Training

Eight subject matter experts participated in the standard setting process. These panelists were the same committee from the August 2025 Alignment Study (WIN, 2026b, 2026d). Panelists represented across higher education levels and had expertise in K-12 and adult education, workforce experience, and corrections experience. They represented different states; one participant was from South Carolina.

Prior to conducting the review, panelists received an orientation and training session led by the WIN psychometrician. Training included an overview of the purpose of standard setting, the role of subject matter experts, the definition and interpretation of the performance level descriptors, security requirements related to assessment materials, and instruction on implementing the Modified Angoff methodology. Sample items were reviewed, and panelists practiced making judgments using the established procedures before beginning. Figure 2 shows a sample score sheet provided during training that illustrates the process for individual panelist decisions for each item.

Figure 2. Panelist Judgement Score Sheet [Sample from Training Materials]

	ROUND 1			
Question #	level 2	level 3	level 4	level 5
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				

The training was intended to establish a common understanding of achievement level expectations and judgment procedures.

5.3 Rounds of Judgment and Use of Data

The standard setting process included two rounds of independent judgments (WIN, 2026d). During the first round, panelists independently reviewed items and recorded Yes/No judgments regarding whether a borderline examinee at each achievement level would answer the item correctly. Panelists made these judgments using the performance level descriptors, assessment forms, and scoring materials provided during training. Individual judgments were collected, and agreement rates were calculated across panelists.

Following completion of the first round, panelists participated in facilitated discussions focused on areas of disagreement and items that generated the greatest variability in ratings. To support these discussions, panelists were provided with summary information showing the distribution of cut score recommendations, average recommendations, and measures of variability across panelists. Individual panelist identities were protected through the use of numeric identifiers. Discussions focused on the interpretation of performance level descriptors, the characteristics of borderline examinees, and the rationale underlying differing judgments. Importantly, panelists were not encouraged to seek consensus regarding cut scores; rather, discussions were intended to improve shared understanding of the achievement level expectations and judgment task.

Following discussion, panelists completed a second round of independent judgments. After Round 2, summary statistics and panelist distributions were again reviewed and discussed. The documentation notes that panelists did not express concerns regarding the resulting recommendations after the second round and that the degree of change between rounds was relatively small. Based on the stability of results and supporting research indicating that two rounds frequently produce defensible results, the decision was made not to conduct a third round of judgments.

The documentation describes how panel recommendations were reviewed and transformed into operational cut scores that were subsequently adopted for operational use. The process included trained subject matter experts, defined performance level descriptors, training methods with borderline examinees, multiple rounds of judgment, discussion, and evaluation of panelist convergence. Documentation additionally describes how resulting panel recommendations were transformed into operational cut scores and adopted without modification (WIN, 2026d).

5.4 Evaluation and Recommendations

Evaluation

The documentation reviewed provides evidence that the standard setting process was a formal testing event, facilitated by a psychometrician, informed by established performance/ achievement level descriptors, and supported by educator and workforce expert judgments. The standard setting report provides evidence of panelist convergence through documentation of individual ratings, means, and standard deviations across both rounds. The inclusion of multiple rounds, structured discussion, performance level descriptor review, and quantitative feedback are consistent with accepted practices intended to improve the quality and consistency of expert judgments and the resulting cut score recommendations. Although the documentation acknowledges limitations associated with panel size and the absence of impact data at the time of the standard setting, the overall evidence indicates that the panel recommendations were

translated into operational cut scores through a systematic, documented, and professionally accepted process suitable for operational use.

The standard setting process is consistent with generally accepted practices for criterion-referenced assessments and provides sufficient evidence supporting the defensibility of achievement level classifications and associated score interpretations. No significant concerns were identified regarding the standard setting methodology or implementation procedures.

Recommendations

All panelists were drawn from southeastern states (South Carolina, Kentucky, Tennessee, and Florida). Future standard setting studies may benefit from including subject matter experts from additional geographic regions to further broaden the range of perspectives informing achievement level expectations.

In addition, future reviews could incorporate post-administration impact data and stakeholder feedback to provide supplementary evidence regarding the appropriateness and consequences of the achievement level classifications after operational implementation.

6. Scale Scores, Achievement Levels, and Reporting

Following administration and scoring, examinee responses are transformed into achievement-level classifications and reported in a manner intended to support interpretation of foundational workforce-readiness skills. The intended interpretation of assessment results is that examinees possess varying levels of foundational math, reading, and data literacy skills associated with workplace preparation and success in occupational training.

6.1 Scoring and Scale Scores

Items are scored dichotomously, with responses marked either correct or incorrect.

Once responses are scored, raw scores are transformed to a reporting scale. The technical report indicates that item response theory methodologies are used to support score scaling and comparability across operational forms, while the standard-setting report documents the conversion of raw-score cut scores to the operational reporting scale through test characteristic curve and theta-based transformation procedures. Reliability estimates were strong, ranging from 0.84 to 0.91, and classification accuracy estimates ranged from approximately 0.75 to 0.78 (WIN, 2026a), providing evidence that scores are sufficiently consistent for their intended uses.

6.2 Achievement Levels

The WIN Work Ready Academic Skills Assessments include five achievement levels for each content area: Level 1, Level 2, Level 3, Level 4, and Level 5. The standard setting event defined the characteristics of borderline examinees and established cut scores, and raw cut scores were established for each achievement level across the three assessments.

Table 1 shows a sample set of achievement levels and descriptors for math.

Table 1. Math Achievement Levels and Anchors [Sample from Standard Setting]

Achievement Level	Achievement Level Descriptor (Anchors)
1	A participant who has earned an Achievement Level 1 in Work Ready Math has the foundational math skills to succeed in preparatory programs for O*NET Job Zone 1 career pathways. Participants can apply basic problem solving strategies and mathematical understandings to solve 1-step mathematical problems in the context of Job Zone 1 occupations.
2	A participant who has earned an Achievement Level 2 in Work Ready Math has the foundational math skills to succeed in preparatory programs for O*NET Job Zone 2 career pathways. Participants can apply problem solving strategies and mathematical understandings to solve mathematical problems in the context of Job Zone 2 occupations.
3	A participant who has earned an Achievement Level 3 in Work Ready Math has the foundational math skills to succeed in preparatory programs for O*NET Job Zone 3 career pathways. Participants can apply problem solving strategies and mathematical understandings to solve mathematical problems in the context of Job Zone 3 occupations.
4	A participant who has earned an Achievement Level 4 in Work Ready Math has the foundational math skills to succeed in preparatory programs for O*NET Job Zone 4 career pathways. Participants can apply problem solving strategies and advanced mathematical understandings to solve mathematical problems in the context of Job Zone 4 occupations.
5	A participant who has earned an Achievement Level 5 in Work Ready Math has the foundational math skills to succeed in preparatory programs for O*NET Job Zone 5 career pathways. Participants can apply advanced, complex mathematical reasoning to solve problems in the context of Job Zone 5 occupations.

