Wisconsin Forward Exam

Spring 2020
Technical Report

WISCONSIN DEPARTMENT OF PUBLIC INSTRUCTION

Submitted to
Wisconsin Department of Public Instruction
August 2020

DATA RECOGNITION
DRC
CORPORATION

Copyright

Developed and published under contract with the Wisconsin Department of Public Instruction by Data Recognition Corporation, 13490 Bass Lake Road, Maple Grove, MN 55311.

Foreword

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

TABLE OF CONTENTS

Copyright.....1

Foreword.....2

Executive Summary.....5

E.1 Overview of the Wisconsin Forward Exam.....5
E.2 Note about Cancelled Test Administration.....6
E.3 Content of the Technical Report.....6

Part 1: Overview.....7

1.1 Historical Background.....7
1.2 Uses of Test Scores.....9
1.2.1 Test-Level Scores.....10
1.2.2 Standard-Level Subscores and Performance Levels.....11
1.3 Technical Report Structure.....12

Part 2: Validity Framework.....14

2.1 Sources of Validity Evidence.....14
2.2 Summary of Validity Evidence for Wisconsin Forward Exam Based on Test Content.....16

Part 3: Test Content and Test Development.....17

3.1 Test Blueprints.....18
3.2 Test Design.....18
3.2.1 English Language Arts.....18
3.2.2 Mathematics.....19
3.2.3 Science.....19
3.2.4 Social Studies.....19
3.3 Universal Design.....20
3.4 Item Development Process.....21
3.4.1 Reading Passage and Item Reviews—Summer 2018.....22
3.5 Field-Testing—Spring 2019.....23
3.5.1 Statistical Analysis of Field Test Data.....23
3.5.2 Item Data Review—Summer 2019.....24
3.6 Form Development Process.....25
3.6.1 Item Selection.....25
3.7 Item and Form Quality Reviews.....27
3.7.1 DPI Approvals.....28
3.8 Summary.....28

Part 4: Summary Recommendations.....40

References.....42

APPENDICES

Appendix A: Summer 2018 Item Review Training Slides.....43

Appendix B: Summer 2019 Field Test Data Review Training Slides.....72

LIST OF TABLES

PART 3

Table 3-1 English Language Arts Test Blueprints for Grades 3–8 30
Table 3-2 Mathematics Test Blueprints for Grades 3–8 31
Table 3-3 Science Test Blueprints for Grades 4 and 8 32
Table 3-4 Social Studies Test Blueprints for Grades 4, 8, and 10 32
Table 3-5 Item Type Descriptions for Items on the Wisconsin Forward Exam........ 33
Table 3-6 English Language Arts Test Design 34
Table 3-7 Mathematics Test Design 35
Table 3-8 Science Test Design 36
Table 3-9 Social Studies Test Design 36
Table 3-10 Elements of Universal Design 37
Table 3-11 College- and Career-Ready Item Bank Development Activities 38
Table 3-12 Items Reviewed at Summer 2018 Item Review 38
Table 3-13 Items Reviewed at Summer 2019 Data Review 39

Executive Summary

This document provides an overview of the Wisconsin Forward Exam and a summary of work leading to the Spring 2020 administration of the English Language Arts (ELA), Mathematics, Science, and Social Studies assessments. An explanation of assessment cancellation is also included.

The Wisconsin Forward Exam assessments are designed to measure students' knowledge of ELA and Mathematics in grades 3 through 8, Science in grades 4 and 8, and Social Studies in grades 4, 8, and 10. The assessments are aligned with Wisconsin Academic Standards. The test forms for the Spring 2020 ELA, Mathematics, and Science administrations were developed by Data Recognition Corporation (DRC) using DRC's College- and Career-Ready item bank. The Spring 2020 Social Studies assessments contained Wisconsin-owned items and were going to be reused from the previous administration.

All assessments except for Braille and accommodated paper-based forms were planned to be administered online.

E.1 Overview of the Wisconsin Forward Exam

The Wisconsin Forward Exam is designed to measure Wisconsin Academic Standards, which define the knowledge and skills students need in each grade level to succeed later in college, other postsecondary training, and careers.

The Wisconsin ELA and Mathematics grade-level tests have undergone multiple alignment changes since their first administration in the 2005–06 school year, with the latest changes occurring in the 2015–16 administration, which was also the first administration year of the tests under the Wisconsin Forward Exam program. The current ELA and Mathematics assessments are aligned to the Wisconsin Academic Standards. The new reporting scales for the ELA and Mathematics tests were established after the Spring 2016 test administration, and the new performance level cut scores were set for these assessments in Summer 2016. The ELA and Mathematics results from the 2015–16 school year are considered a new baseline for year-to-year student performance comparisons. The subsequent assessments, including the last test administration in the 2018–19 school year, were statistically linked to the established scales, allowing for year-to-year test score comparability.

The Science assessments have been on a different trajectory. A change to the Science test blueprint and design was made for the Spring 2019 operational test administration. New Science tests, aligned to the new Wisconsin Standards for Science and the Next Generation Science Standards, were developed and administered to Wisconsin students for the first time in Spring 2019. Due to the change of standards, new scales were developed for the new Science tests, and new performance level cut scores were set in Spring 2019, which was the most recent Science test administration.

The Social Studies assessments continue to be aligned with the Wisconsin Model Academic Standards. New scales were developed for the Social Studies tests under the new Wisconsin

Forward Exam program in Spring 2016. Following the new scale development, the new performance level cut scores were set for Social Studies in Summer 2016. The subsequent Social Studies assessments, including the last test administration in the 2018–19 school year, were statistically linked to the established scales, allowing for year-to-year test score comparability.

All Wisconsin assessments are developed for online administration and contain various item types, including multiple-choice (MC), multi-select (MS), technology-enhanced (TE), evidence-based selected response (EBSR), short-answer (SA), and, in ELA, text-dependent analysis (TDA) items. Braille, print-on-demand, and Spanish translation forms that contain the same items as regular online operational test forms are also available to students who need them.

E.2 Note about Cancelled Test Administration

The Wisconsin Forward Exam test administration window was scheduled to last from March 23 to May 1, 2020. However, due to the COVID-19 pandemic, all schools were closed in March 2020, before the testing window opened. While Wisconsin students continued to receive instruction and learn online through the rest of the school year, the Wisconsin Forward Exam was cancelled for the 2019–20 school year. The Wisconsin Department of Public Instruction received a waiver from the federal government and state laws were rewritten to suspend the requirement for students to participate in the end-of-grade standardized assessments. Students in grades K–12 are expected to return to school for the 2020–21 school year, and the next administration of the Wisconsin Forward Exam is planned for Spring 2021.

E.3 Content of the Technical Report

A typical Technical Report documents all aspects of the test development, administration, and reporting cycle. While the test forms were developed for the Spring 2020 test administration, the test administration itself was cancelled and students did not get an opportunity to demonstrate their knowledge and skills in ELA, Mathematics, Science, or Social Studies. No test scores were reported in Spring 2020. Therefore, this document is an abbreviated Technical Report with its content limited to the description of the test content and development.

Part 1 of this Technical Report includes an overview of the Wisconsin testing program and the types of test scores available for the Wisconsin Forward Exam. Part 2 of the report provides the validity framework and a summary of the validity evidence based on the content of the Wisconsin Forward Exam test forms developed for the Spring 2020 administration. Part 3 of the report includes details on the test content, test design, and test development cycle. Part 4 presents some recommendations for the Spring 2021 test administration for DPI consideration.

Due to the cancellation of the assessments, the remaining aspects of the Wisconsin Forward Exam administration and reporting cycle, including standardized administration, scoring, psychometric data analysis, standard setting, construct validity and reliability studies, and reporting of test results, are not included in this Technical Report.

Part 1: Overview

The abbreviated *Wisconsin Forward Exam Spring 2020 Technical Report* documents the processes and procedures applied in the development of the Spring 2020 assessments. This report also provides evidence in support of content-based test validity in adherence to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). This report demonstrates that the Spring 2020 Wisconsin Forward Exam test forms were developed in accordance with standards and best practices of educational assessment.

1.1 Historical Background

The Improving America's Schools Act of 1994 required that states establish challenging academic standards as well as aligned annual assessments. The Goals 2000: Educate America Act and the Elementary and Secondary Education Act (ESEA) spelled out additional requirements to ensure that citizens receive coherent information about whether and to what degree students are meeting rigorous academic standards. This Technical Report is an important part of meeting those requirements.

Wisconsin students in grades 4, 8, and 10 began taking the Wisconsin Knowledge and Concepts Examination (WKCE) norm-referenced assessments in the 1997 school year. At that time and in the following years, *TerraNova*™ tests developed by CTB/McGraw-Hill (1997, 2000, 2009) were used. The selection of those tests was partly predicated on an awareness of the academic standards being developed. In January 1998, the Wisconsin Model Academic Standards (WMAS) were adopted. These new standards were the work of the Governor's Commission on Wisconsin Model Academic Standards, chaired by then Lieutenant Governor Scott McCallum and the Wisconsin Department of Public Instruction (DPI). The assessments aligned to WMAS would measure student performance in the same subjects as the *TerraNova* tests.

Beginning in the 2005–06 school year, the federal No Child Left Behind Act (NCLB) required all states to test all students in Reading and Mathematics in grades 3 through 8 and once in high school (in grade 10 under Wisconsin law § 118.30). Based on the NCLB legislation, student performance, reported in terms of proficiency categories, was used to determine the Adequate Yearly Progress (AYP) of students at the school, district, and state levels. Beginning with the 2007–08 school year, states were also required to administer Science assessments at least once in grades 3–5, once in grades 6–9, and once in grades 10–12.

It was within this policy context that the WKCE was constructed, as a criterion-referenced test, for the Fall 2005 administration, replacing the previously existing norm-referenced WKCE in Reading and Mathematics. The criterion-referenced WKCE was designed specifically for Wisconsin students to measure their performance on the WMAS. These assessments were designed to evaluate students' knowledge and to measure achievement in the basic skills taught in schools at grades 3–8 and 10. The Fall 2013 WKCE was the ninth administration of these assessments and the last administration of Reading, ELA, and

Mathematics. The assessments in Science and Social Studies under the existing WKCE model continued to be administered until Fall 2014.

A major change in the Wisconsin assessments occurred for the 2014–15 test administration. First, the ELA and Mathematics assessments were moved from the Fall testing window to the Spring testing window. Second, the new ELA and Mathematics tests for grades 3–8 developed for the Spring 2015 administration consisted of new Smarter Balanced Assessment Consortium (SBAC) items aligned to the Common Core State Standards (CCSS). Thus, the 2014–15 ELA and Mathematics assessments were not comparable content- and construct-wise to the assessments administered in prior years. Third, while the prior years' assessments included CTB's *TerraNova* items that yielded norm-referenced scores, the 2014–15 assessments did not include such items. Fourth, the regular versions of the 2014–15 assessments were administered as fixed forms in the online mode, in contrast to the previous assessments, which were all administered in the paper-and-pencil mode. Fifth, technology-enhanced items were introduced in the 2014–15 online test administration. Last, the student test scores for ELA and Mathematics were reported on SBAC scales, and the students were classified into performance levels based on SBAC cut scores. Further details on the structure and reporting of the Spring 2015 ELA and Mathematics assessments (called the Wisconsin Badger Exam) can be found at https://dpi.wi.gov/assessment/historical/smarter.

The ELA and Mathematics assessments underwent yet another change in the 2015–16 administration year. The Wisconsin DPI partnered with Data Recognition Corporation (DRC) to develop new ELA and Mathematics assessments for grades 3–8 for the Spring 2016 administration. The items contained in these assessments were drawn from DRC's nationally field-tested College- and Career-Ready (CCR) item bank and aligned with Wisconsin Academic Standards for ELA and Mathematics. The new assessment program is called the Wisconsin Forward Exam, and the new ELA and Mathematics tests were first administered online in Spring 2016. Since the new assessments did not contain any items from the 2014–15 Wisconsin Badger Exam tests, the new scales were not statistically linked to the previous scales. The new reporting scales for the ELA and Mathematics tests were developed after the Spring 2016 test administration, and the new performance level cut scores were set for these assessments in Summer 2016.

Science (grades 4 and 8) and Social Studies (grades 4, 8, and 10) assessments have been on a different trajectory, and they have continued to be aligned with the WMAS. However, the test administration for these assessments was moved from the Fall window to the Spring window in the 2015–16 administration year. The items contained in the Science and Social Studies tests were mainly drawn from the pool of previously administered items, but new items were also included. Several of the previously administered items were edited to improve item quality and reflect test content changes over time. Despite the fact that many Science and Social Studies items in the Spring 2016 administration came from the previous item pool, statistically linking the Spring 2016 forms to the previous forms was not recommended due to the change of the testing window and the numerous changes to the items themselves. Instead, similar to what was done for the ELA and Mathematics assessments, new scales were developed for the Science and Social Studies tests under the new Wisconsin Forward Exam program. Following the new scale development, the new performance level cut scores were set for Science and Social Studies in Summer 2016.

Details regarding development, scaling, reporting, and standard setting for all Spring 2016 assessments are included in the *Wisconsin Forward Exam Spring 2016 Technical Report* available at https://dpi.wi.gov/assessment/forward/resources.

Spring 2020 was intended to be the fifth administration year for the Wisconsin Forward Exam in ELA, Mathematics, and Social Studies, using the test blueprint and test design developed for the Spring 2016 test administration. The new ELA and Mathematics tests were developed with adherence to Wisconsin's standards and, with a few exceptions, consisted of items administered to Wisconsin students in Spring 2018 and Spring 2019 as part of the operational test or a field test. Previously administered operational test items were selected to serve as linking items between the Spring 2019 test administration and the next administration, allowing the newly developed ELA, Mathematics, and Social Studies assessments to be placed on the Wisconsin Forward Exam scales using statistical equating procedures. (Test equating allows for direct comparison of student scores within a content area and for evaluation of year-to-year student performance change.) The Social Studies operational test forms intended for the Spring 2020 test administration were the same test forms that were administered in Spring 2018.

Spring 2020 was also planned to be the second administration year for the new Wisconsin Forward Exam in Science, aligned to the new Wisconsin Standards for Science (WSS) and the Next Generation Science Standards (NGSS). The new Science assessments focus on content understanding linked to work with science and engineering practices and crosscutting concepts as detailed in the National Research Council Framework for K–12 Science Education (https://www.nap.edu/read/13165/chapter/1). The items contained in the Science tests were drawn from the pool of items aligning to the new WSS and NGSS that were field-tested in Spring 2018 and 2019, as well as operational test items administered in Spring 2019.

This Technical Report documents all aspects of form development in preparation for the Spring 2020 test administration. A brief content summary of the report is provided later in this part of the report.

1.2 Uses of Test Scores

While the Wisconsin Forward Exam was not administered in Spring 2020, this section of the Technical Report serves as an overview of the uses of the test scores that students receive in a typical school year.

Validity is the overarching component of the Wisconsin Forward Exam program. The following excerpt on validity is from the *Standards for Educational and Psychological Testing* (hereafter the *Standards*) (AERA, APA, & NCME, 2014):

> Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence . . . include evidence of careful test construction;

> adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question. (p. 22)

The validity of a test score ultimately rests on how that test score is used. To understand whether a test score is being used properly, one must first understand the purpose of the test. The intended uses of the Wisconsin Forward Exam scores include the following:

- Identifying students' strengths and areas in need of improvement
- Communicating expectations for all students
- Evaluating school-, district-, and state-level programs
- Informing stakeholders (i.e., teachers, school administrators, district administrators, DPI staff members, parents, and the public) about the status of the progress toward meeting the academic achievement standards of the state
- Meeting the requirements of the state's accountability program

The Wisconsin Forward Exam reported scores include the test-level scores (scale scores and performance levels) and standard-level (objective) scores (Standard Performance Index [SPI] scores and performance levels).

1.2.1 Test-Level Scores

At the test level, an overall scale score that is based on student performance on the entire test is reported. In addition, an associated level of performance is reported. These scores indicate, in varying ways, a student's achievement in ELA, Mathematics, Science, or Social Studies. Test-level scores are typically reported at four levels: state, school district, school, and student.

Two types of test-level scores are reported to indicate a student's achievement on the Wisconsin Forward Exam: (1) the scale score and (2) its associated level of performance.