The achievement levels are directly linked to learning objectives, performance (achievement) level descriptors, and O*NET job zone expectations.

6.3 Reporting and Intended Uses of Results

Assessment results are intended to communicate examinees' levels of foundational workforce-readiness skills in mathematics, reading, and data literacy. They are designed to determine whether examinees possess the foundational academic skills necessary to succeed in training associated with various occupational pathways.

The WIN documentation reviewed primarily focused on assessment development and technical characteristics. Limited documentation was available regarding operational score reports, stakeholder-facing interpretive guides, or other reporting resources intended to support score use and interpretation.

6.4 Evaluation and Recommendations

Evaluation

The documentation reviewed indicates that WIN implemented scoring, scale score reporting, and achievement level classification procedures that are consistent with generally accepted psychometric practices. Assessment forms were developed and equated using common-item methodologies and psychometric analyses designed to maintain consistency of score interpretations across administrations.

Reported reliability and classification accuracy estimates provide evidence that assessment results are sufficiently consistent for their intended purposes. Additional psychometric analyses, including dimensionality analyses and differential item functioning, further support the appropriateness of the scoring model and resulting score interpretations.

The documented relationship among learning objectives, achievement level descriptors, standard setting judgments, and cut scores provides evidence supporting meaningful interpretation of achievement classifications. The achievement level framework appears aligned with the intended use of the assessments as measures of foundational workforce-readiness skills rather than preparation for specific occupations (WIN 2026a, 2026d).

Overall, the scoring, scaling, and achievement-level classification procedures provide reasonable evidence supporting the consistency, interpretability, and intended uses of assessment results.

Recommendations

The WIN documentation describes the intended uses and interpretations of assessment results; however, limited evidence was available regarding operational score reports, user guides, or stakeholder-facing materials intended to support score interpretation.

Future technical documentation and reviews could include sample score reports, interpretative guides, and related reporting resources to provide additional evidence that score meanings, limitations, and appropriate uses are communicated clearly to students, educators, workforce partners, and other stakeholders.

7. Conclusion

Based on the documentation reviewed, the WIN Team implemented a comprehensive assessment development process that included blueprint development, item development and review, field testing, psychometric analysis, alignment studies, standard setting, and score scaling activities. The procedures described are generally consistent with accepted assessment development practices and provide evidence supporting the validity and defensibility of score interpretations for the intended purposes of the WIN Work Ready Assessments.

Noted strengths of the assessment program include the structured item review process, extensive field testing sample sizes, documented psychometric analyses, formal alignment study, and well-accepted standard setting methodology. These activities collectively contribute evidence supporting content validity, score reliability, and achievement level interpretations.

Recommendations provided throughout this report focus primarily on expanding technical documentation related to blueprint governance, form construction procedures, score reporting resources, and recommended administration practices. These recommendations are intended to enhance transparency and documentation rather than address substantive concerns regarding assessment quality. Specific opportunities include providing additional information about blueprint approval processes, form construction procedures, and reporting materials that support appropriate score interpretation and use. Although administration and test security materials and training are customized for each state, WIN may also benefit from documenting a core set of recommended practices that apply across states.

Overall, the evidence reviewed supports the conclusion that the WIN Work Ready Assessments were developed through a process consistent with professional assessment practices and are appropriately prepared for operational use within the intended context.

References

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

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting*. SAGE Publications, Inc., https://doi.org/10.4135/9781412985918.

Laitusis, C. C., & Karvonen, M. (2025). A case for reimagining universal design of assessment systems. *Educational Measurement: Issues and Practice, 44,* 5-14. https://doi.org/10.1111/emip.12674.

Nunally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117.

Worldwide Interactive Network, Inc. (WIN Learning). (2026a). WIN Work Ready Assessments technical report (Version 6.2026) [Confidential technical report].

Worldwide Interactive Network, Inc. (WIN Learning). (2026b). WIN Academic Skills Assessments alignment report 2025. (Updated May 2026) [Confidential technical report].

Worldwide Interactive Network, Inc. (WIN Learning). (2026c). WIN Academic Skills Assessments field testing summary [Confidential technical report].

Worldwide Interactive Network, Inc. (WIN Learning). (2026d). WIN Academic Skills Assessments standard setting report 2025 (Updated May 2026) [Confidential technical report].

Mary Frances Early College of Education
K–12 Assessment Solutions
UNIVERSITY OF GEORGIA

WIN Work Ready Assessments

Evaluation of Math, Reading, Data

2026

Gwinnett Campus:
2530 Sever Road
Suite 149
Lawrenceville, GA 30043

Contact Us:
Toll free: (888) 392-8977
Web: https://coe.uga.edu/directory/k-12-assessment-solutions

Table of Contents

Summary 3
1. WIN Work Ready Assessments 5
1.1 South Carolina Career Ready Test Overview 5
1.2 Review Process 6
2. Test Blueprint Review 7
2.1 Coverage of Learning Objectives 7
2.2 Depth of Knowledge Distribution 8
3. Evaluation of Overall Validity 9
3.1 Coverage by Achievement Level 9
3.2 Alignment to Learning Objectives 11
3.3 Depth of Knowledge 12
3.3.1 Distribution of DOK Levels 12
3.3.2 Review Items for DOK Level 14
3.4 Internal Structure 21
3.4.1 CTT Item Statistics 21
3.4.2 1-PL Item Statistics 23
3.4.3 Dimensionality 24
3.4.4 Reliability 25
3.4.5 Measurement Invariance 25
3.4.6 Equating Multiple Forms 26
4. Achievement Level Classifications 27
References 28

Summary

This document reviews the WIN Work Ready Math, Reading, and Data Assessments and evaluates the extent to which these assessments provide valid and reliable evidence for measuring foundational employability skills associated with career readiness. The review examines the alignment of the assessments of their intended learning objectives and achievement levels, evaluates evidence related to content validity and internal structure, and considers the reliability and appropriateness of achievement level classifications used to support credential attainment and reporting.

The review of the WIN Work Ready Assessments yielded several key findings. Overall, the test blueprints appropriately define the knowledge and skills measured by each assessment and reflect the progression of achievement levels aligned to O*NET Job Zones. The operational forms were consistent with blueprint specifications, and all learning objectives identified on the blueprints were represented across the reviewed forms. Item reviews indicated that the assessments are generally well aligned to the intended learning objectives, with only a small number of items on the WIN Work Ready Data recommended for content review or objective re-alignment. Overall, the alignment between assessment items and the intended learning objectives was strong and supports the content validity of the assessments.

With respect to cognitive complexity, the WIN Work Ready Assessments incorporate a range of Depth of Knowledge (DOK) levels intended to measure increasingly complex applications of skills and knowledge. The DOK distributions identified in the *Test Maps* appear reasonable and generally consistent with the intended purpose of workforce-readiness assessments. However, item-level reviews indicated that a number of items classified as DOK 3, particularly within the Reading and Data assessments, function operationally at a DOK 2 level because the correct responses can often be identified through straightforward interpretation or application of readily available information. While these findings do not raise concerns regarding overall assessment quality, a broader review of DOK classifications and distributions is recommended to ensure that the cognitive demands of operational forms reflect the intended blueprint specifications.

Evidence based on the internal structure of the assessments indicated that the WIN Work Ready Assessments generally function as intended. Classical Test Theory analyses demonstrated appropriate ranges of item difficulty and strong item discrimination across the Math, Reading, and Data assessments. Item response theory analyses showed that item difficulties span the ability continuum, supporting the validity of score interpretations and classification decisions. Dimensionality analyses provided evidence that each assessment measures a single dominant construct, and reliability estimates were strong. The multiple forms equating process was sound, supporting the use of multiple forms for each assessment. Due to insufficient sample sizes, DIF and subgroup analyses are incomplete, but are strongly recommended for additional evidence regarding fairness and validity. Additionally, model fit evaluations and post-equating checks are recommended for continued assurance of valid score comparisons.

The evaluation also examined achievement level classifications for the WIN Work Ready Assessments. Achievement level cut scores were established using a recognized standard-setting methodology and applied to 1-PL-based ability estimates to classify students into five

achievement levels corresponding to established O*NET Job Zones. Classification quality indices indicated acceptable classification accuracy and consistency. When feasible given blueprint and content constraints, we recommend adding more difficult items to better scale and classify higher ability examinees.

Overall, the WIN Work Ready Assessments demonstrate sound technical quality and provide valid and reliable evidence regarding the prerequisite career readiness skills required for success in occupations aligned to the corresponding O*NET Job Zones. The assessments are generally aligned to their intended learning objectives, exhibit appropriate psychometric characteristics, and support meaningful achievement level classifications linked to career readiness outcomes. Although the assessments are functioning as intended, targeted refinements, including reviewing flagged content and DOK alignment issues, conducting differential item functioning analyses when feasible, and expanding item coverage at higher difficulty levels, would further strengthen the validity, interpretability, and utility of assessment results.

1. WIN Work Ready Assessments

This report reviews the WIN Work Ready Assessments, a suite of career-readiness assessments covering academic skills measuring Math, Reading, and Data. This review evaluates the three component subtests, or employable skills, of the WIN Work Ready Assessments. These subtests are utilized for the *South Carolina Career Ready Test* and are conducted on behalf of the South Carolina Oversight Committee in collaboration with the South Carolina Department of Education.

1.1 South Carolina Career Ready Test Overview

The *South Carolina Career Ready Test* is a statewide assessment administered to all eleventh-grade students to measure foundational workplace skills identified as essential for career readiness. The assessment evaluates students in four key areas: soft skills, math, reading, and data analysis.

The South Carolina Career Ready Test incorporates the WIN Work Ready Assessments to measure students' proficiency in employability skills. South Carolina students who pass the WIN Work Ready Assessments earn the South Carolina Work Ready Credential, demonstrating mastery of foundational math, reading, and data analysis skills. Students who also pass the soft skills component earn the South Carolina Work Ready Essential Soft Skills Credential, which validates competencies in effective communication, professionalism, teamwork and collaboration, critical thinking, and problem solving. Both credentials are endorsed by the National Work Readiness Council and are designed to help educators, workforce development professionals, and employers determine students' readiness for post-secondary education, technical training, apprenticeships, on-the-job training, and employment opportunities.

The mathematics, reading, and data assessments are assigned a scale score and an achievement level score from 1 to 5 based on the U.S. Department of Labor O*NET Job Zones, which categorize occupations based on the education, experience, and training typically required for successful job performance. Students who achieve Levels 2 through 5 earn a corresponding Work Ready Credential, reflecting readiness for occupations within those job zones.

The WIN Work Ready Assessments are organized around learning objectives aligned to achievement levels and O*NET Job Zones. Within each assessment, related skills are grouped into strands that represent broad domains of workplace competencies and help define the knowledge and skills measured by each assessment. Together, these components guide the development of authentic assessment items that measure students' ability to apply career readiness skills in workplace-relevant contexts.

1.2 Review Process

With the support of the Education Oversight Committee, experts from the University of Georgia evaluated the WIN Work Ready Assessments for reliability and validity in assessing the prerequisite career readiness skills required to successfully complete preparatory programs for jobs categorized by the corresponding achievement level / job zone. The evaluation process follows best practices in educational measurement, as detailed by the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014).

The following materials were provided by the WIN Team for evaluation:

- 100A and 200B Test Forms
- For each Work Ready assessment:
    - Blueprint and Test Maps
- WIN Work Ready Assessments Technical Report

The Test Maps included metadata about individual items and psychometric indices. The test contractor calculated all parameters; no additional estimation of item or test parameters was conducted. The items reviewed for content validity were presented in the new/current edition of all three-component assessment, and the psychometric review is based on the summaries from the Technical Report which reported on the new editions released December 2025. The n-counts are relatively small and will be updated as more data is collected.

This report was prepared by the University of Georgia and examines critical elements of the WIN Work Ready Assessments test designs and summarizes findings and recommendations for each.

2. Test Blueprint Review

A test blueprint review is crucial for an assessment's validity, fairness, and reliability. It aligns assessment items to learning objectives and cognitive complexity levels such as Depth of Knowledge (DOK). Traditionally, such as in state assessments, test blueprints are content based maps and used to define how many items are dedicated to each content domain. The WIN Work Ready Assessments use a progression of skills and cognitive demand. They are designed with Achievement Levels 1-5 aligned to O*NET Job Zones 1-5 with roughly equal items aligned to each achievement level.

2.1 Coverage of Learning Objectives

The learning objectives stated in the test blueprints encompass all the learning objectives that underpin both the curriculum and the assessments. The learning objectives reflect the most common (most in-demand) foundational work readiness skills required across the nearly 1,000 occupations classified by O*NET. The learning objectives are comprehensive for the curriculum and the assessments but are not occupation specific and therefore do not reflect all O*NET defined discrete skills.