Scale Scores

A scale score indicating a student's performance is determined for each content area. The overall scale score for a content area quantifies the achievement being measured by the ELA, Mathematics, Science, or Social Studies test. In other words, the scale score represents the student's level of performance, where higher scale scores indicate higher levels of performance on the test and lower scale scores indicate lower levels of performance.

Levels of Performance

A student's performance on the ELA, Mathematics, Science, or Social Studies Wisconsin Forward Exam is reported in one of four levels of performance: *Below Basic*, *Basic*, *Proficient*, or *Advanced*. The cut scores for the levels of performance for ELA, Mathematics, and Social Studies were recommended by Wisconsin educators at the standard setting workshop in June 2016. The cut scores for Science assessments were established during the standard setting workshop in May 2019. The cut scores reflect the expectations of Wisconsin educators of what

Wisconsin students should know and be able to do in ELA, Mathematics, Science, and Social Studies (see *Wisconsin Forward Exam Spring 2019 Technical Report* posted at https://dpi.wi.gov/assessment/forward/resources#documentation for a brief description of the Wisconsin Forward Exam standard setting).

Use of Test-Level Scores

The Wisconsin Forward Exam scale scores and performance levels provide summary evidence of student achievement in ELA, Mathematics, Science, and Social Studies. Classroom teachers may use these scores as evidence of student achievement in these content areas. At the aggregate level, district and school administrators may use this information for activities such as curriculum planning.

1.2.2 Standard-Level Subscores and Performance Levels

The standard-level subscores (i.e., the SPI scores) indicate student performance on individual content standards and can be interpreted as an estimate of the number of items a student would be expected to answer correctly if there had been 100 similar items for a given reporting category. The SPI scores are criterion-referenced scores, in that they estimate how much a student knows in a clearly defined skill domain (i.e., the criterion). The SPI scores are computed for content standards measured by at least four items.

Based on their SPI scores, students are classified in one of the four content category performance levels: *Below Basic*, *Basic*, *Proficient*, or *Advanced*. The SPI cut scores separating these performance levels are derived as expected percentages of possible score points for a given standard (content category) for students whose total test score is at the corresponding total test cut score (*Basic*, *Proficient*, or *Advanced*).

Use of the Standard-Level Subscores

The purpose of reporting SPI scores on the Wisconsin Forward Exam is to show the relationship between the overall achievement being measured (represented by the test score) and the skills within each of the content standards associated with the content area. Teachers may use the SPI scores for individual students as indicators of strengths and needs, but the SPI scores are best corroborated by other evidence, such as homework, class participation, diagnostic test scores, or observation.

District and school administrators may compare their results by content standard and grade level with the state results to better understand students' strengths and needs within a particular content area and grade level. Caution should be exercised when comparing standard-level subscores across years because different items will contribute to these subscores and these items may vary in difficulty between test forms or test administrations.

1.3 Technical Report Structure

A typical Technical Report documents major activities of the testing cycle. This abbreviated Technical Report provides comprehensive details only on the process of and activities related to the test form development in preparation for the Spring 2020 test administration. An overview of the parts included in the *Wisconsin Forward Exam Spring 2020 Technical Report* and a short description of parts not included in this report is provided below.

Part 2 of the Technical Report discusses the concept of validity evidence. This Technical Report provides content-based evidence that supports the validity of the Wisconsin Forward Exam ELA, Mathematics, Science, and Social Studies test forms.

Part 3 of this report describes the test blueprint, test design, item development process, test form development process, and some aspects of the content-related validity of the Wisconsin Forward Exam. More specifically, it describes how DRC and DPI collaborated to ensure that the appropriate content was included in the Wisconsin Forward Exam and to ensure that the test items adequately sampled the domain of content knowledge necessary to make legitimate inferences about student performance. The Wisconsin Academic Standards were the basis of the test blueprints and item specifications for their respective content areas. Wisconsin educators were involved in reviewing the items in all content areas to ensure the appropriateness of the test to the standards. The first item review, for grades 3–8 in ELA and Mathematics and grades 4, 8, and 10 in Social Studies, occurred in December 2015. The first item review for new assessments in Science grades 4 and 8 occurred in August 2017. Each year after that, new items were reviewed and added to the Wisconsin pool of items for future field-testing. The item reviews served to establish the accessibility of the items and reading passages. Simultaneously, DRC created the test specifications documents that were later approved by DPI and will continue to serve as a foundation for item and test development. Additional item reviews, supported by the item data, occurred after each field test administration and were conducted by DPI content experts. The purpose of these reviews was to refine the pool of items from which the subsequent operational test forms would be selected.

Part 3 also presents the Wisconsin Forward Exam design and discusses key development tasks related to creating the Wisconsin Forward Exam forms intended for Spring 2020 test administration. Item selection was based on the approved test blueprints. DRC’s CCR item bank contained a sufficient number of items to fulfill the test design needs for the ELA, Mathematics, and new Science assessments. Social Studies test forms consisted of Wisconsin-owned items. Part 3 also discusses the process of selecting operational test items and the process of obtaining DPI approvals. As detailed in Part 3, in addition to the operational test items, there were numerous unique field test items on each form. Selection of the test forms intended for the Spring 2020 test administration was done using the approved test blueprints, test designs, and psychometric specifications as guides.

Although parts describing test administration, scoring, psychometric data analysis and standard setting, studies of reliability and validity, and assessment reporting and results, are not included in this Technical Report, a brief description of the typical content of these parts is provided below.

Part 4 of the Technical Report was intended to describe the processes and activities implemented and information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students.

Part 5 was intended to document the scoring process for different item types: scanning of multiple-choice and multi-select items; autoscoring of technology-enhanced, short-answer, and evidence-based selected response items; and artificial intelligence scoring and handscoring of text-dependent analysis items.

Part 6 was intended to describe characteristics of the sample of student data used for data analysis and to present the classical and item response theory model (IRT) procedures implemented to analyze the Wisconsin Forward Exam test data.

Part 7 was intended to provide a brief overview of the standard setting process, during which the performance level cut scores were set for the ELA, Mathematics, and Social Studies tests in Summer 2016 and for Science tests in Spring 2019.

Part 8 was intended to include detailed results of Wisconsin Forward Exam reliability studies related to the test's internal consistency, student performance level classification consistency and accuracy, and inter-rater reliability for TDA items on the ELA assessments.

Additional construct-related validity evidence supporting the Wisconsin Forward Exam, including differential item functioning, principal component analysis, correlations among content standards, and a relationship between the Wisconsin Forward Exam scores and external variables, were planned to be presented in Part 9.

Part 10 was intended to include short descriptions of reports provided to end users, including individual student reports and aggregate reports, as well as the test results of the Spring 2020 Wisconsin Forward Exam administration.

While key findings of the Wisconsin Forward Exam administration cycle are presented in the body of the report, recommendations for subsequent administrations are typically presented in Part 11, which is the last part of the Technical Report. Because this Technical Report does not include content related to test administration, scoring, data analysis and standard setting, studies of test reliability and validity, or assessment reporting and results, the recommendations for the next test administration are presented in Part 4.

Part 2: Validity Framework

Validity is the overarching component of the Wisconsin Forward Exam ELA, Mathematics, Science, and Social Studies assessments. The following excerpt is from the *Standards for Educational and Psychological Testing* (hereafter the *Standards*) (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014):

> Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence . . . include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question. (p. 22)

As stated by the *Standards,* the validity of a testing program hinges on the interpretation of the test scores. In this part of the Technical Report, the sources of the validity evidence for any standardized educational assessment are discussed first, followed by a summary of the validity evidence based on the content of the Wisconsin Forward Exam test forms developed for the Spring 2020 administration.

2.1 Sources of Validity Evidence

The sources of validity evidence described in the *Standards* (AERA et al. 2014, pp. 26–31) include evidence based on test content, evidence based on response processes, evidence based on internal test structure, evidence based on relationship with other variables, and evidence based on consequences of testing. These sources of validity evidence are briefly described below.

Validity evidence based on test content can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure (AERA et al. 2014, p.14). It refers to traditional forms of content validity evidence and is supported by a correspondence between test content and a specification of the content domain. This type of evidence can be demonstrated through consistent adherence to test blueprints, through a high-quality test development process that includes the review of items for accessibility to English language learners and students using testing accommodations, and through alignment studies.

Validity evidence based on response process relies to large degree on the evaluation of the cognitive processes of examinees responding to various types of items and the relationship between these processes and the construct being measured. Direct evidence based on response processes typically comes from analyses of individual responses or responses from test takers from various groups making up the intended test-taking population about their performance strategies or responses to specific items (AERA et al. 2014, p.15). Such evidence can be gathered through cognitive labs conducted as part of the field test data analysis. Validity evidence based on response process is also supported by a relationship between the item type, format, and content and the construct being measured. For example, if a test is intended to measure a certain set of skills or knowledge, it is important to determine whether the items included in the test are, in fact, designed to measure these skills. In addition, evaluation of student written responses (e.g., text-dependent analysis) further contributes to the validity evidence based on response processes. In such cases, validity evidence includes the extent to which the processes of item response scoring, whether by a human reader or by an artificial intelligence engine, are consistent with the intended interpretation of scores. For example, scorers are expected to apply particular criteria in scoring students' responses and not be influenced by factors that are irrelevant to the intended interpretation of the scores (AERA et al., 2014, pp. 15–16). Recruitment and training of human scorers, as well as monitoring the artificial intelligence scoring process and results, contribute to the validity evidence based on response processes.

Validity evidence based on internal test structure refers to the fact that "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16). Such analyses may include statistical analyses of items and subscores conducted to investigate the dimensionality of an assessment. Procedures for gathering such evidence may include factor analysis for single assessments and evaluation of the continuity of the construct across grades for vertically scaled assessments. Internal test structure can also be evaluated using indices of measurement precision such as test reliability, decision accuracy and consistency, generalizability coefficients, and standard errors of measurement. Evaluation of the correlation coefficients that measure the relationship between the content standard (domain) scores and studies of whether test items may function differently for different subgroups of students are additional sources of validity evidence based on internal test structure.

Validity evidence based on relationships to other variables refers to "evidence about the degree to which these relationships are consistent with the construct underlying the proposed test score interpretations" (AERA et al., 2014, p. 16). In educational testing, such evidence is often gathered through studies of correlations between the test scores and measures of different or similar constructs. As stated in the *Standards* (AERA et al., 2014, pp. 16–17), relationships between test scores and other measures intended to assess the same or similar constructs provide convergent evidence, whereas relationships between test scores and measures of different constructs provide discriminant evidence.

Validity evidence based on the consequences of testing is ultimately determined by the stakeholders. Stakeholders decide the purpose and interpretation of scores within their system of reporting and accountability. DRC provides information about test content and technical quality but does not decide the use of test scores. As such, the validity evidence based on consequences of testing has not been addressed in the Wisconsin Forward Exam Technical Reports published to date.

2.2 Summary of Validity Evidence for Wisconsin Forward Exam Based on Test Content

Part 3 of the Technical Report documents evidence of the content-related validity demonstrated through each Wisconsin Forward Exam assessment’s consistent adherence to the assessment blueprints, which were constructed by DPI based on the Wisconsin Academic Standards. This part of the report also presents the test design and describes the key development tasks related to Wisconsin ELA, Mathematics, Science, and Social Studies operational test forms intended for the Spring 2020 test administration. This part documents the involvement of Wisconsin educators, DPI, and DRC in the item review and test development process. The test development process and the involvement of Wisconsin educators in that process forms an important part of the validity of the entire Wisconsin Forward Exam program. The knowledge, expertise, and professional judgment offered by Wisconsin educators ultimately ensure that the content of the Wisconsin Forward Exam forms an adequate and representative sample of appropriate content and that the content forms a legitimate basis from which to derive valid conclusions about student achievement. The blueprint and design, as well as the item and test development activities, described in Part 3 explain how specific development processes provide evidence in support of the validity of an intended interpretation of test scores, primarily based on the test content and through the use of expert professional judgment from Wisconsin educators and from DRC test development specialists. The foundational documents—test blueprints and test designs—developed and approved during the initial phases of test development served as critical guides throughout the development and field-testing of items. These documents contributed to ensuring that each form of the test accurately measured the content in consistent and stable ways, thus providing evidence supporting the test scores’ use as indicators of student achievement of Wisconsin standards.

Part 3 provides evidence to support the validity of an intended interpretation of test scores based on test content of the Wisconsin Forward Exam and addresses AERA, APA, & NCME (2014) Standards 3.1, 3.2, 3.9, 4.0, 4.1, 4.7, and 4.12.

Part 3: Test Content and Test Development

The purpose of this section is to describe how DRC, DPI, and Wisconsin educators collaborated through a series of test development processes to ensure that appropriate content was included in the Wisconsin Forward Exam and to ensure that test items adequately sampled the domain of content knowledge necessary to make accurate inferences about student performance. Part 3 documents the test blueprints, test designs, item development process, review and field-testing of new items, and the test form development process for the Spring 2020 administration.

This part of the Technical Report is particularly relevant to American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014) Standards 3.1, 3.2, 3.9, 4.0, 4.1, 4.7, and 4.12. Each of these Standards and the way each Standard is addressed will be presented in this section of the report. AERA, APA, & NCME (2014) Standard 4.0 states the following:

> Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population. (p. 85)

The test blueprint and item development activities described in this part explain how specific development processes provided evidence to support test validity, primarily content validity, through the use of expert professional judgment from Wisconsin DPI and from DRC test development specialists. The foundational documents—test blueprints and test designs—developed and approved during the initial phases of the project served as critical guides throughout the development of the test forms. These documents contributed to ensuring that each test form accurately measured the content in consistent and stable ways, thus providing evidence supporting the test's use as an indicator of student achievement of Wisconsin standards.

The Wisconsin Forward Exam ELA, Mathematics, Science, and Social Studies domains are generally defined as the knowledge and skills that are identified within the Wisconsin Academic Standards for these content areas. The framework of Wisconsin Academic Standards, in turn, is based on prior consensus among DPI, Wisconsin educators, and experienced subject matter experts that the standards represent what is important for teachers to teach and students to learn.

Evidence of validity based on test content includes information about the test specifications, including the test design and test blueprint. Test development involves creating a design framework from the statement of the construct to be measured. The primary consideration in the development of the Wisconsin Forward Exam test specifications was the assessment's alignment with the Wisconsin Academic Standards. The constraints of the assessment program and state policy decisions were also taken into consideration in the development of the test specifications.

The Wisconsin Forward Exam test specifications consist of a test blueprint and a test design for each grade level and content area. In partnership with DRC, DPI created test blueprints and test designs. DRC and DPI content experts scrutinized each blueprint to ensure optimal content coverage and efficient use of time and resources.

3.1 Test Blueprints

AERA, APA, & NCME (2014) Standard 4.1 states the following:

> Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s). (p. 85)

The key structural aspect of the Wisconsin Forward Exam for ELA, Mathematics, Science, and Social Studies is the assessment blueprint that specifies the target score points for each grade and content strand or domain. These assessment blueprints were developed by staff members at Wisconsin DPI who made recommendations for the test content for each grade and content area, seeking to ensure optimal content coverage of the Wisconsin Forward Exam assessments. In general, each blueprint represents content sampling proportions that reflect the intended emphasis on instruction and mastery in each content area and grade level. Specifications for a range of items by standard and item type demonstrated the desired proportions within the summative assessment. In summary, the Wisconsin Forward Exam assessment blueprint at a given grade and content area provides guidance on how the standards are measured.

The test blueprints specify the number of item points for each reporting category and subskill as well as the allowable depth-of-knowledge (DOK) levels for the respective reporting categories. The process used for developing the blueprints for the Wisconsin Forward Exam was a collaborative effort between DRC and DPI. The DPI-approved blueprints can be found in Tables 3-1 through 3-4 for ELA, Mathematics, Science, and Social Studies, respectively.

3.2 Test Design

The test design for the 2020 operational assessments included the use of items reviewed and approved by Wisconsin educators and DPI. Information concerning the item development process can be found in Section 3.3. Various item types were developed and included in the Wisconsin Forward Exam in order to assess students' understandings of the standards. A description of item types included in the Wisconsin Forward Exam is presented in Table 3-5. The following sections provide detailed information about the test design of the content areas assessed in the Spring 2020 Wisconsin Forward Exam assessments.

3.2.1 English Language Arts

Table 3-6 shows the ELA test design, including the number of passages, items, and points at each grade level that were used in the core and embedded field test positions. There was one common set of core operational items in each of the eight field test forms at each grade level. Table 3-6 also identifies the various item types that appeared on the ELA forms, including the points for item scoring. Detailed descriptions of the item types are provided in Table 3-5 of this report.

The ELA section of the Forward Exam was divided into four sessions: text-dependent writing prompt, writing/language, listening, and reading. Students were able to take the sessions in any order. Recommended testing times for all sessions were included in the test design document as well as in the test administration manual.

3.2.2 Mathematics

Table 3-7 shows the Mathematics test design, including the number of items and points at each grade level that were used in the core and embedded field test positions. There was one common set of core operational items in each of the eight field test forms at each grade level.

The Mathematics section of the exam was divided into two testing sessions, with students able to take the sessions in either order. In grades 3–5, no calculator was allowed for any of the Mathematics items. In grades 6–8, no calculator was allowed for the first session, and the second session allowed students to use an embedded calculator. Recommended testing times for both sessions were included in the test design document as well as in the test administration manual.

3.2.3 Science

Table 3-8 shows the Science test design, including the number of items and points at each grade level that were used in the core and embedded field test positions. There was one common set of core operational items in each of the twenty field test forms at each grade level.

The Science section of the exam was divided into three testing sessions, with students able to take the sessions in any order. Recommended testing times for all sessions were included in the test design document as well as in the test administration manual.

3.2.4 Social Studies

Table 3-9 shows the Social Studies test design, including the number of items and points at each grade level that were used in the core and embedded field test positions. There was one common set of core operational items in each of the fifteen field test forms at grade 4 and thirteen field test forms at both grades 8 and 10. The Social Studies exam included two test sessions that could be administered in either order. Recommended testing times for both sessions were included in the test design document as well as in the test administration manual. The Social Studies exams included custom items developed specifically for the Wisconsin Forward Exam.

3.3 Universal Design

Assessments that are universally designed allow for the participation of the widest possible range of students, resulting in more valid inferences about student performance. Universally designed grade-level assessments may reduce the need for accommodations by reducing or eliminating access barriers associated with the tests themselves. Table 3-10 presents the elements of universal design that were implemented on the Wisconsin Forward Exam (Thompson, Johnstone, & Thurlow, 2002).

These elements of universal design are relevant to both item development and form construction. This section addresses how the elements of universal design were addressed in the construction of the Spring 2020 test forms in compliance with AERA, APA, & NCME (2014) Standard 3.1, which states the following:

> Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (p. 63)

A goal of universal design is to measure the performance of students with a wide range of abilities and skills, ensuring that students with diverse learning needs receive opportunities to demonstrate competence on the same content. To accommodate the greatest number of students for the Wisconsin Forward Exam, the assessments include simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility. These design components are addressed primarily through the physical layout and formatting of the online test forms as well as the paper-based test forms used for accommodations. The page specifications define how directions and test items are placed on the pages, the location and appearance of headers and footers, the spacing between an item stem and the answer choices, and other page elements to ensure a consistent, legible appearance of online forms and paper-based test forms. Written instructions at the beginning of each test session are clearly and simply stated, and the wording of such instructions is standardized as much as possible across content areas and grade levels to ensure clarity and consistency.

3.3.1 Accommodations

AERA, APA, & NCME (2014) Standard 3.9 states the following:

> Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs. (p. 67)

Students with disabilities or students who are English language learners may be provided with test administration accommodations based on their Individualized Education Plans. Accommodation code definitions can be found in the Accessibility Guide available on the DPI website at the following address: https://dpi.wi.gov/assessment/forward/accommodations.

Braille and Large Print test versions were also available for each grade and content area to enable students who are blind or visually impaired to participate in the Wisconsin Forward Exam. Braille and Large Print forms for all grades and content areas were created by DRC test developers and consisted of the same content that was included in the regular operational online test forms. Specific recommendations on how to transcribe items into Braille were provided by an independent Braille expert who collaborated with the Braille publisher to produce the Braille version of the Wisconsin Forward Exam assessment and teacher's notes that accompany the Braille forms.

3.4 Item Development Process

ELA, Mathematics, and Science test items included in the Spring 2020 Wisconsin Forward Exam were selected from DRC's College- and Career-Ready (CCR) item bank. DRC's CCR item bank contains nationally field-tested CCR items that support the next generation of standards and assessments. CCR items are aligned to the College and Career Readiness standards in ELA and Mathematics grades 3–8. Science items are aligned to Wisconsin's Standards for Science and enhanced by the Next Generation Science Standards based on the National Research Council's Framework for K–12 Science Education. The item bank is designed to support states like Wisconsin that have adopted, or are preparing to adopt, more rigorous content standards, curricula, and assessments that better prepare students for college and careers.

Alignment to standards, grade-level appropriateness, depth of knowledge (DOK), item/task level of complexity, estimated difficulty level, relevancy of context, rationale for distractors, style, accuracy, and correct terminology were major considerations in the item development process. DRC's item development process for the CCR item bank followed the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). DRC's item development work was and continues to be designed to produce reliable and instructionally valid tests that reflect the complete range of performance articulated in the AERA, APA, and NCME *Standards*.

Furthermore, DRC's item development work adheres to the Principles of Universal Design (Thompson, Johnstone, & Thurlow, 2002) and reflects how items and tests must lend themselves to accessibility by diverse groups of students. Members of DRC's item development team have received direct training from the National Center on Educational Outcomes. Therefore, DRC employs the Principles of Universal Design throughout all stages of both the item development process and the test development process.

All DRC's ELA, Mathematics, and Science items that appear on the Wisconsin Forward Exam were reviewed for content and for fairness not only by DRC's content experts but also by a panel of external experts and, more recently, by Wisconsin educators. The external reviewers have a broad range of experience in the educational field. All the reviewers have bachelor's-level, master's-level, or doctoral-level degrees and teaching experience in their specific area of expertise. Table 3-11 provides a high-level sequence of the activities that occurred in the development of the DRC CCR item bank.

Wisconsin-owned Social Studies items were developed by DRC content specialists. The items appearing in operational positions were aligned to Wisconsin's Model Academic Standards for Social Studies and were repeated from the Spring 2018 test administration. Items appearing in the field test positions on the Spring 2020 forms were aligned to the Wisconsin Social Studies Standards (2018). Social Studies items underwent reviews by DRC content experts as well as by DRC bias and sensitivity experts. All Social Studies items were also reviewed and approved by committees of Wisconsin educators.

The efforts by DRC in developing items are in alignment with multiple best practices of the test industry and, in particular, support the following AERA, APA, & NCME (2014) Standards:

> **Standard 3.1** Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (p. 63)
>
> **Standard 3.2** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (p. 64)

As stated earlier, Wisconsin licensed ELA, Mathematics, and Science items from DRC's CCR item bank for the Spring 2020 test administration. Due to the state-specific nature of the Social Studies standards, DPI owns the items for that content area. Details regarding the development of the items in the CCR bank created prior to their field-testing on the Forward Exam are provided in the *Wisconsin Forward Exam Spring 2016 Technical Report*, available on the DPI website at https://dpi.wi.gov/assessment/forward/resources#documentation.

3.4.1 Reading Passage and Item Reviews—Summer 2018

Test items typically begin their life cycle two years prior to their operational administration. New ELA, Mathematics, Science, and Social Studies passages and items were first reviewed and approved for placement on the Wisconsin Forward Exam by both DPI and Wisconsin educators. For these reviews, educators from across the state convened in Madison, Wisconsin, to review items in an online format so that items could be evaluated in the same testing engine and style in which items are presented to students during the actual administration. ELA, Mathematics, and Social Studies item reviews were held the week of August 6, 2018. The review of Science items occurred the week of August 13, 2018. An example of the training PowerPoint presentation used at the reviews can be found in Appendix A of this report.

Table 3-12 shows the number of items taken to the item review by grade and content area. Using the approved test blueprints as a guide, DRC content specialists determined the focus of the items that would be taken to item review. Using an electronic tally sheet, Wisconsin educators made determinations about standard alignment, depth-of-knowledge levels, and key(s). They noted any bias and sensitivity concerns and had the opportunity to determine whether items were accepted as is or accepted with revisions. They also had the opportunity to register a “dissenting view,” which indicated that the committee preferred the item not be selected to appear on the Wisconsin Forward Exam in a field test position.

Items and passages that were approved by the Wisconsin educators were then included in the next field test administration in Spring 2019. The purpose of the Spring 2019 field test was to expand the pool of items eligible for inclusion in subsequent operational test forms, such as the Spring 2020 Forward Exam.

3.5 Field-Testing—Spring 2019

Items approved for the field test administration during the Summer 2018 item review were field-tested in Spring 2019 during the operational test administration. Field test items were fully embedded in the operational forms, and students were not able to distinguish between the operational and field test items. The field test items were embedded in several test forms administered in each grade and content area. Each test form contained the same operational test items and unique field test items. The test forms were spiraled at the student level within a grade and a content area. A total of 401 items were field-tested for ELA. A total of 188 items were field-tested for Mathematics. A total of 199 items were field-tested for Science, and a total of 524 items were field-tested for Social Studies in the Spring 2019 test administration.

3.5.1 Statistical Analysis of Field Test Data

Following the field test data acquisition, the field test data analyses were conducted. The analyses included classical item analysis, differential item functioning (DIF) analysis, and item response theory (IRT). The classical item analysis included the computation and evaluation of the following statistics: item *p*-values (difficulty), item-total test correlation, percentage of students selecting incorrect responses, point-biserial correlation for incorrect responses for the multiple-choice (MC) items, score point distribution for items worth more than 1 point, and omit rates for all items. Details on classical item analysis methodology can be found in the *Wisconsin Forward Exam Spring 2019 Technical Report*, Part 8, posted at the following address: https://dpi.wi.gov/assessment/forward/resources#documentation.

DIF was conducted for all field test items to examine potential item bias and to determine whether item performance differences between identifiable subgroups were due to factors other than student ability, making the items unfairly difficult for a particular subgroup in the student population. DIF analyses were conducted based on gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency status. More details on the DIF methodology can be found in the *Wisconsin Forward Exam Spring 2019 Technical Report*, Part 10, posted at the following address: https://dpi.wi.gov/assessment/forward/resources#documentation.

As the last step of the field test data analyses, the field test items were calibrated and equated to operational test scales using the IRT methodology (explained in detail in the *Wisconsin Forward Exam Spring 2019 Technical Report*, Part 6). Note that ELA, Mathematics, Science, and Social Studies field test items were equated to their respective operational test scales using a common student design. All operational test items contained in the Spring 2019 operational test forms served as anchor items to place the field test items on the operational test scores using the Stocking and Lord procedure.

The field test item statistics are used as a means of detecting items that deserve closer scrutiny, rather than being a mechanism for automatic retention or rejection. To this end, a set of criteria was used as a screening tool to identify items that needed a closer review. For an item to be flagged for an additional review, the criteria included any of the following:

- *p*-value <0.20 or >0.90
- item-total test correlation (point biserial for MC items) <0.15
- positive point biserial on a distractor for an MC item
- omit rate >3%
- large DIF

Items flagged for any of the above reasons were reviewed by the content area specialists prior to their review by DPI.

3.5.2 Item Data Review—Summer 2019

In the preceding section, it was stated that test development content area specialists used certain statistics from item and DIF analyses of the 2019 field tests to identify items for further review. Specific flagging criteria for this purpose were specified in the previous section. In addition to items flagged for poor statistics, several items with statistics just above the threshold or potential content-related concerns were also reviewed at the data review meeting. Items of extremely poor statistical quality were regarded as unacceptable and needed no further review. Such items were excluded from the Wisconsin item pool prior to the data review with DPI. The intent was to capture all items that needed an additional review based on their statistical properties or item content; thus, the criteria employed for identification of field test items needing an additional review tended to over-identify rather than under-identify potential item issues.

The data review of the items was conducted by DPI staff and DRC content specialists, who were broken out into content area and/or grade-level groups. The data review took place in Madison, Wisconsin, in July 2019 (ELA and Science) and August 2019 (Mathematics and Social Studies). In these sessions, reviewers were first trained by a representative from DRC's staff with regard to the statistical indices used in item evaluation. This was followed by a discussion with examples concerning reasons that an item might be retained regardless of the statistics. The review process involved a brief exploration of possible reasons for the statistical profile of an item (e.g., possible bias, grade appropriateness, instructional issues) and a decision regarding acceptance. DRC content area test development specialists facilitated the review of the items. Each group reviewed the pool of field test items and made recommendations on each item and/or scenario/passage. The training presentation used at the ELA and Science data review meeting may be found in Appendix B. (Similar training was conducted for Mathematics and Social Studies.) A summary of the data review results, including the number of items that were field-tested, the number and percentage of items with statistical flags, and the number and percentage of items rejected by DPI during the data review, is presented in Table 3-13. Items accepted for subsequent use in the Wisconsin Forward Exam were included in the pool of items for Spring 2020 operational test form selection.

3.6 Form Development Process

The creation of test forms involved the expertise of multiple DRC departments and DPI. The activities that contributed to the creation of the test forms are described below.
The Wisconsin Forward Exam test development process complied with the following AERA, APA, & NCME (2014) *Standards*:

> **Standard 4.1** Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s). (p. 85)
>
> **Standard 4.7** The procedures used to develop, review, and try out items and to select items from the item pool should be documented. (p. 87)
>
> **Standard 4.12** Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications. (p. 89)

The DRC team worked cooperatively with DPI content and assessment specialists to select passages and prompts with associated content-specific items for the online assessments. The DRC team constructed forms that complied with the approved test blueprints and form construction guidelines. DRC used an integrated team approach to test development, which included content area specialists, psychometricians, and scoring specialists working as a unit in collaboration with DPI content experts.

3.6.1 Item Selection

New operational test forms were developed for ELA, Mathematics, and Science for the Spring 2020 test administration. As a first step in building the online assessments, the DRC team prepared all eligible items in DRC’s item banking system, which is called IDEAS. The form, format, extent, and organization of items in their respective test sessions were determined in consultation with DPI.

Following the preparation of all necessary materials and resources, forms construction began. The construction of the test forms themselves was a collaborative effort within DRC’s integrated development team of assessment specialists, psychometric services specialists, and scoring specialists.

Before test forms were created, passages, item/performance tasks, and artwork were carefully selected. The following process was used for item selection:

- Using the pool of vendor-owned items for ELA, Mathematics, and Science and Wisconsin-owned items for Social Studies, DRC test development specialists first selected items to match the approved test blueprints.

- DRC test development specialists checked to see that each item clearly aligned with the standards where applicable and that each item with available item statistics met psychometric guidelines for inclusion in the test.
- DRC test development specialists verified that each item met technical quality requirements for well-crafted items, including that each item
  - had one clearly correct answer (or answers if the item was multi-select);
  - used clear and concise wording;
  - was grammatically correct;
  - had an appropriate range of difficulty;
  - was free of any offensive, inappropriate, or biased content; and
  - met the Principles of Universal Design and maximum accessibility.

In addition to content requirements, the following statistical criteria were used in item selection:

- Test length and item types match the DPI-approved test design.
- Content coverage matches the DPI-approved test blueprint.
- Items had acceptable statistics which included:
  - *p*-value between 0.20 and 0.90
  - Item-total test correlation >0.15
  - Omit rates <3%
  - Acceptable fit statistics (no misfit flag)
  - No large DIF—If an item with large DIF had to be included in the test to maintain blueprint coverage, the item was examined to determine whether any content reason exists for the DIF flag (sometimes items demonstrate statistical bias but no content reason can be determined for the bias).

The statistical properties of the Spring 2019 test forms were used as targets for selection of the Spring 2020 ELA, Mathematics, and Science test forms. The item selection for these content areas was conducted in two phases.

In the first phase, the anchor (linking) items were selected. The anchor items are used for the statistical linking of the new forms to previous test forms on already established test scales. The anchor items on the Spring 2020 test forms were selected mainly from the Spring 2019 operational item pool (with a small number of anchor items being selected from the Spring 2018 operational tests for ELA and Mathematics). The anchor set was selected as a “mini” version of the full operational test for each grade level and content area in regard to its length, content coverage, and psychometric properties.

The length of the anchor sets was at least one-third of the length of the total test. The items included in the anchor sets met the same blueprint specifications as the full test in regard to the percentage of score points measuring each content standard. In addition, the psychometric properties of the anchor sets matched the corresponding properties of the target forms as closely as possible. Anchor selections were reviewed and approved by a DRC psychometrician.