Because all learning objectives are represented on the test blueprints, the Subject Matter Experts (SMEs) did not need to evaluate and ensure balanced weighting across content domains (reporting categories). Rather, each achievement level comprises approximately 20% of each assessment. Table 1 through Table 3 summarize the number of learning objectives and items aligned with each achievement level as reflected in the provided *New Blueprint and Test Form Composition*.

Table 1. Test Blueprint for Work Ready Math

Achievement Level	Test Blueprint	
	Number of Objectives	Number of Items per Achievement Level
1	7	8
2	7	8
3	7	8
4	6	8
5	4	8
Total Number of Items		**40**

Table 2. Test Blueprint for Work Ready Reading

Achievement Level	Test Blueprint	
	Number of Objectives	Number of Items per Achievement Level
1	5	8
2	6	8
3	5	8
4	5	8
5	4	8
Total Number of Items		**40**

Table 3. Test Blueprint for Work Ready Data

Achievement Level	Test Blueprint	
	Number of Objectives	Number of Items per Achievement Level
1	4	6
2	5	6
3	6	7
4	7	7
5	3	6
Total Number of Items		**32**

Recommendation: N/A

2.2 Depth of Knowledge Distribution

The WIN Work Ready Assessments use a progression of Achievement Levels 1-5 aligned to O*NET Job Zones 1-5 (National Center for O*NET Development, 2026), which are the foundations for scoring, standard setting, credential attainment, and reporting. The framework for each of the assessments has increasing cognitive complexity built into the learning objectives. Each achievement level builds on the prior level, with the application of skills increasing with each level.

The assessments do not specifically employ Webb's Depth of Knowledge (DOK) as an operational assessment framework. The test blueprints and score reports do not classify items by DOK level or provide DOK distributions. However, in the Test Maps, all items are assigned to a DOK level, indicating that DOK is actively used during item development and blueprinting.

Recommendation: N/A

3. Evaluation of Overall Validity

Content validity is essential for ensuring an assessment accurately measures the intended knowledge and skills (Bandalos, 2018). It involves a thorough evaluation of the items and domains to ensure they represent the target domain. This review ensures that the information gathered from administering the assessment is relevant and minimizes construct-irrelevant variance. Furthermore, content specification and item review help to ensure that the full range of the construct(s) is measured, minimizing construct underrepresentation.

To verify content validity, subject matter experts (SMEs) compared the WIN Work Ready Assessments with the provided *New Blueprint and Test Form Composition* tables (December 2025). The assessments were reviewed for item alignment to learning objectives and Depth of Knowledge (DOK) levels. Two SMEs independently reviewed each item for objective alignment. A third SME resolved any discrepancies. The panel of SMEs then held a consensus meeting to finalize alignment recommendations. The internal structure of the assessments was reviewed by an educational measurement expert at the University of Georgia.

3.1 Coverage by Achievement Level

Items on the WIN Work Ready Assessments were reviewed for alignment to achievement level and evaluated against each blueprint. First, the provided *Test Maps* were evaluated against each *New Blueprint and Test Form Composition* tables. Then, all items were reviewed for alignment to each achievement level.

Table 4 through Table 6 summarize the number of items aligned with each achievement level as reflected in the provided *Test Map*s.

Table 4. Coverage by Achievement Level for Work Ready Math

Achievement Level	Test Blueprint		Test Map	
	Number of Objectives	Number of Items per Achievement Level	Number of Items per Achievement Level	% of Assessment
1	7	8	8	20%
2	7	8	8	20%
3	7	8	8	20%
4	6	8	8	20%
5	4	8	8	20%
Total Number of Items			**40**	**100%**

Table 5. Coverage by Achievement Level for Work Ready Reading

Achievement Level	Test Blueprint		Test Map	
	Number of Objectives	Number of Items per Achievement Level	Number of Items per Achievement Level	% of Assessment
1	5	8	8	20%
2	6	8	8	20%
3	5	8	8	20%
4	5	8	8	20%
5	4	8	8	20%
Total Number of Items			**40**	**100%**

Table 6. Coverage by Achievement Level for Work Ready Data

Achievement Level	Test Blueprint		Test Map	
	Number of Objectives	Number of Items per Achievement Level	Number of Items per Achievement Level	% of Assessment
1	4	6	6	19%
2	5	6	6	19%
3	6	7	7	22%
4	7	7	7	22%
5	3	6	6	19%
Total Number of Items			**32**	**100%***

**due to rounding at each level, total does not equal 100%*

The number of items on the WIN Work Ready Assessments by achievement level align with the number stated on the test blueprints.

Math: All learning objectives on the test blueprint are represented on both forms of the math assessment.

Reading: All learning objectives on the test blueprint are represented on both forms of the reading assessment.

Data: All learning objectives on the test blueprint are represented on both forms of the data assessment.

Recommendation: N/A

3.2 Alignment to Learning Objectives

All items on the WIN Work Ready Assessments were reviewed for alignment to learning objectives. Table 7 highlights items flagged during review. There were no items flagged for learning objective alignment in the Work Ready Math and Work Ready Reading assessments.

Table 7. Alignment to Learning Objectives for Work Ready Data

Form	Item #	Objective on Test Map	Suggested Re-alignment	Notes
100A	20	DGI4.1	DAD4.5	Item does not align to the sources stated in DGI4.1: interviews, surveys, and focus groups. Recommend alignment to DAD4.5: trends and relationships between variables.
200B	11	NA	NA	Item does not have a correct answer, review content. The key is stated as D but there is no error; Katie Kite's rental fee is correct. The rental is 8 days, so 1 week + 1 day = $700.
200B	20	DGI4.1	DAD4.5	Same item has Form A, Item 20.
200B	21	NA	NA	The purpose and structure of the graphic is unclear. It appears to function as a range map or spatial display rather than a traditional chart or graph (such as coordinate grid), but key elements of the display are difficult to interpret. While Answer A appears to be the best choice because changing the order of the ranges may improve readability, the item may benefit from clearer visual design or a more explicit indication of the graphic's purpose.
200B	29	NA	NA	The correct answer is C, but choice B may also be plausible because the trend for admitted students could reasonably be extrapolated to approximately 5,100 in Year 6. The graph may support more than one defensible prediction. Consider revising the values, answer choices, or

				graph scale to ensure a single best answer is clearly supported by the data.

Recommendation: Overall, the items on the WIN Work Ready Assessments are well aligned with the stated learning objectives. Review alignment and content flag for the Data assessments.

3.3 Depth of Knowledge

All items are assigned to a DOK level in the Test Maps only, suggesting that WIN uses DOK as an internal measure of cognitive complexity during item development and blueprinting, while achievement levels are used for external reporting and credential decision-making.

Items on the Test Map are classified as DOK 1, DOK 2, or DOK 3. This evaluation uses the following descriptions as guidelines for DOK. As DOK levels increase, the cognitive demand on students also increases; higher DOK levels require more than just recalling facts; they require deeper understanding, application, analysis, and synthesis.

- Level 1. Recall and Reproduction: This level requires students to recall basic facts, information, definitions, terms, or perform simple, routine procedures.
- Level 2. Skills and Concepts: This level requires engaging in mental processing beyond simple recall. Students need to apply concepts, use skills, and make decisions. It requires understanding and using knowledge.
- Level 3. Strategic Thinking: This level requires deep understanding, planning, using evidence, and more complex reasoning. Students must analyze, evaluate, and draw conclusions. The cognitive demands are more abstract and require justification.
- Level 4. Extended Thinking: This level requires students to make connections, relate ideas within or among content areas, and select or devise an approach to solve a problem. It often involves extended time and requires synthesis and in-depth analysis.

Items on the WIN Work Ready Assessments were reviewed for DOK distribution and evaluated against each Test Map.

3.3.1 Distribution of DOK Levels

Table 8 through Table 10 summarize the number and percentage of items aligned with each DOK level as reflected in the provided Test Map.

Table 8. DOK Distribution for Work Ready Math (Test Map)

DOK	Form 1		Form 2	
	Number of Items	% of Assessment	Number of Items	% of Assessment
1	7	17.5%	6	15%
2	27	67.5%	28	70%
3	6	15%	6	15%
Total	**40**	**100%**	**40**	**100%**

Table 9. DOK Distribution for Work Ready Reading (Test Map)

DOK	Form 1		Form 2	
	Number of Items	% of Assessment	Number of Items	% of Assessment
1	11	27.5%	13	32.5%
2	13	32.5%	13	32.5%
3	16	40%	14	35%
Total	**40**	**100%**	**40**	**100%**

Table 10. DOK Distribution for Work Ready Data (Test Map)

DOK	Form 1		Form 2	
	Number of Items	% of Assessment	Number of Items	% of Assessment
1	4	12.5%	2	6.25%
2	16	50%	18	56.25%
3	12	37.5%	12	37.5%
Total	**32**	**100%**	**32**	**100%**

Recommendation:

The DOK distributions for the WIN Work Ready Assessments appear reasonable and consistent with the intended purpose of a workforce-readiness assessment. While the intended proportion of DOK 3 items is substantial in each section, item-level reviews flag many items that function operationally as DOK 2. As a result, a full form-level analysis is recommended to determine whether the delivered assessments adequately reflect the cognitive complexity is intended.

> Math: The Math assessments demonstrate balanced DOK distribution with strong emphasis on application of math concepts and procedural problem-solving (DOK 2). The assessments focus on applying formulas, proportional reasoning, interpretation of information, data, graphs, and decision-making.

Reading: The Reading assessments suggest rigorous assessments with a substantial emphasis on DOK 3 reasoning. The item review in the following section suggests that many designated DOK 3 items function operationally as DOK 2. As a result, the operational forms may not fully achieve the intended distribution of higher complexity.

Data: The Data assessments place substantial emphasis on DOK 3 reasoning and evidence-based interpretation. However, similar to Reading, several items classified as DOK 3 appeared to function operationally as DOK 2. Consequently, the implemented forms may contain a lower proportion of DOK 3 items than intended.

3.3.2 Review Items for DOK Level

All items on the WIN Work Ready Assessments were reviewed for DOK level.

Evaluation: Table 11 through Table 13 highlights items flagged during review.

Table 11. Alignment to DOK for Work Ready Math

Form	Item #	DOK on Test Map	Suggested DOK	Notes
100A	22	1	2	Item asks students to find the volume of cylinder, then compute costs. Students apply knowledge and calculate using multiple steps. Item is similar to #21, classified as DOK 2.
200B	3	1	2	Item asks students to analyze multiple pieces of information, verify totals, identify discrepancies, and evaluate a business decision. The item requires more than recall or simple computation. Item is similar to Form 100A #8, classified as DOK 2.

Table 12. Alignment to DOK for Work Ready Reading

Form	Item #	DOK on Test Map	Suggested DOK	Notes
100A	20	3	2	Item asks students to analyze how word choice affects readers, but the distractors are relatively weak and only one interpretation is supported. Students are not required to evaluate multiple reasonable interpretations.

Form	Item #	DOK on Test Map	Suggested DOK	Notes
100A	22	3	2	Item asks students to analyze purpose and audience to identify where the passage would most likely appear, but there is only one clear fit. The distractors are implausible, limiting strategic reasoning.
100A	25	3	2	Item asks students to determine the purpose of a company letter. The item approaches DOK 3, but students are not required to justify how specific details or sections of the letter support the purpose.
100A	26	3	2	Item asks students to analyze how a word contributes to the meaning of the letter. The item stem targets abstract vocabulary analysis, but only one interpretation aligns with the context, limiting the need for deeper evaluation.
100A	29	3	2	Item asks students to identify words contributing to an optimistic tone. To increase the rigor, item could require students to justify their conclusion or compare multiple plausible tone interpretations and determine which is better supported by evidence.
100A	30	3	2	Item asks students to determine how the author supports a claim. Item approaches DOK 3 because students evaluate evidence; however, the distractors are easily eliminated and do not require students to evaluate the relative strength of multiple plausible pieces of evidence.
100A	33	3	2	Item asks students to identify evidence supporting a statement/idea. The item stem resembles evidence-based reasoning, but students are only required to match the statement to a supporting detail. Students do not evaluate multiple or competing pieces of evidence.
100A	34	3	2	Item asks students to determine which visual would add to the meaning of the

Form	Item #	DOK on Test Map	Suggested DOK	Notes
				passage. Item stem approaches DOK 3 reasoning, but only one choice clearly aligns with the passage's focus, reducing overall complexity.
100A	35	3	2	Item asks students to determine the format of a passage and explain its impact on readers. The second part of the item stem ("effect") suggests DOK 3 reasoning, but the choices combine structure and effect, allowing students to identify the correct choice without deeper reasoning.
100A	37	3	2	Item asks students to determine how an expert quote supports the author's purpose. Item stem approaches DOK 3 analytical thinking, but students are not required to explain why the quote supports the purpose better than alternative evidence.
200B	10	2	1	Item asks students to recall the meaning of "Open Enrollment" using information stated directly in the memorandum. Students locate the definition explicitly provided in the passage.
200B	20	3	2	Item asks students to analyze how word choice affects the tone, but the distractors are relatively weak and only one interpretation is supported. Students are not required to evaluate multiple reasonable interpretations or justify conclusions.
200B	22	3	2	Item asks students to identify the tone of passage. The item stem is attempting higher-level tone analysis, but the correct response is supported by explicit features of the passage, and the distractors are comparatively weak.