In the second phase of the item selection process, non-anchor operational items were selected for ELA, Mathematics and Science. With the exception of ELA TDA items, the non-anchor operational items came from the Spring 2017, 2018, and 2019 Wisconsin Forward Exam operational and field test item pool for ELA and Mathematics and from the Spring 2018 and 2019 Wisconsin Forward Exam field test and operational item pool for Science. TDA items included in the Spring 2020 ELA assessments were not previously field-tested in Wisconsin.

The non-anchor operational items were selected using the item selection guidelines presented earlier in this section. Full form selections were reviewed and approved by a DRC psychometrician.

After the selection of all operational items, the new field test items were added to each form in each grade and content area. In constructing the final forms, the DRC content area test development specialists followed the guidelines provided below:

- Forms included adequate standards coverage as required by test blueprints.
- No item in a form "clued" another item on that same form.
- Forms were diverse in terms of artwork and graphics.
- Forms included a wide range of topics and a variety of questions.
- Correct answer distributions were reasonable across MC items on the form.
- Forms did not contain any items that had been released to the public.
- DPI reviewed and gave final approval of all online test forms.

No item selection was conducted for Social Studies assessments. Instead, the operational portions of the assessments administered in Spring 2018 were intended to be reused in the Spring 2020 test administration. Similar to other content areas, new field test items were embedded in each test form in each grade of Social Studies.

3.7 Item and Form Quality Reviews

In all phases of the item and form development process, content area test development specialists and editorial specialists reviewed items and passages for technical quality; alignment with the standards; issues of bias, fairness, and sensitivity; depth of knowledge; estimated difficulty; and adherence to the Principles of Universal Design. The aim for this team approach was to conduct a multi-tiered internal review of all passages and items prior to submission for review by DPI and then, with approval from DPI prior to submission, for review by Wisconsin educators to ensure that all items align with Wisconsin's standards and adhere to DPI's standards for high-quality items.

DRC content and editorial teams reviewed all passages and items to ensure that they possessed the following characteristics:

- content alignment or congruence with the knowledge and skills specified in the standards;
- a range of estimated difficulty levels;
- appropriate grade-level vocabulary, subject matter, and assumed student knowledge;

- freedom from issues or concerns regarding bias, sensitivity, or fairness;
- accessibility, following the Principles of Universal Design; and
- correct grammar, usage, and structure/format.

As part of DRC's internal review of the items and test forms, the test development team members and graphic specialists ensured that item art could be reproduced clearly and accurately when electronically displayed and when used in the print-on-demand forms.

Test specifications were reviewed to identify any potential display requirements that may present challenges in an electronic display environment. Display tolerances are impacted by line thickness, percentage of screening for shading, specialized fonts and symbols, photographs, and color. These are defined in the early stages of the item and test development process to help guide the delineation of style requirements and specifications.

Item art was produced using transparent vector graphics that allow for adjustments without the breakdown of image clarity, which is common with lower-quality formats, and provide for the online accommodation of alternate background colors. The DRC multi-tiered quality assurance process made certain that converted item art was carefully compared to the original format throughout the test development and production process.

In reviewing forms in the online environment, multiple reviewers checked passages and items on the multiple electronic platforms on which students took the test to ensure a smooth testing experience.

3.7.1 DPI Approvals

DPI had the opportunity to review passages and items to be placed on the Spring 2020 Wisconsin Forward Exam during the following phases:

- prior to item content review in Summer 2018
- at item content review in Summer 2018
- during review of flagged field test data in Summer 2019
- during the Spring 2020 form construction

Prior to the opening of the testing window, all online forms were made accessible to DPI for review in DRC's secure INSIGHT testing engine.

3.8 Summary

In summary, the Spring 2020 Wisconsin Forward Exam test forms adhered to the Wisconsin test blueprints and test designs for each grade level and content area. The items included in the Spring 2020 Wisconsin Forward Exam were reviewed by DRC, DPI, and Wisconsin educators for issues regarding accessibility, bias, sensitivity, and content. During the reviews, experts identified (1) issues that could negatively affect a student's ability to access stimuli and items, (2) content in stimuli and items that could unfairly affect a student's response because of his or her background, (3) developmental appropriateness, and (4) the alignment of stimuli and items to the content specifications. Item content was checked for the accuracy of the content, answer keys, and scoring rules. Following Spring 2019 field-testing, items flagged for accessibility, bias and sensitivity, and/or other content concerns were further reviewed by DRC and DPI to determine whether these flagged items should be removed from the Wisconsin item pool prior to the form construction of the Wisconsin Forward Exam. In addition, item statistics from the Spring 2019 operational and field test administration were used to refine the item pool used in the selection of Spring 2020 Wisconsin Forward Exam forms. The efforts and procedures used in the development of the Spring 2020 Wisconsin Forward Exam forms balanced the content and psychometric requirements for form development. The content of the Spring 2020 test forms adhered to the test blueprint requirements. The psychometric properties of the new test forms were comparable to the psychometric properties of the Spring 2019 forms for all content areas. Overall, the process implemented in the Spring 2020 operational form development was in alignment with multiple best practices of the test industry.

Table 3-1 English Language Arts Test Blueprints for Grades 3–8

Domain (Reporting Category)	**Depth of Knowledge**	**Total Points by Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Reading		22	24	24	24	24	24
Key Ideas and Details	grade 3: 1–3 grades 4–8: 2–3	6–12	6–12	6–12	6–12	6–12	6–12
Craft and Structure/Integration of Knowledge and Ideas	all grades: 2–3	4–10	4–10	4–10	4–10	4–10	4–10
Vocabulary Use—Includes Language Standards 4 and 5	grades 3–5: 1–3 grades 6–8: 2–3	4–6	4–6	4–6	4–6	4–6	4–6
Literature		about 60%	about 60%	about 60%	about 50%	about 50%	about 50%
Informational Text		about 40%	about 40%	about 40%	about 50%	about 50%	about 50%
Writing/Language		24	24	24	24	24	24
Text Types and Purposes/ Text-Dependent Analysis	all grades: 2–3	10–14	10–14	10–14	10–14	10–14	10–14
Research	all grades: 2–3	6–8	6–8	6–8	6–8	6–8	6–8
Language Conventions	all grades: 1–3	6–8	6–8	6–8	6–8	6–8	6–8
Listening	all grades: 2–3	7	8	8	8	8	8
ELA Points Total		**53**	**56**	**56**	**56**	**56**	**56**

Table 3-2 Mathematics Test Blueprints for Grades 3–8

Reporting Category	**Depth of Knowledge**	**Total Points by Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Operations and Algebraic Thinking	grade 3: 1–3 grades 4–5: 1–2	8–10	9–11	8–10			
Number and Operations in Base Ten	grades 3–5: 1–3	7–9	8–10	8–10			
Number and Operations—Fractions	grades 3–5: 1–3	7–9	9–11	8–10			
Measurement and Data	grades 3–5: 1–3	9–11	9–11	9–11			
Geometry	grades 3–4: 1–2 grades 5–8: 1–3	6–8	6–8	8–10	6–8	9–11	9–11
Ratios and Proportional Relationships	grades 6–7: 1–3				6–8	7–9	
The Number System	grades 6–7: 1–3 grade 8: 1–2				10–12	6–8	7–9
Expressions and Equations	grades 6, 8: 1–3 grade 7: 1–2				10–12	9–11	9–11
Statistics and Probability	grade 6: 1–2 grades 7–8: 1–3				9–11	10–12	7–9
Functions	grade 8: 1–3						9–11
Mathematics Points Total		**42**	**46**	**46**	**46**	**46**	**46**

Table 3-3 Science Test Blueprints for Grades 4 and 8

Reporting Category	**Depth of Knowledge**	**Total Points by Grade**	
		4	**8**
Practices and Crosscutting Concepts in Life Science	grades 4, 8: 2–3	8–12	8–12
Practices and Crosscutting Concepts in Physical Science	grades 4, 8: 2–3	8–12	8–12
Practices and Crosscutting Concepts in Earth and Space Science	grades 4, 8: 2–3	8–12	8–12
Practices and Crosscutting Concepts in Engineering	grades 4, 8: 2–3	8–12	8–12
Science Total Points		**40**	**40**

Table 3-4 Social Studies Test Blueprints for Grades 4, 8, and 10

Reporting Category	**Depth of Knowledge**	**Total Points by Grade**		
		4	**8**	**10**
Geography: People, Places, and Environments	all grades: 1–3	7–11	8–12	9–11
History: Time, Continuity, and Change	all grades: 1–3	6–10	10–15	11–14
Political Science and Citizenship: Power, Authority, Governance, and Responsibility	grade 4: 2–3 grades 8, 10: 1–3	5–9	5–7	11–14
Economics: Production, Distribution, Exchange, and Consumption	all grades: 1–3	5–9	5–7	7–10
The Behavioral Sciences: Individuals, Institutions, and Cultures	all grades: 2–3	5–9	4–6	7–10
Social Studies Total Points		**38**	**40**	**50**

Table 3-5 Item Type Descriptions for Items on the Wisconsin Forward Exam

Item Type	Name	Description
EBSR	Evidence-Based Selected Response	Each evidence-based selected response item has two parts, and each two-part item is designed to elicit an evidence-based response from a student who has read a literature text passage, an informational text passage, or a writing concept. In part one, which is similar to a multiple-choice item, the student analyzes a passage or writing concept and chooses the best answer from four response options. In part two, the student uses evidence from the passage or writing concept to select one or more answers based on the response to part one. Each of these items is worth one point.
MC	Multiple Choice	Each multiple-choice item has four response options, only one of which is correct. Multiple-choice items are used to assess a variety of skill levels, from short-term recall of information to inference and problem solving. Each of these items is worth one point.
MS	Multiple Select	Each multiple-select item requires a student to evaluate information presented and respond by choosing two or more correct responses. Multiple-select items can be used to assess multiple skills and concepts in a given content area.
SA	Short Answer	Each short-answer item requires a student to enter a short numeric or algebraic response. These items are designed to assess a student's ability to formulate a solution to a pure or applied mathematics problem without the assistance of response options. The short-answer items are scored on a 0–1-point scale using item-specific autoscoring rules.
TE	Technology Enhanced	Each technology-enhanced item is designed to elicit evidence of a broad range of student understanding. A student interacts with the enhanced features of these computer-delivered, auto-scorable test items to show understanding of skills and concepts. Item types such as drag-and-drop, hot-spot, number line and coordinate graphing, data displays, matching interaction, and drop-down menus are just some of the technology-enhanced item types presented to students. The technology-enhanced items are scored on a 0–2-point scale using item-specific scoring rules.
TDA	Text-Dependent Analysis	Each text-dependent analysis item is a text-based analysis based on a passage or a multiple-passage set that each student has read during the assessment. Both literary and informational texts are addressed through this item type. Students must draw on basic writing skills while inferring and synthesizing information from the passage in order to develop a comprehensive, holistic essay response. The demands required of a student's reading and writing skills in response to a TDA item coincide with the similar demands required for a student to be college and career ready. The TDA prompts are scored using a holistic scoring guideline on a 1–4-point scale. A weight of 2 is applied to the item scores in the computation of the student total test raw scores and scale scores. That is, the TDA prompts contribute up to 8 raw score points toward the student total test raw score. This item type is supported by all Wisconsin ELA standards across all grades for both Reading Literature and Reading Informational Texts and by Writing standards 1, 2, 3, 4, and 9 across all grades. The TDA items are scored using artificial intelligence (AI) scoring, with an appropriate level of human scoring to validate the AI algorithms for all TDA items used in the Wisconsin ELA grades 3–8 assessments.

Table 3-6 English Language Arts Test Design

Test Design		**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Number of Passage Sets	**Literature**	2	2	2	2	2	3
	Informational	2	3	4	3	3	2
	Listening	3	3	3	3	3	3
Number of Core (OP) Items	**Item Types: MC/TE (1 pt)**	27	28	30	24	22	28
	Item Types: MS/TE/EBSR (2 pts)	9	10	9	12	13	10
	Item Type: TDA (4 pts x 2)	1	1	1	1	1	1
	Total Core Items	37	39	40	37	36	39
Total Core Points		53	56	56	56	56	56
Embedded Field Test (FT)	**Number of Forms**	8	8	8	8	8	8
	Passages (Reading + Listening)	2	2	2	2	2	2
	FT Items per Form	10	10	9	8	8	8
	Total Field Test Items	72	72	72	64	63	64
Total Items (Core + FT) per Form		47	49	49	45	44	47
Total Estimated Testing Time (minutes)		130	130	130	130	130	130

Note: TDA items are scored using a 1–4-point scoring rubric. A weight of 2 is applied to item scores in the computation of the student total test raw scores and scale scores.

Table 3-7 Mathematics Test Design

Test Design		**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Number of Core (OP) Items	**Item Types: MC/SA (1 pt)**	35	40	40	38	39	37
	Item Type: TE (1 pt)	7	6	6	8	7	9
	Total Core Items	42	46	46	46	46	46
Total Core Points		42	46	46	46	46	46
Embedded Field Test (FT)	**Number of Forms**	8	8	8	8	8	8
	FT Items per Form	8	8	8	8	8	8
	Total Field Test Items	64	64	64	64	64	64
Total Items (Core + FT) per Form		50	54	54	54	54	54
Total Estimated Testing Time (minutes)		90	90	90	105	105	115

Table 3-8 Science Test Design

Test Design		Grade	
		4	**8**
Number of Core (OP) Items	**Item Types: MC/MS/TE/EBSR (1 pt)**	40	40
Total Core Points		40	40
Embedded Field Test (FT)	**Number of Forms**	20	20
	Scenarios/Tasks	10	10
	FT Items per Form	5	5
	Total Field Test Items	94	93
Total Items (Core + FT) per Form		45	45
Total Estimated Testing Time (minutes)		120	120

Table 3-9 Social Studies Test Design

Test Design		Grade		
		4	**8**	**10**
Number of Core (OP) Items	**Item Types: MC/TE/MS (1 pt)**	38	40	50
Total Core Points		38	40	50
Embedded Field Test (FT)	**Number of Forms**	15	13	13
	FT Items per Form	8	8	8
	Total Field Test Items	120	104	103
Total Items (Core + FT) per Form		46	48	58
Total Estimated Testing Time (minutes)		70	70	70

Table 3-10 Elements of Universal Design

Element	Explanation
Inclusive Assessment Population	Tests designed for state, district, or school accountability must include every student except those in the alternate assessment, and this is reflected in assessment design and field-testing procedures.
Precisely Defined Constructs	The specific constructs tested must be clearly defined so that all construct-irrelevant cognitive, sensory, emotional, and physical barriers can be removed.
Accessible, Unbiased Items	Accessibility is built into items from the beginning, and bias review procedures ensure that quality is retained in all items.
Amenable to Accommodations	The test design facilitates the use of needed accommodations.
Simple, Clear, and Intuitive Instructions and Procedures	All instructions and procedures are simple, clear, and presented in understandable language.
Maximum Readability and Comprehensibility	Readability and plain language guidelines are followed (e.g., sentence length and number of difficult words are kept to a minimum) to produce readable and comprehensible text.
Maximum Legibility	Characteristics that ensure easy decipherability are applied to text, tables, figures, illustrations, and response formats.

Table 3-11 College- and Career-Ready Item Bank Development Activities

DRC College- and Career-Ready Item Bank Development Activities
Establish item/passage development specifications and style guides and prepare item writing training manuals.
Determine item development plans.
Train item writers and/or passage developers in the project requirements and specifications.
Develop passages and write items.
Review, edit, code, and track items and produce graphics.
Produce review forms for content and bias/fairness/sensitivity reviews by external reviewers.
Modify items based on external reviewers’ recommendations.
Review and approve field test–ready items and passages.
Develop field test forms and administer field tests.
Internally review field test item data.
Approve items to be included in the item bank.