Form	Item #	DOK on Test Map	Suggested DOK	Notes
200B	25	3	2	Item asks students to identify the intended audience based on phrases. Students use contextual clues and document purpose to make an inference. The item does not require students to justify how multiple details support the audience determination or evaluate alternative audiences.
200B	28	3	2	Item asks students to determine the main purpose of the listed activities. Students synthesize information from a section of the passage, but the answer is a straightforward interpretation of the activities presented. The item does not require evaluation of competing purposes or supporting evidence.
200B	30	3	2	Item asks students to determine which activity would not contribute to the requirements. Students compare an activity against explicitly stated criteria and identify the mismatch. The item does not require strategic reasoning or evaluation.
200B	34	3	2	Item asks students to identify the effect of formatting. The item stem targets analysis of text structure; however, one answer clearly aligns with the purpose of the formatting while the distractors are relatively implausible.
200B	38	3	2	Item asks students to identify evidence supporting the author's claim. The item stem approaches evidence-based reasoning, but students are only required to match a claim to a supporting detail explicitly stated in the passage.

Table 13. Alignment to DOK for Work Ready Data

Form	Item #	DOK on Test Map	Suggested DOK	Notes
100A	1	2	1	Item asks students to identify the purpose of research questions. Students recall a basic concept about information gathering.
100A	19	3	2	Item asks students to interpret a graph and identify the occupation with the strongest employment opportunity. Students compare several data categories, but the item does not require them to evaluate multiple indicators or justify why one occupation represents the best opportunity.
100A	23	3	2	Item asks students to analyze a graph containing multiple variables and identify an industry that meets specific criteria. The item approaches DOK 3 because students must consider multiple data relationships simultaneously; however, students are only required to locate the correct industry rather than explain how the variables interact to support the conclusion.
100A	24	3	2	Item asks students to determine the relationship between advertising and sales. Students interpret the trends and relationships in a graph, but the item does not require DOK 3 level reasoning such as evaluating alternative explanations, justifying the strength of the relationship, or determining limitations in the data. Item is similar to #10 and #29.
100A	26	3	2	Item asks students to determine which conclusion is best supported by a graph. The item approaches DOK 3 because students evaluate data and compare conclusions; however, the distractors are relatively easy to eliminate and do not require students to weigh the relative

				strength of multiple plausible interpretations.
100A	31	3	2	Item asks students to identify an implication supported by survey data. The item stem targets higher-order reasoning by requiring students to move beyond direct observation; however, only one conclusion is plausible, limiting the need for deeper evaluation.
100A	32	3	2	Item asks students to compare data from two time periods and determine the most likely implication. The item approaches DOK 3 because students identify patterns and draw conclusions; however, students are not required to evaluate alternative interpretations of the data or explain the evidence supporting their conclusion.
200B	16	2	3	Item asks students to determine which survey response is related to a research question but not relevant to answering it. Students must evaluate the relationship between the research question and multiple pieces of information and distinguish between related and relevant evidence, aligned with DOK 3 reasoning.
200B	19	3	2	Item asks students to draw a conclusion supported by a graph showing job openings and unemployment. Students interpret trends, but the item does not require them to evaluate the strength of multiple plausible conclusions.
200B	23	3	2	Item asks students to determine the relationship between fuel economy and various driving speeds. Students interpret the overall pattern of the graph, but the item does not require DOK 3 level reasoning such as explaining how the variables interact or justifying one interpretation over competing choices.
200B	24	3	2	Item asks students to determine the relationship shown in a graph comparing age ranges across employment

				arrangements. Item does not require deeper evaluation beyond the visible relationships.
200B	26	3	2	Item asks students to determine the trends shown in a graph of the number of movers and mover rates over time. The answer is straightforward and does not require reasoning beyond the visible relationships.
200B	31	3	2	Item asks students to identify the most likely implication supported by data across multiple graphs. Students move beyond direct observation and make a modest inference from the graphs. However, students are not required to evaluate competing explanations or justify why one implication is better supported than another.

Recommendation:

Overall, the WIN Work Ready Assessments items varied in their alignment with the stated DOK levels. Most flags involved items designated as DOK 3 that appeared to function operationally as DOK 2 because students could arrive at the correct answer through straightforward or concrete interpretation, application of stated criteria, or identification of a clearly supported answer. A recurring pattern was that items intended to measure strategic thinking but did not require students to engage in the evaluation, justification, comparison, or evidence-based reasoning typically associated with DOK 3 items.

The most DOK alignment flags occurred in the Reading and Data assessments.

Math: Review the two flagged items and either update the associated metadata or revise the items to more accurately reflect the intended DOK classification. The Work Ready Math assessment generally demonstrates appropriate alignment between the intended and stated DOK levels. Overall, no substantial DOK alignment concerns were identified.

Reading: Although the test map identifies several Reading items as DOK 3, many function operationally as DOK 2. The item stems often attempt to target strategic thinking or abstract reasoning; however, the answer choices limit the cognitive demand of the task because the key is clearly supported by the text while the distractors are comparatively weak. As a result, students are required to identify and interpret parts of the text rather than evaluate, justify, and compare competing ideas that are typically associated with DOK 3 items.

Review the flagged items and consider revisions that require students to engage in strategic thinking or abstract reasoning. For example, items addressing author purpose, tone, and word choice could be strengthened by asking students to evaluate multiple plausible interpretations and determine which is best supported by the text. These revisions would more closely align item complexity with DOK 3 expectations for strategic thinking, evaluation, and evidence-based reasoning.

Data: Although the test map identifies several Data items as DOK 3, many function operationally as DOK 2. The item stems often attempt to target strategic thinking through data interpretation, trend analysis, prediction, and evidence-based conclusions; however, the answer choices frequently reduce the cognitive demand because the correct response can be identified from a straightforward reading of the graph, chart, or dataset while the distractors are comparatively weak. As a result, students are required to identify patterns, interpret information, and select supported conclusions rather than evaluate, justify, or analyze the relative strength of multiple conclusions/ explanations that are typically associated with DOK 3 items.

Review the flagged items and consider revisions that require students to engage in stronger evidence-based reasoning. For example, items involving trend analysis, predictions, implications, and conclusions could be strengthened by asking students to evaluate multiple plausible interpretations, determine which conclusion is best supported by the available data, compare competing explanations, or analyze how multiple variables interact to influence outcomes. These revisions would more closely align item complexity with DOK 3 expectations for strategic thinking, evaluation, and data-driven reasoning.

In addition to the item-level DOK review, a comprehensive, holistic review is recommended to determine whether each assessment adequately reflects the cognitive demands and distribution parameters specified in the test blueprint. Consider the cumulative impact of any proposed DOK reclassifications, as shifts across multiple items may affect the overall balance of cognitive complexity represented within and across forms.