Table 3-12 Items Reviewed at Summer 2018 Item Review

Grade	Number of Items			
	English Language Arts	Mathematics	Science	Social Studies
3	65	44		
4	65	46	153	49
5	65	47		
6	65	45		
7	65	42		
8	65	44	155	48
10				64
TOTAL	**390**	**268**	**308**	**161**

Table 3-13 Items Reviewed at Summer 2019 Data Review

Content Area	Grade	Number of Items in 2019 Field Test	Field Test Items Flagged for Poor Statistics or DIF		Field Test Items Rejected at Data Review for Statistical or Content-Related Reasons	
			Number of Items	Percentage of All Field Test Items	Number of Items	Percentage of All Field Test Items
English Language Arts	3	75	20	26.7	8	10.7
	4	75	22	29.3	7	9.3
	5	64	19	29.7	6	9.4
	6	64	18	28.1	5	7.8
	7	63	20	31.7	6	9.5
	8	60	14	23.3	5	8.3
Mathematics	3	31	8	25.8	7	22.6
	4	32	10	31.3	8	25.0
	5	32	7	21.9	6	18.8
	6	32	10	31.3	7	21.9
	7	32	15	46.9	16	50.0
	8	29	15	51.7	10	34.5
Science	4	99	25	25.3	23	23.2
	8	100	32	32.0	24	24.0
Social Studies	4	16	3	18.8	3	18.8
	8	16	4	25.0	4	25.0
	10	20	2	10.0	2	10.0

Part 4: Summary Recommendations

The last section of this report presents some recommendations for the Spring 2021 test administration for DPI consideration.

The 2021 Wisconsin Forward Exam administration will be the fifth administration of the assessment. In this fifth administration, the assessment results will be reported on the existing scales and students will be classified into the proficiency levels using the cut scores established in the most recent standard settings, allowing for longitudinal tracking of student performance in ELA, Mathematics, and Social Studies. Using the same scales and the same cut scores for Wisconsin assessments allows for monitoring student growth across administration years. New test scales were established, and new performance level cut scores were set for Science assessments after the Spring 2019 test administration. The Spring 2019 assessment results will serve as the new baseline for monitoring student performance in Science across years. The 2021 Wisconsin Forward Exam administration will be the second administration of the Science assessments that measure the new Science standards with the test scores that are reported on the new Science scales.

Given the cancellation of the Spring 2020 test administration, special attention should be given to the assessment results in Spring 2021. It is not known whether the COVID-19 pandemic and resulting school closures, economic impacts, and trauma of the recent events may have long-lasting effects on students and their families. While it is difficult to anticipate all unintended consequences of the COVID-19 pandemic, it is possible that it may affect student achievement in subsequent academic years and contribute to long-standing achievement gaps between students from different ethnic groups, students with and without disabilities, students whose families are and are not socioeconomically disadvantaged, limited English proficiency and fully English proficient students, and students who typically need or do not need testing accommodations. Therefore, any significant shifts in student performance between Spring 2019 and Spring 2021 should be carefully examined, and all possible contributing factors should be taken into consideration while making decisions for accountability purposes.

Because the test forms developed for the Spring 2020 administration were not administered and remain secure, these forms can be administered in the future. DPI may also consider reusing previous operational test forms in Spring 2021 test administration. Reusing previously administered test forms will allow for the evaluation of changes in student performance between test administrations not only at the total test level (using total test scale scores) but also at the content standard and item level (using item level scores for items common in both administrations). If previous operational test forms are reused in the Spring 2021 test administration, the field test positions on these forms will be filled with new field test items.

DRC will review all test items in all forms for content and sensitivity issues that may be directly or indirectly related to the COVID-19 pandemic and the social and economic events that followed. Operational test items (and passages, if needed), flagged for content and sensitivity issues should be replaced, if possible, given the test blueprint constraints, with items that have neutral content, have acceptable field test statistics, and measure the same standards.

In addition, because of the missed test administration in Spring 2020, DRC recommends increasing the number of field test items and field test forms in Spring 2021. Embedded field-testing for all content areas should continue to be used in order to build a high-quality Wisconsin item bank for future form development.

DRC recommends continuing to use an artificial intelligence (AI) engine in the scoring of text-dependent analysis items for its efficiency and accuracy. As indicated in past Technical Reports, the AI scores were in good agreement with scores by trained human scorers.

References

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

CTB/McGraw-Hill. (1997). *TerraNova* (1st ed.). Monterey, CA: Author.

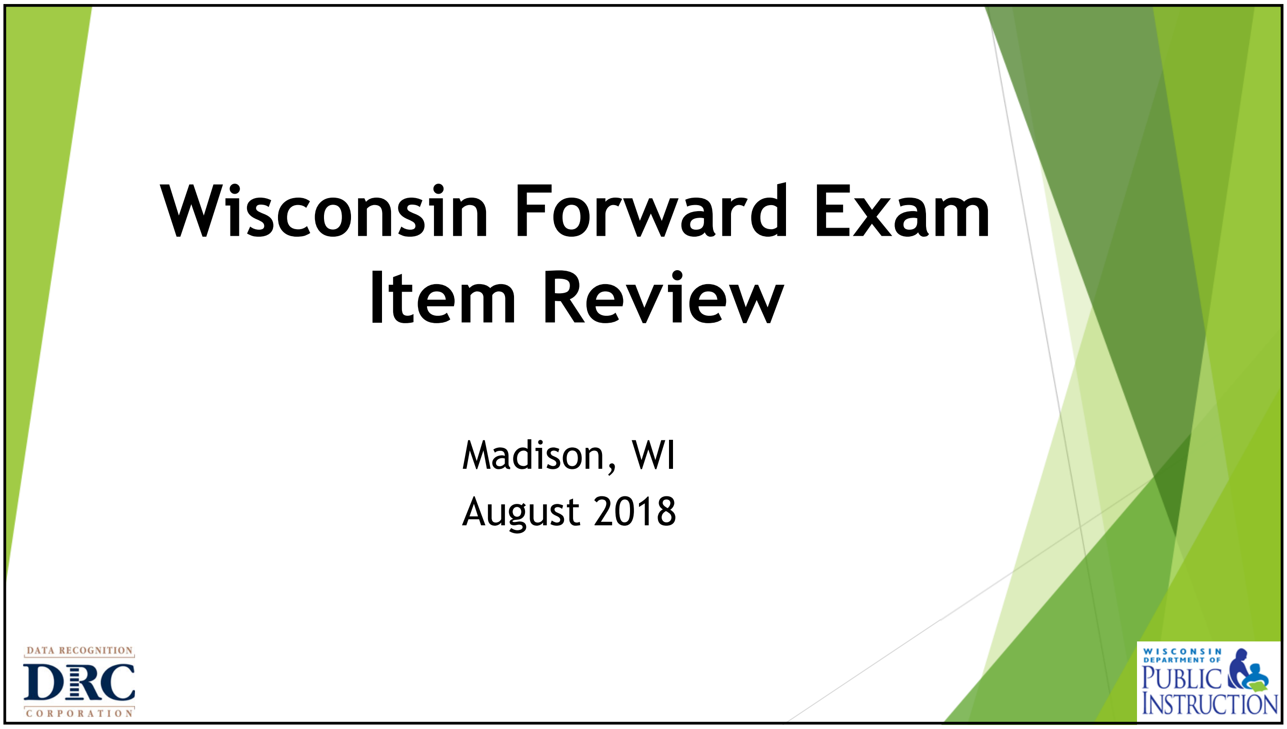
CTB/McGraw-Hill. (2000). *TerraNova* (2nd ed.). Monterey, CA: Author.

CTB/McGraw-Hill. (2009). *TerraNova 3rd Edition Technical Addendum: Forms E and F*. Monterey, CA: Author.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessment* (Synthesis Report No. 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Appendix A

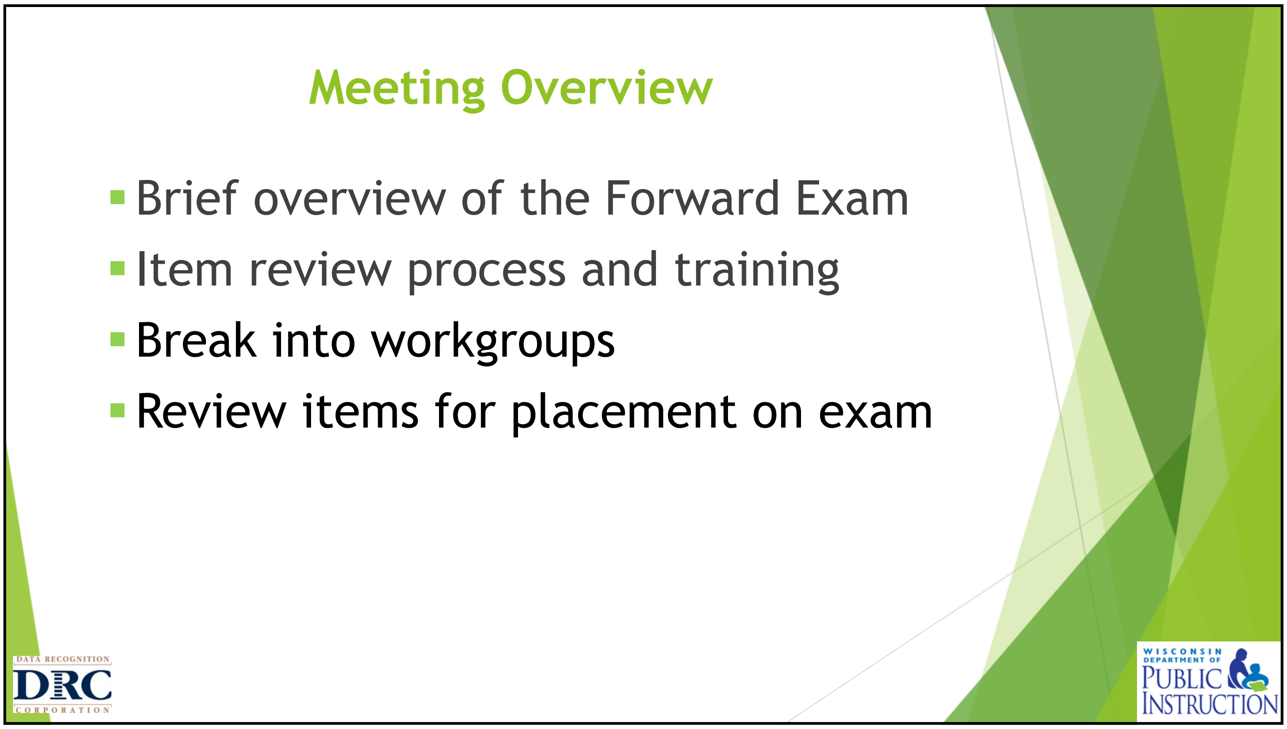
Summer 2018 Item Review Training Slides



Wisconsin Forward Exam Item Review

Madison, WI
August 2018

1



Meeting Overview

- Brief overview of the Forward Exam
- Item review process and training
- Break into workgroups
- Review items for placement on exam

2

Roles & Responsibilities

Participants

- Item Review

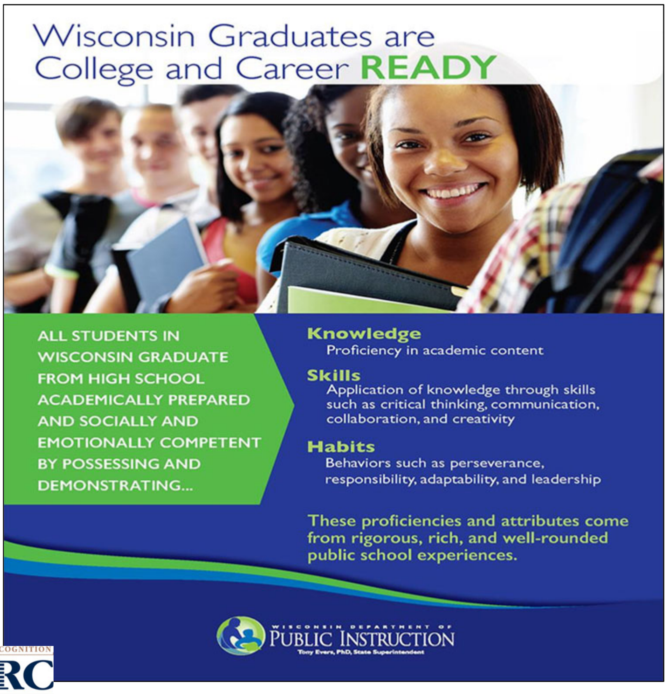
DRC Facilitators

- Lead the group through the agenda
- Encourage interaction
- Lead discussions
- Collect secure materials

DPI and DRC

- Answer questions

3



Wisconsin Graduates are College and Career **READY**

ALL STUDENTS IN WISCONSIN GRADUATE FROM HIGH SCHOOL ACADEMICALLY PREPARED AND SOCIALLY AND EMOTIONALLY COMPETENT BY POSSESSING AND DEMONSTRATING...

Knowledge
Proficiency in academic content

Skills
Application of knowledge through skills such as critical thinking, communication, collaboration, and creativity

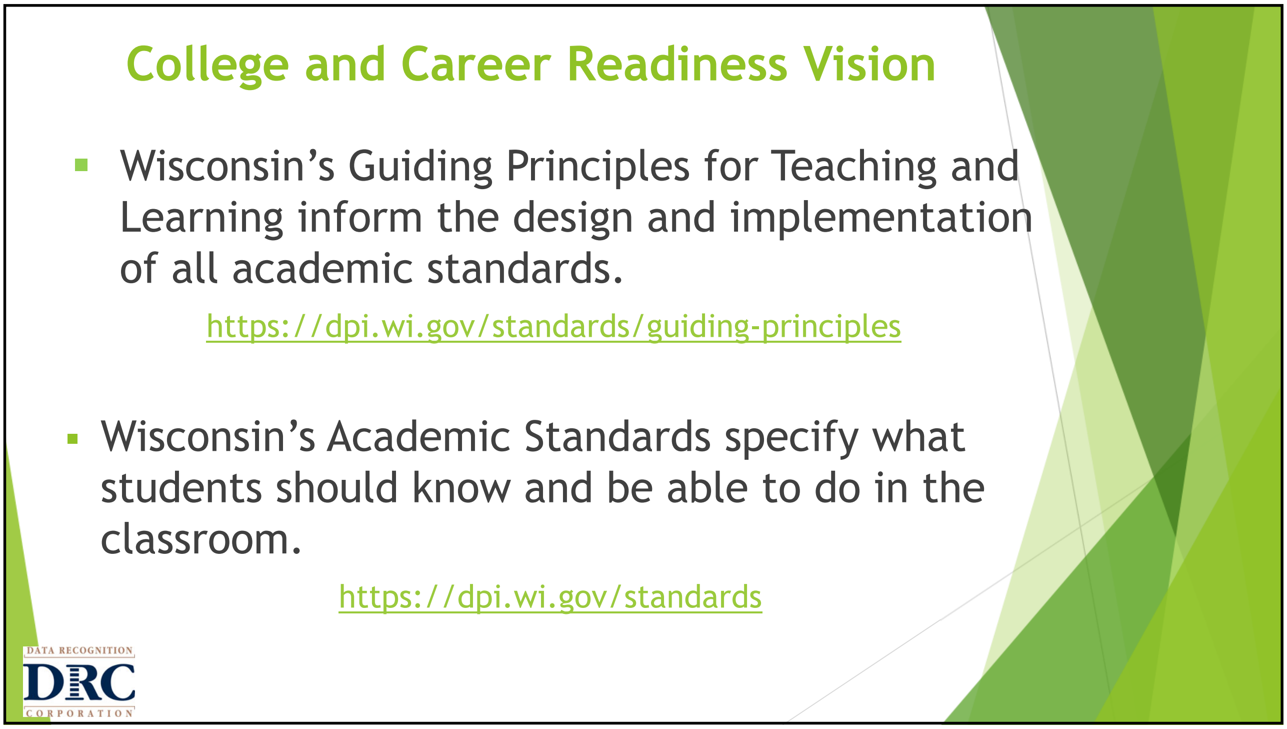
Habits
Behaviors such as perseverance, responsibility, adaptability, and leadership

These proficiencies and attributes come from rigorous, rich, and well-rounded public school experiences.

WISCONSIN DEPARTMENT OF PUBLIC INSTRUCTION
Tony Evers, PhD, State Superintendent

Wisconsin's Definition of College and Career Readiness

4



College and Career Readiness Vision

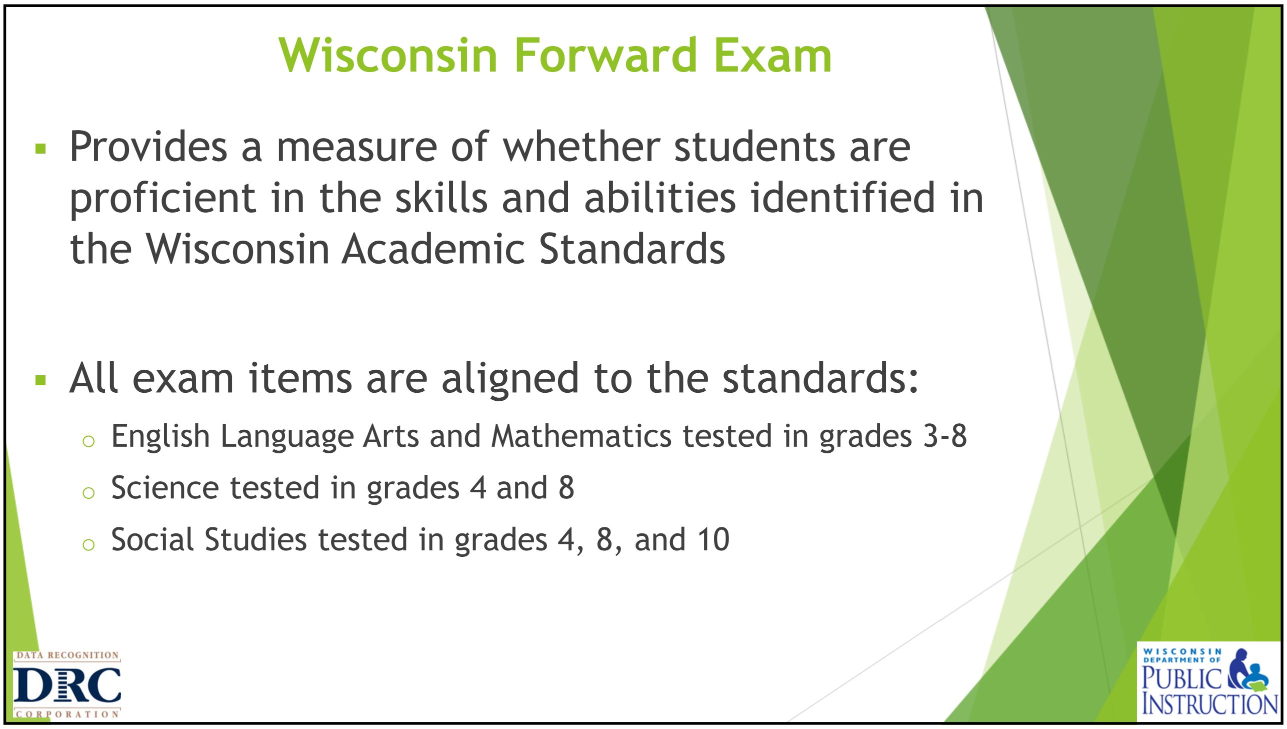
- Wisconsin’s Guiding Principles for Teaching and Learning inform the design and implementation of all academic standards.

  https://dpi.wi.gov/standards/guiding-principles

- Wisconsin’s Academic Standards specify what students should know and be able to do in the classroom.

  https://dpi.wi.gov/standards

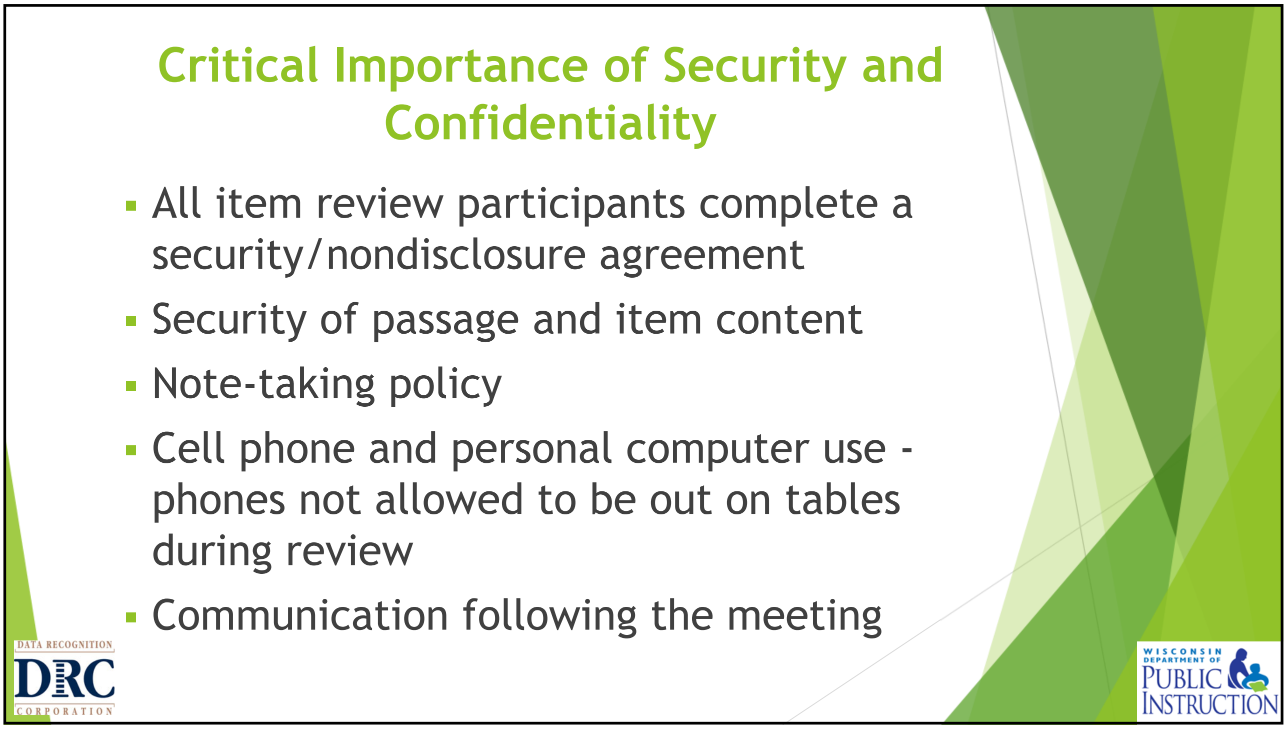
5



Wisconsin Forward Exam

- Provides a measure of whether students are proficient in the skills and abilities identified in the Wisconsin Academic Standards
- All exam items are aligned to the standards:
  - English Language Arts and Mathematics tested in grades 3-8
  - Science tested in grades 4 and 8
  - Social Studies tested in grades 4, 8, and 10

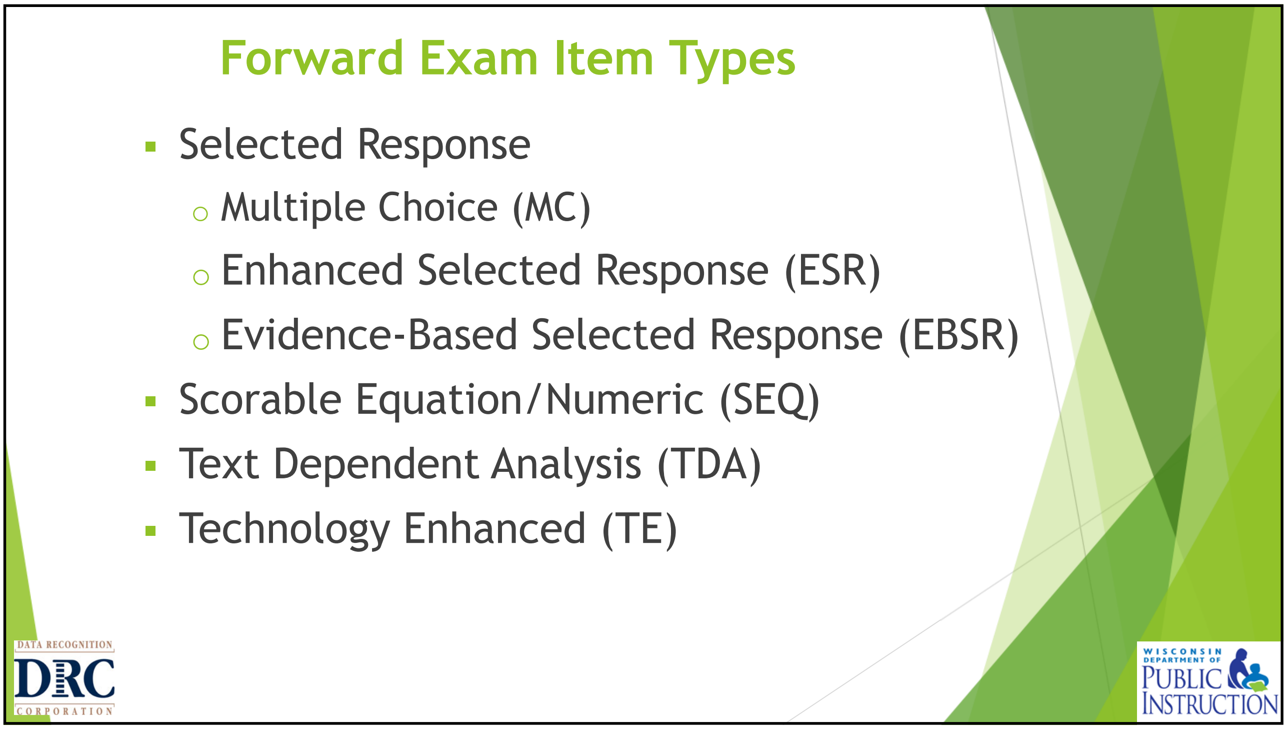
6



Critical Importance of Security and Confidentiality

- All item review participants complete a security/nondisclosure agreement
- Security of passage and item content
- Note-taking policy
- Cell phone and personal computer use - phones not allowed to be out on tables during review
- Communication following the meeting

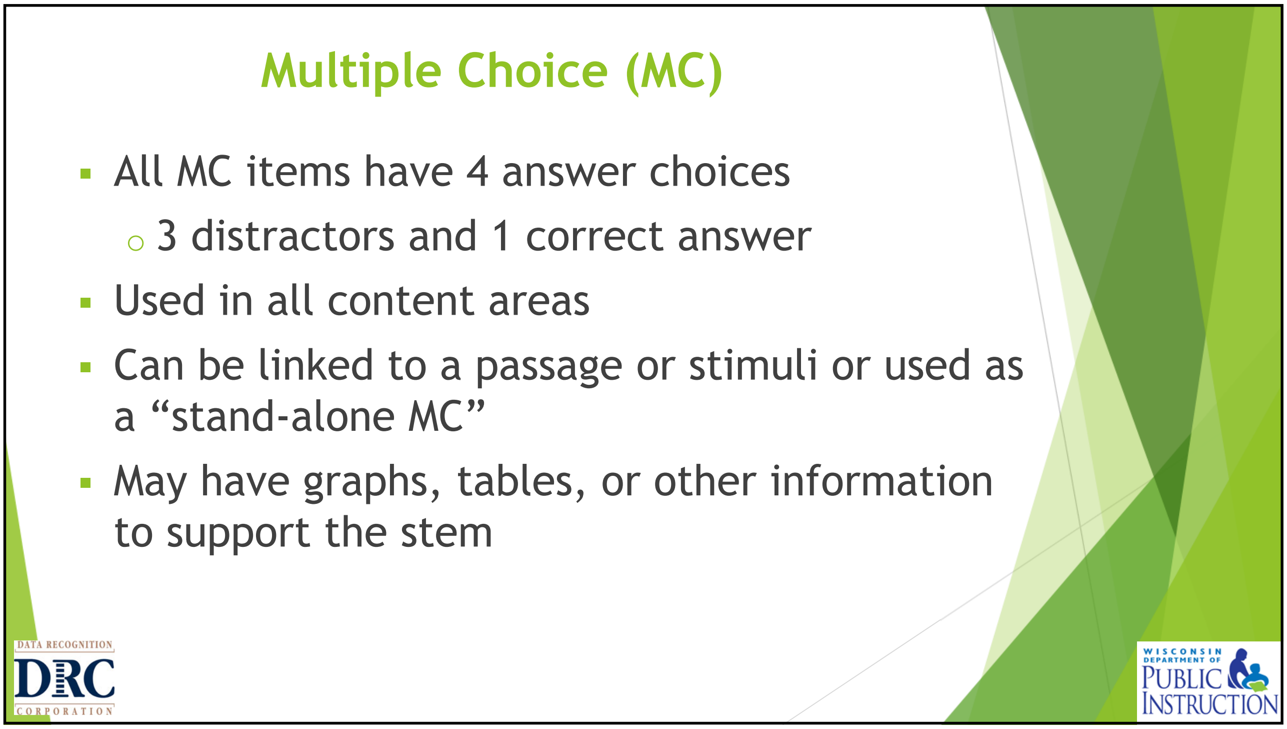
7



Forward Exam Item Types

- Selected Response
  - Multiple Choice (MC)
  - Enhanced Selected Response (ESR)
  - Evidence-Based Selected Response (EBSR)
- Scorable Equation/Numeric (SEQ)
- Text Dependent Analysis (TDA)
- Technology Enhanced (TE)

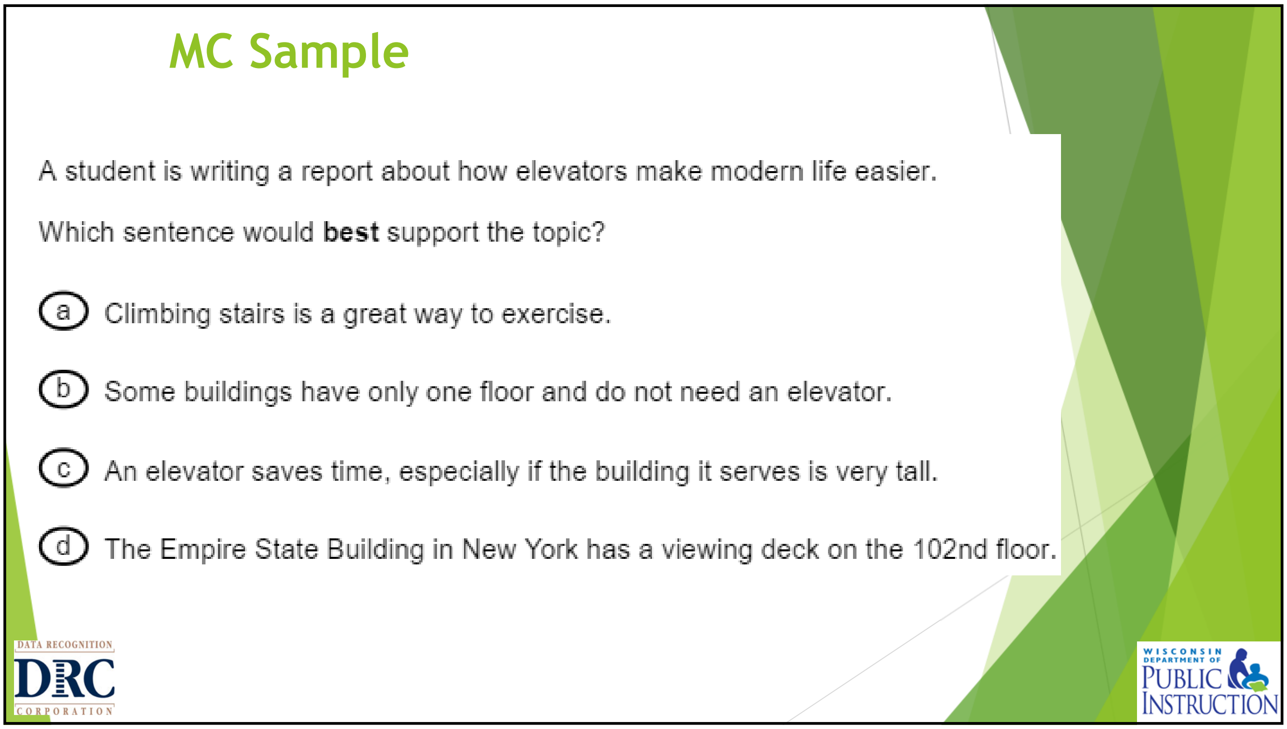
8



Multiple Choice (MC)

- All MC items have 4 answer choices
  - 3 distractors and 1 correct answer
- Used in all content areas
- Can be linked to a passage or stimuli or used as a “stand-alone MC”
- May have graphs, tables, or other information to support the stem

9



MC Sample

A student is writing a report about how elevators make modern life easier.

Which sentence would **best** support the topic?