3.4 Internal Structure

According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014), validity evidence based on internal structure relates to how test items, individually and collectively, align with the construct(s) being measured. To that end, this evaluation included summaries of classical test theory and Rasch model item analyses, dimensionality, reliability, and measurement invariance. No additional analysis was conducted. Rather, materials provided by SCDE and WIN, including the technical report, field testing summary, and psychometric statistics files, are summarized and interpreted.

3.4.1 CTT Item Statistics

A classical test theory (CTT) item analysis is conducted as a baseline check for the appropriateness of items. To support psychometric analyses and ability estimation, items need to

be sufficiently difficult for the population and function properly. CTT item difficulty is reported as the proportion of examinees who answered an item correctly. Lower values indicate harder items (e.g., 0.15 (15%) of the examinees answered the item correctly), and higher values indicate easier items (e.g., 0.85 (85%) of examinees answered the item correctly). For these assessments, a wide range of difficulty values should be observed to indicate utility across the breadth of examinee ability levels. Additionally, values near 0.50 provide strong information (Bandalos, 2018). One measure of CTT item quality is the item-total score (also termed point-biserial) correlation. The item-total correlation is a quantification of the degree to which individual items separate examinees with low and high scores. Values greater than 0.30 are considered satisfactory, while values less than 0.20 indicate low discrimination and suggest revision of the item (Nunnally & Bernstein, 1994). Negative values indicate that higher scoring examinees tend to get the item incorrect more often than low scoring examinees, which is a flag for immediate item review.

Evaluation: Results from CTT analysis are displayed in Table 14.

Overall, each assessment has an appropriate range of item difficulties to properly assess students across the span of ability levels. Additionally, most item-total correlations are above 0.30, indicating that items discriminate adequately. The field test summary document indicates stringent and appropriate criteria for item flagging, review, and revision, including flags for items that are too easy for the sample, too difficult for the sample, and items with low discriminatory power. Altogether, the difficulty and item-total correlation estimates provide preliminary evidence that the items are functioning properly for the math, reading, and data assessments.

Recommendation: N/A

Table 14. Classical Test Theory Item Statistics

Assessment	Number of Items	Mean *p*-value (range)	Mean Item-Total Correlation (range)	Percentage of Items Flagged
Work Ready Math (Field Test)	75	0.48 (0.08 - 0.96)	0.31 (-0.23 - 0.64)	31%
Work Ready Reading (Field Test)	215	0.61 (0.14 - 0.96)	0.37 (-0.06 - 0.59)	14%
Work Ready Data (Field Test)	67	0.53 (0.12 - 0.89)	0.31 (-0.05 - 0.57)	24%
Work Ready Math (Operational Form 1)	40	0.57 (0.22 - 0.96)	0.41 (0.23 - 0.55)	8%*
Work Ready Math (Operational Form 2)	40	0.57 (0.20 - 0.95)	0.43 (0.21 - 0.64)	8%*

Work Ready Reading (Operational Form 1)	40	0.65 (0.36 - 0.89)	0.43 (0.24 - 0.59)	0%
Work Ready Reading (Operational Form 2)	40	0.65 (0.38 - 0.94)	0.42 (0.22 - 0.57)	5%*
Work Ready Data (Operational Form 1)	32	0.61 (0.31 - 0.87)	0.38 (0.17 - 0.57)	3%
Work Ready Data (Operational Form 2)	32	0.59 (0.34 - 0.84)	0.38 (0.15 - 0.58)	3%

Note. Blue shaded rows are field tests. Green shaded rows are operational. *Some flagged (easy) items were included to cover learning objectives.

3.4.2 1-PL Item Statistics

The one-parameter logistic (1-PL) item response theory model was used for calibration and scaling. While it makes strong assumptions about items, the 1-PL model has preferrable measurement properties including sum-score sufficiency, invariant item ordering, and a common item-to-ability scale (Engelhard, 2013). These properties are useful for interpretation of items, ability estimates, and performance level classifications. The 1-PL model includes ability estimates and item parameter estimates. The 1-PL model assumes that each item discriminates equally and estimates a difficulty parameter for each item. This difficulty parameter is a location parameter, indicating the point on the ability scale where an examinee has a 50% chance of answering the item correctly. For the WIN assessment purposes, difficulty parameters should span the ability distribution (e.g., -3 to 3) to reliably locate all examinees.

When applying the 1-PL model, it is important to assess the degree to which the data fits the model. 1-PL model fit statistics, including infit and outfit, residual-based item fit statistics, and chi-square tests can quantify the differences between observed responses and model-predicted responses.

Evaluation: Results from 1-PL model psychometric analyses are displayed in Table 15. Overall, each assessment has a sufficient range of item difficulties to properly assess students across the span of ability levels.

Table 15. 1-PL IRT Model Item Parameters

Assessment	Number of Items	Mean 1-PL Difficulty (Range)
Work Ready Math (Field Test)	75	0.23 (-3.54 - 3.62)
Work Ready Reading (Field Test)	215	0.00 (-3.00 - 2.81)
Work Ready Data (Field Test)	67	0.32 (-2.09 - 3.04)
Work Ready Math (Operational Forms)	40	-0.83 (-3.98 - 1.28)
Work Ready Reading (Operational Forms)	40	-1.10 (-3.54 - 0.83)
Work Ready Data (Operational Forms)	32	-0.60 (-2.22 - 1.03)
Note. Blue shaded rows are field tests. Green shaded rows are operational.		

Recommendation: Each operational form's model and item fit statistics should be included for a complete evaluation. Additionally, if feasible within blueprint constraints, more difficult items (b > 1.2) could be included to more precisely scale and classify higher ability examinee.

3.4.3 Dimensionality

The item response theory models used to scale the WIN assessments have an underlying assumption of unidimensionality. That is, each test measures a single domain. If this is not the case, and more than one factor contributes to item responses, then the validity and interpretation of scale scores are called into question. To assess dimensionality, common approaches are exploratory factor analysis (EFA) and principal components analysis (PCA). Within both frameworks, when the first component accounts for substantially more variance than subsequent components, it is an indicator that the first factor is dominant and that the unidimensionality assumption has been satisfied (Zopluoglu & Davenport, 2017).

Evaluation: Results from EFA visualizations (scree plots) of the dimensionality of each assessment indicated that for each assessment, the first factor was dominant, and the unidimensionality assumption was met.