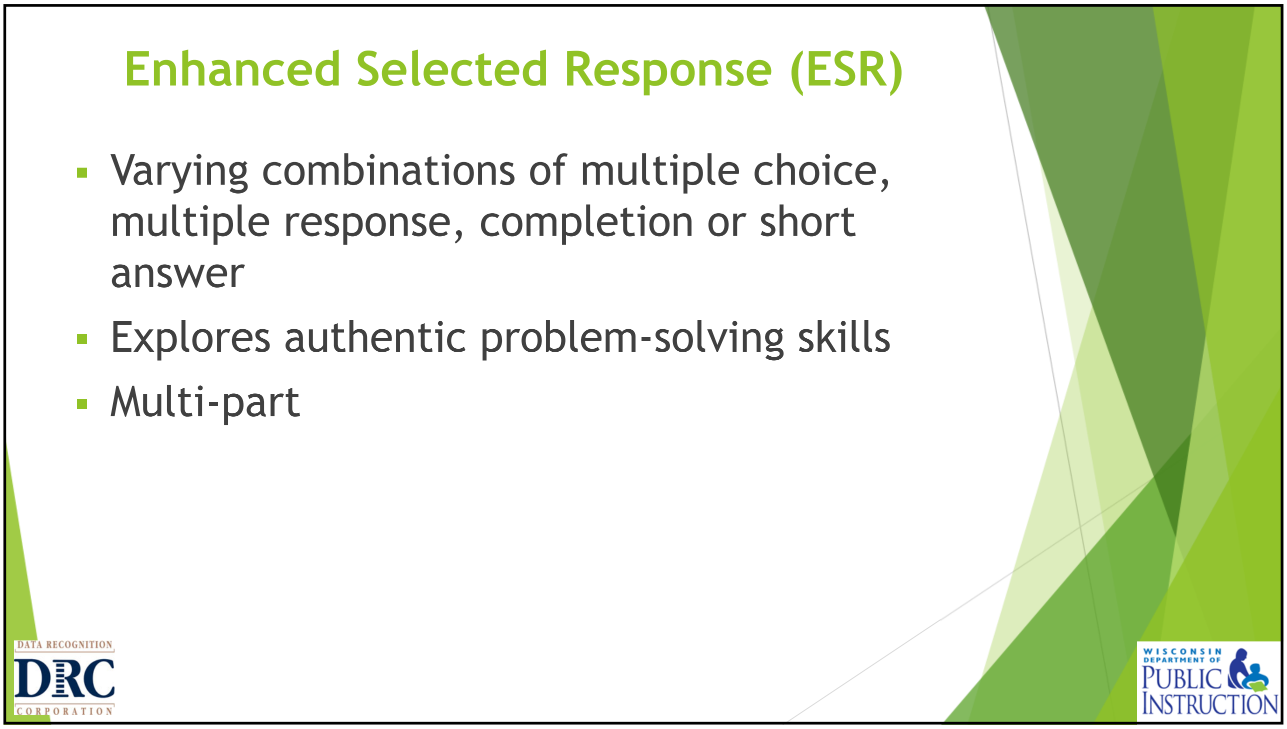
ⓐ Climbing stairs is a great way to exercise.

ⓑ Some buildings have only one floor and do not need an elevator.

ⓒ An elevator saves time, especially if the building it serves is very tall.

ⓓ The Empire State Building in New York has a viewing deck on the 102nd floor.

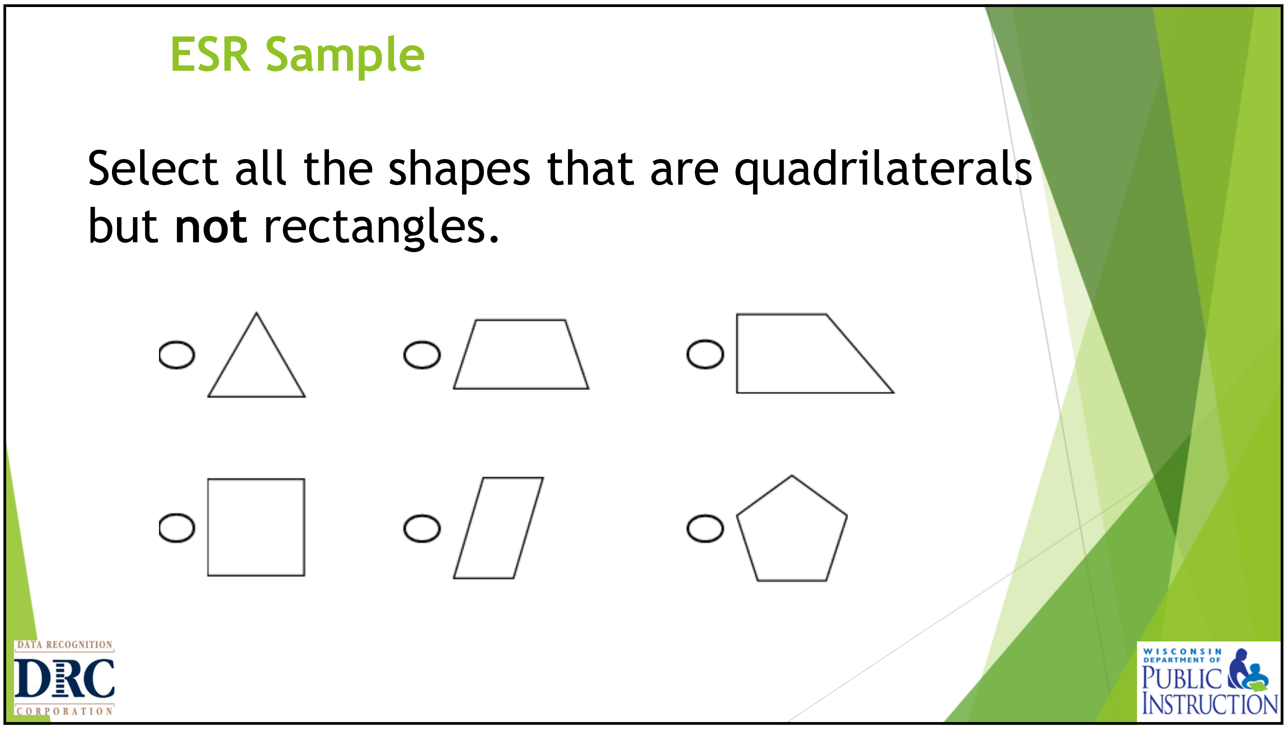
10



Enhanced Selected Response (ESR)

- Varying combinations of multiple choice, multiple response, completion or short answer
- Explores authentic problem-solving skills
- Multi-part

11



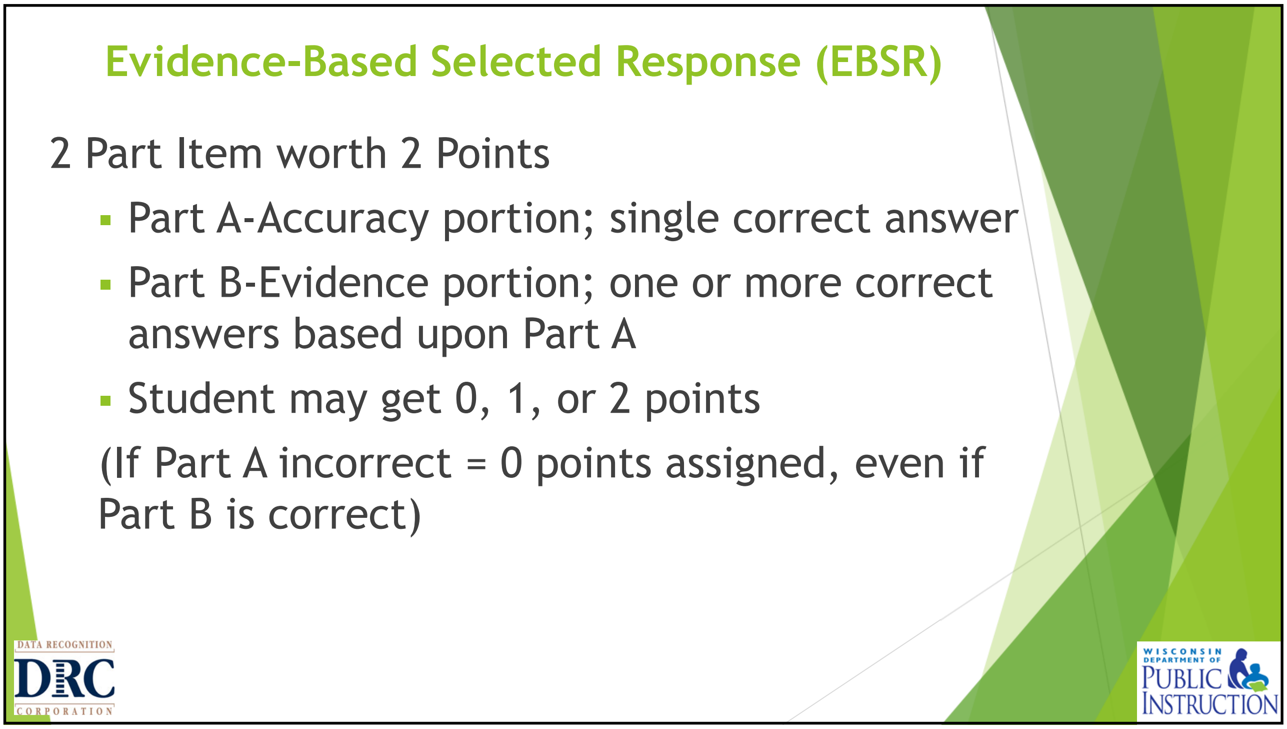
ESR Sample

Select all the shapes that are quadrilaterals but **not** rectangles.

○ ○ ○

○ ○ ○

12



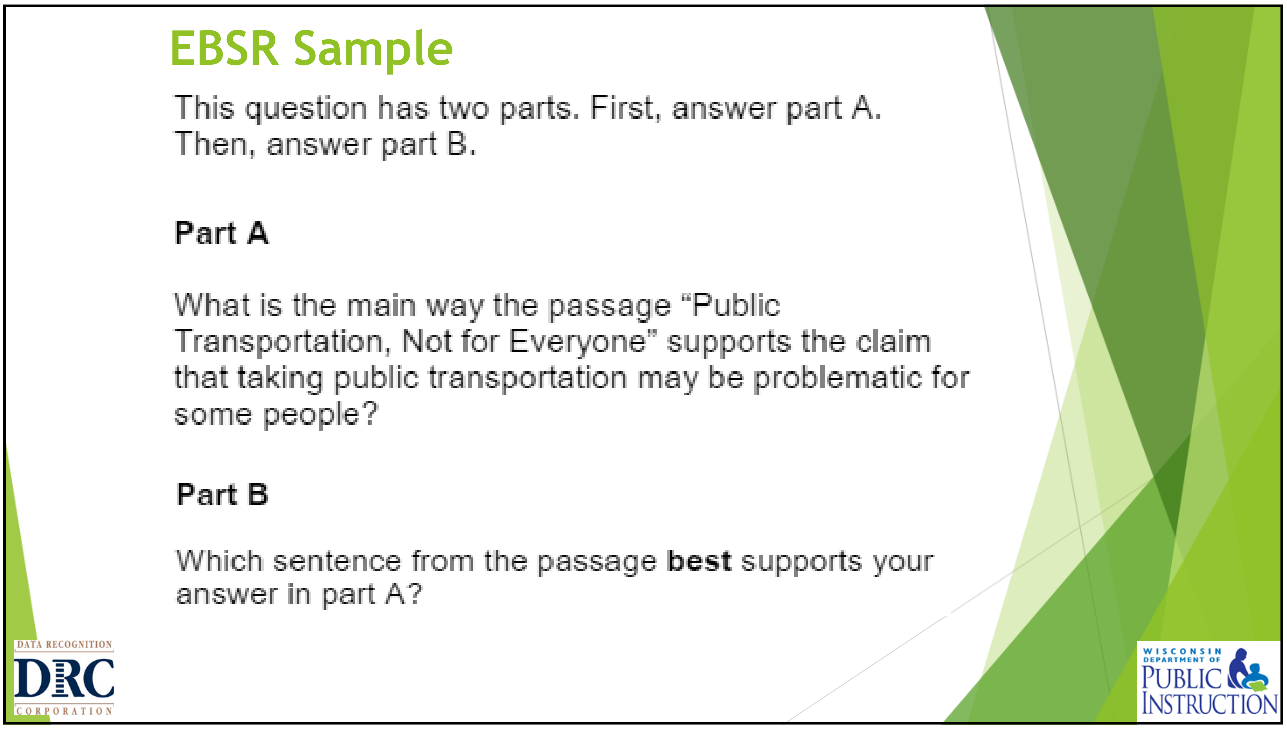
Evidence-Based Selected Response (EBSR)

2 Part Item worth 2 Points

- Part A-Accuracy portion; single correct answer
- Part B-Evidence portion; one or more correct answers based upon Part A
- Student may get 0, 1, or 2 points

(If Part A incorrect = 0 points assigned, even if Part B is correct)

13



EBSR Sample

This question has two parts. First, answer part A. Then, answer part B.

Part A

What is the main way the passage “Public Transportation, Not for Everyone” supports the claim that taking public transportation may be problematic for some people?

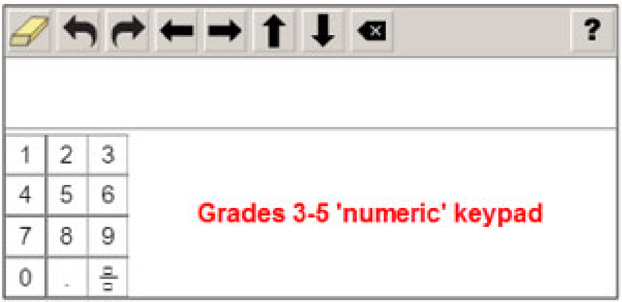
Part B

Which sentence from the passage **best** supports your answer in part A?

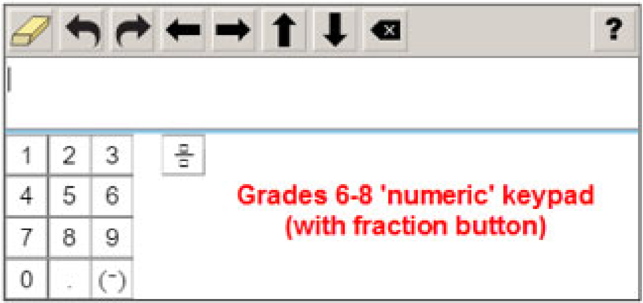
14

Scorable Equation/Numeric Item Type (SEQ)

- Used in Mathematics Items
- Grade-level specific keypad that allows for more guided input of student responses



Grades 3-5 'numeric' keypad

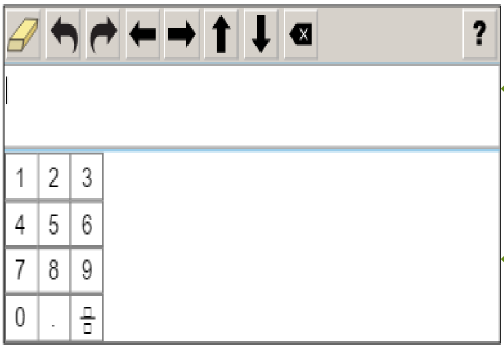


Grades 6-8 'numeric' keypad (with fraction button)

15

SEQ Sample

A rectangular section of a kitchen wall will be tiled. What is the area, in square feet, of the section of wall that will be tiled?