Recommendation: N/A

3.4.4 Reliability

Test score reliability is concerned with the consistency and precision of scores and is a function of the amount of measurement error (Wells & Wollack, 2003). Reliability is a necessary condition for validity because if scores are highly variable and error-ridden, they cannot be said to measure the construct(s) accurately. There are several ways to quantify reliability in the CTT framework, including Cronbach's alpha and the standard error of measurement. Cronbach's alpha ranges from 0 to 1 and quantifies the degree to which the items consistently measure the target domain. In high-stakes settings, alpha should be approaching or above 0.90 (Wells & Wollack, 2003). The standard error of measurement (SEM) provides an interval estimate around raw scores.

Evaluation: For the WIN assessments, Cronbach's alpha reliability estimates were strong, ranging from 0.84 - 0.91. SEM values ranged from 2.42 to 2.69. Raw score SEMs in this range are adequate considering there were 32 – 40 operational items. In summary, reliability is adequate for the WIN assessment scaling purposes.

Recommendation: Reliability estimates could be provided at the subgroup level for additional evidence of fairness and validity. Reliability estimates should also be provided at the achievement level cut-scores (e.g., conditional SEMs at cut scores).

3.4.5 Measurement Invariance

Test fairness is a fundamental aspect of conducting group comparisons and ensuring the validity of assessments, particularly when examining differences based on gender, ethnicities, culture, or treatment conditions. To achieve test fairness, it is essential to detect and prevent any form of unfairness throughout the entire testing process, including test design, development, administration, and scoring (Camilli, 2006). When a test is free of systematic bias, measurement invariance has been met.

Differential item functioning (DIF) analysis plays a crucial role in addressing test fairness by identifying potentially biased items in a test. DIF procedures assess whether examinees from different subgroups, who possess the same underlying ability or trait, have different probabilities of endorsing an item (Angoff, 1993). By identifying items that function differently across groups, DIF analysis helps to minimize the impact of factors unrelated to the construct being measured (Sireci & Rios, 2013). Biased items systemically advantage or disadvantage a specific subgroup because of factors irrelevant to the intended construct. By addressing DIF, the fairness and validity of the test can be enhanced, ensuring that an item is unbiased and measures the same construct across groups.

From a psychometric perspective, DIF is commonly analyzed using methods that compare item performance across groups. For the WIN assessments, subgroups of interest could be gender (male/female), racial/ethnic groups, disability status (no/yes), multilingual status (no/yes), and different form accommodations. To measure DIF, the Mantel-Haenszel (MH) delta statistic

quantifies the difference in item response distribution for two groups. Researchers at ETS developed thresholds to interpret MH delta values (Zwick, et al., 1999).

Evaluation: In the test development phase, item writers followed guidelines for fairness and sensitivity to minimize bias under Universal Design principles, and each item was reviewed by professional editors and the executive content team. Due to insufficient subgroup sizes, DIF analysis has not yet been completed.

Recommendation: Evaluation to be completed after DIF analysis is conducted.

3.4.6 Equating Multiple Forms

For each of the three WIN assessments, WIN administers two forms. This is done primarily for security purposes. When multiple forms of an assessment are administered, steps must be taken to ensure that all forms provide valid and reliable scores, are comparable, and do not bias against any individuals. Equating is the statistical process of adjusting scores from different forms so they can be used interchangeably (Kolen & Brennan, 2014). For example, if Form A is more difficult than Form B, then a score of 75 on Form A may correspond to a score of 84 on Form B. Equating is the statistical process of finding the equivalent score adjustments.

For the WIN assessments, all items are calibrated using equivalent groups and are therefore on the same scale. Using the 1-PL IRT model, each total score corresponds to an ability estimate and scale score. With the total score – ability correspondence for each form, the forms can be equated by mapping ability estimates on each form to the corresponding scores on each form.

Evaluation: The process outlined in the technical report is sound. For each assessment, forms were very similar in terms of difficulty, so equating adjustments were minimal. The true score equating process was applied appropriately using the 1-PL IRT model. Based on the process applied, different forms can be used interchangeably with no concerns with threats to validity or bias.

Recommendation: After future administrations, post-equating checks should be conducted to ensure that the 1-PL difficulty parameter estimates remain appropriate and support a valid equating process.

4. Achievement Level Classifications

For summative assessments, in addition to scale scores, it is often useful to provide achievement level classifications that are coupled with interpretable descriptors of skill and understanding development for each achievement level. In an item response theory framework, cut scores for each performance level must be determined and applied to classify examinees. For the WIN assessments, cut scores were determined in a standard setting process and applied to classify examinees into one of five achievement levels:

- **Job Zones 1-2:** occupations that need very little to some preparation
- **Job Zone 3:** occupations that need medium preparation
- **Job Zone 4:** occupations that need considerable preparation
- **Job Zone 5:** occupations that need extensive preparation

This evaluation concerns the degree to which these classifications are valid and reliable. Data for this section comes from the technical report (WIN, 2026).

Evaluation: A modified Angoff standard setting was used to determine the WIN assessment achievement level cut scores. This method is commonly applied in industry and accepted as a valid procedure for standard settings (Cizek & Bunch, 2007). The cut scores were then applied to the ability estimates derived from the 1-PL scaling process to classify students into the achievement levels. Classification accuracy estimates ranging from 0.75 - 0.78 and consistency indices ranging from 0.66 - 0.71 indicate acceptable consistency for classification into the five achievement levels.

Recommendation: As noted earlier, there were very few items located near Level 3/4 and Level 4/5 operational cut scores. More difficult items could be added to improve accuracy and consistency of Level 4 and Level 5 classifications.

References

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

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Lawrence Erlbaum Associates.

Bandalos, D. L. (2018). Measurement theory and applications for the social sciences. Guilford Publications: New York.

Camilli, G. (2006). Test fairness. *Educational measurement*, *4*, 221-256.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting*. SAGE Publications, Inc., https://doi.org/10.4135/9781412985918

Engelhard, G. (2013). *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences*, 1st edition. Routledge/Taylor & Francis Group.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices,* 3rd ed. Springer Science + Business Media.

National Center for O*NET Development. O*NET OnLine Help: Job Zones. *O*NET OnLine*. Retrieved from https://www.onetonline.org/help/online/zones

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*, 3rd ed. McGraw-Hill.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2-3), 170-187.

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*, *28*(March), 1-9.

Wells, C. S., & Wollack, J. A. (2003). An Instructor's Guide to Understanding Test Reliability. Testing & Evaluation Services. University of Wisconsin.

Worldwide Interactive Network (2026). *WIN Work Ready Assessments: Technical Report, Version 6.2026*. Retrieved from WIN associates.

Zopluoglu, C., & Davenport, E. C. (2017). A note on using eigenvalues in dimensionality assessment. *Practical Assessment, Research, and Evaluation, 22*(7). Available online: http://pareonline.net/getvn.asp?v=22&n=7.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28.