Student Response Area

Grade 5 Keypad

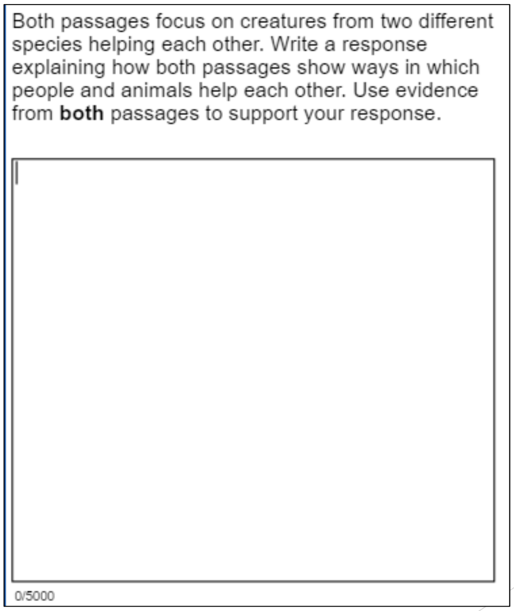
16

Text Dependent Analysis (TDA)

- Used in ELA assessment
- Based on a passage
- Used for both literature and informational texts
- Writing skills tested include inferring, analyzing, and synthesizing information from the passage
- Scored using a holistic scoring guide
- Character counter feature

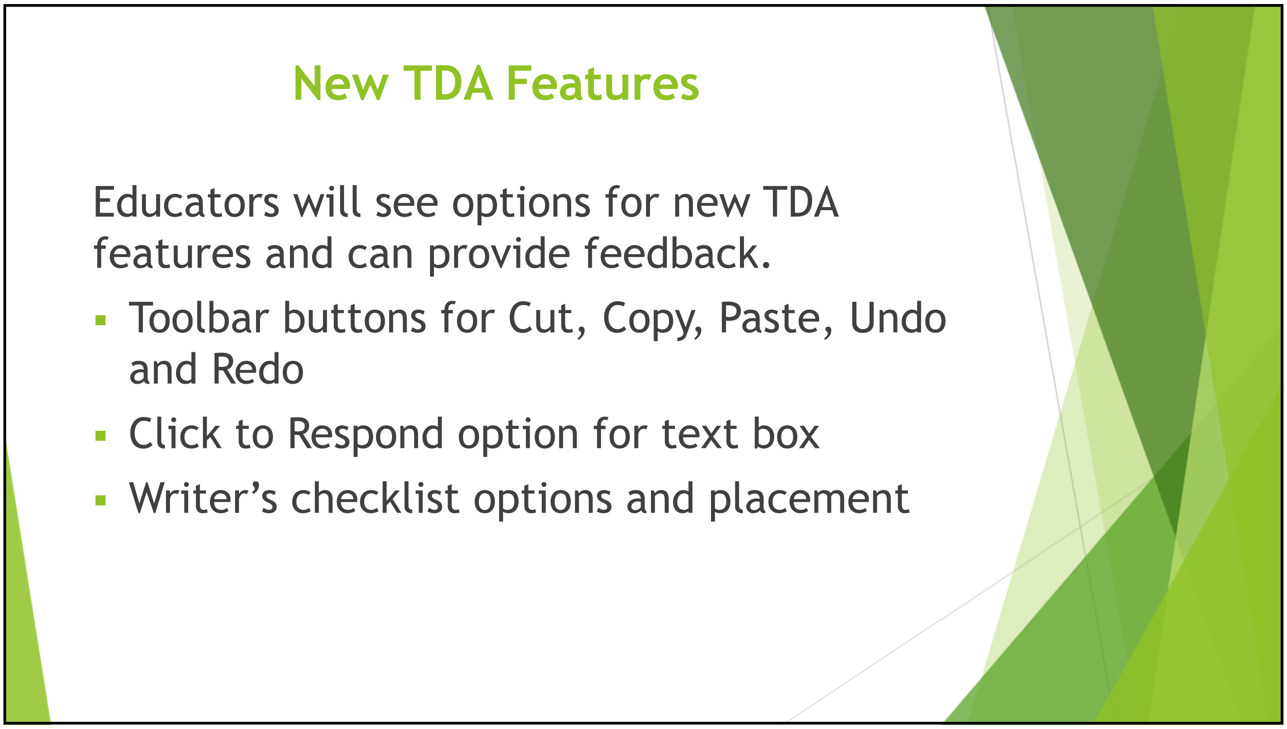
17

TDA Sample



Both passages focus on creatures from two different species helping each other. Write a response explaining how both passages show ways in which people and animals help each other. Use evidence from **both** passages to support your response.

18

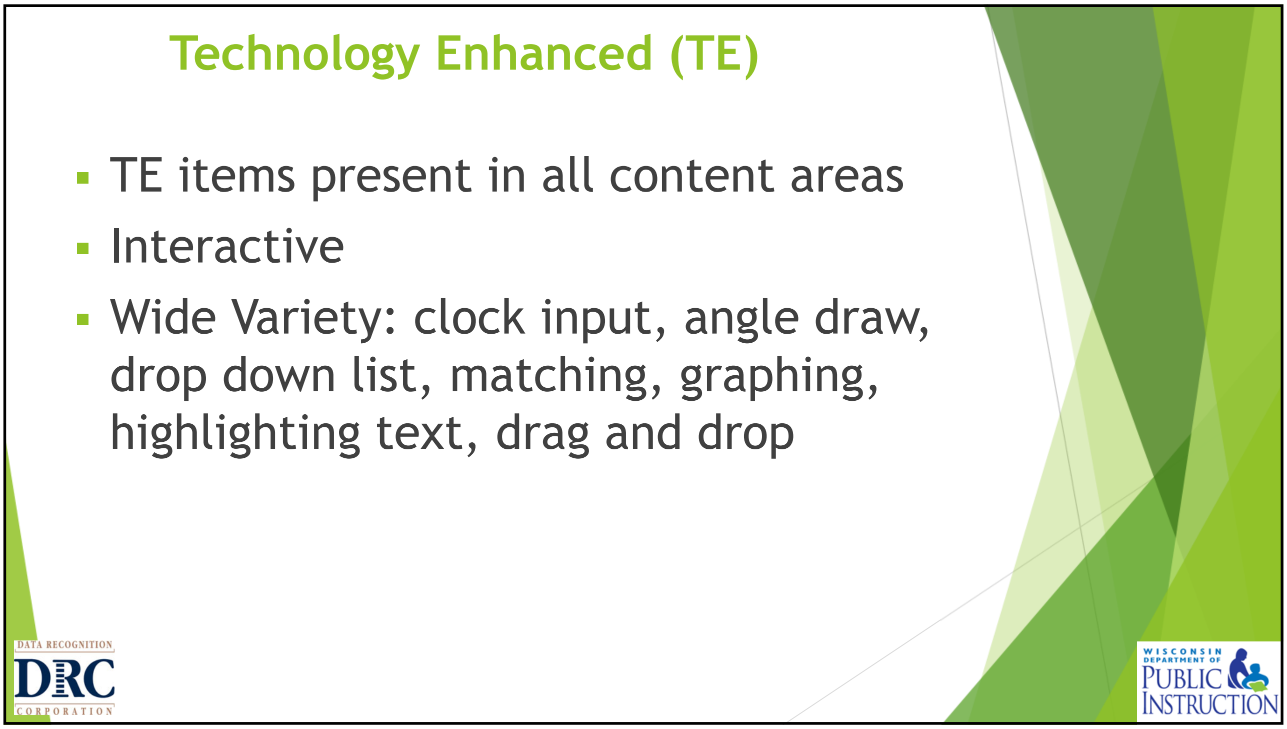


New TDA Features

Educators will see options for new TDA features and can provide feedback.

- Toolbar buttons for Cut, Copy, Paste, Undo and Redo
- Click to Respond option for text box
- Writer’s checklist options and placement

19



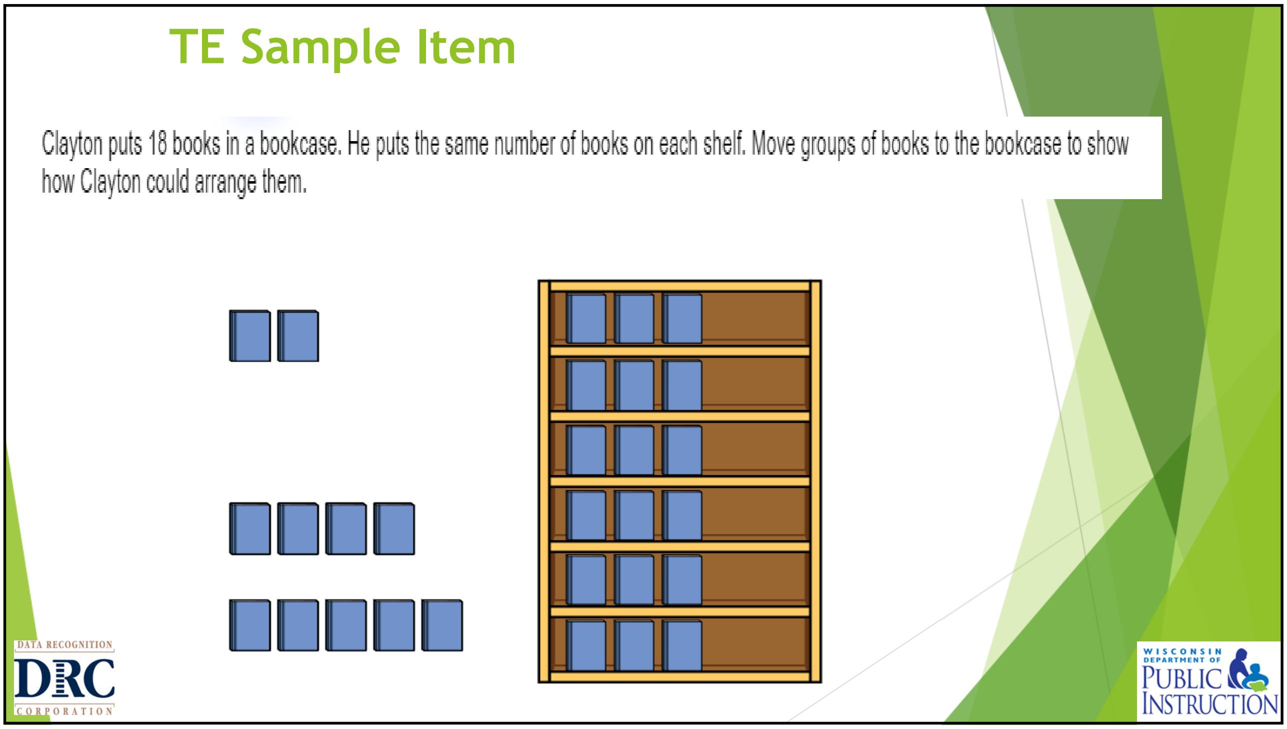
Technology Enhanced (TE)

- TE items present in all content areas
- Interactive
- Wide Variety: clock input, angle draw, drop down list, matching, graphing, highlighting text, drag and drop

DATA RECOGNITION DRC CORPORATION

WISCONSIN DEPARTMENT OF PUBLIC INSTRUCTION

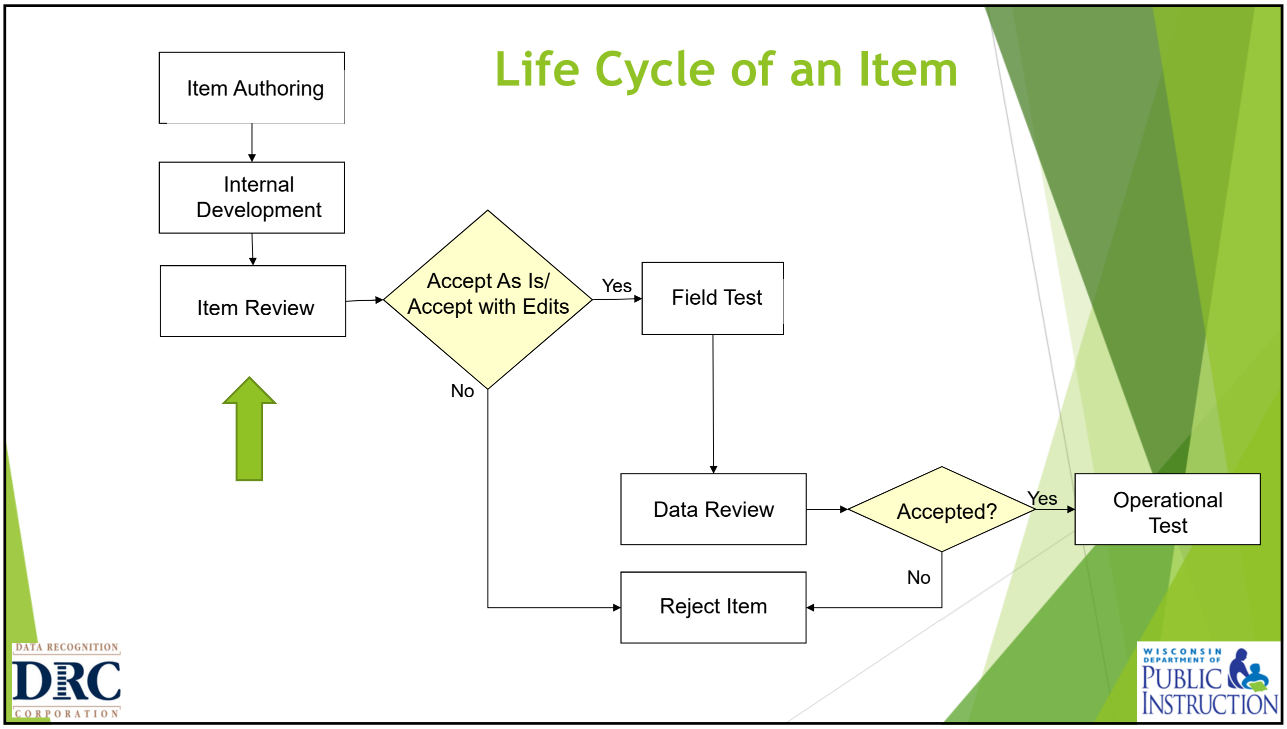
20



TE Sample Item

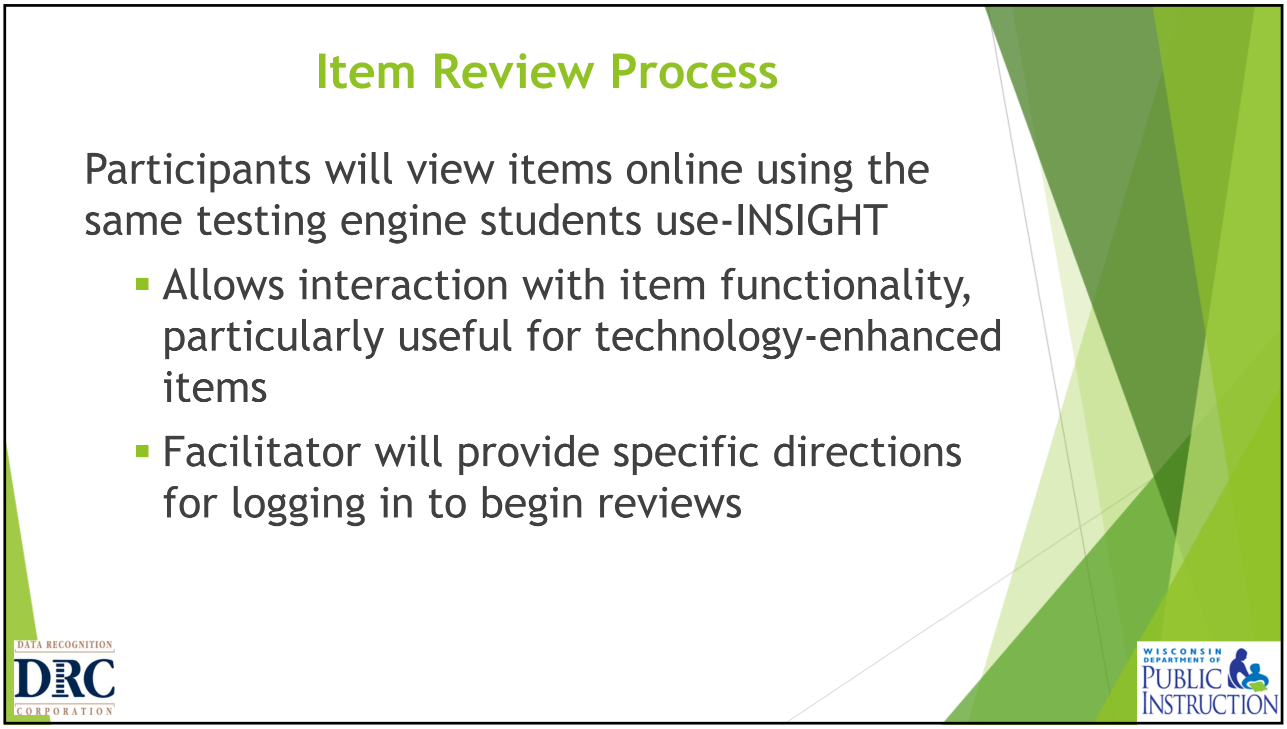
Clayton puts 18 books in a bookcase. He puts the same number of books on each shelf. Move groups of books to the bookcase to show how Clayton could arrange them.

21



Life Cycle of an Item

22



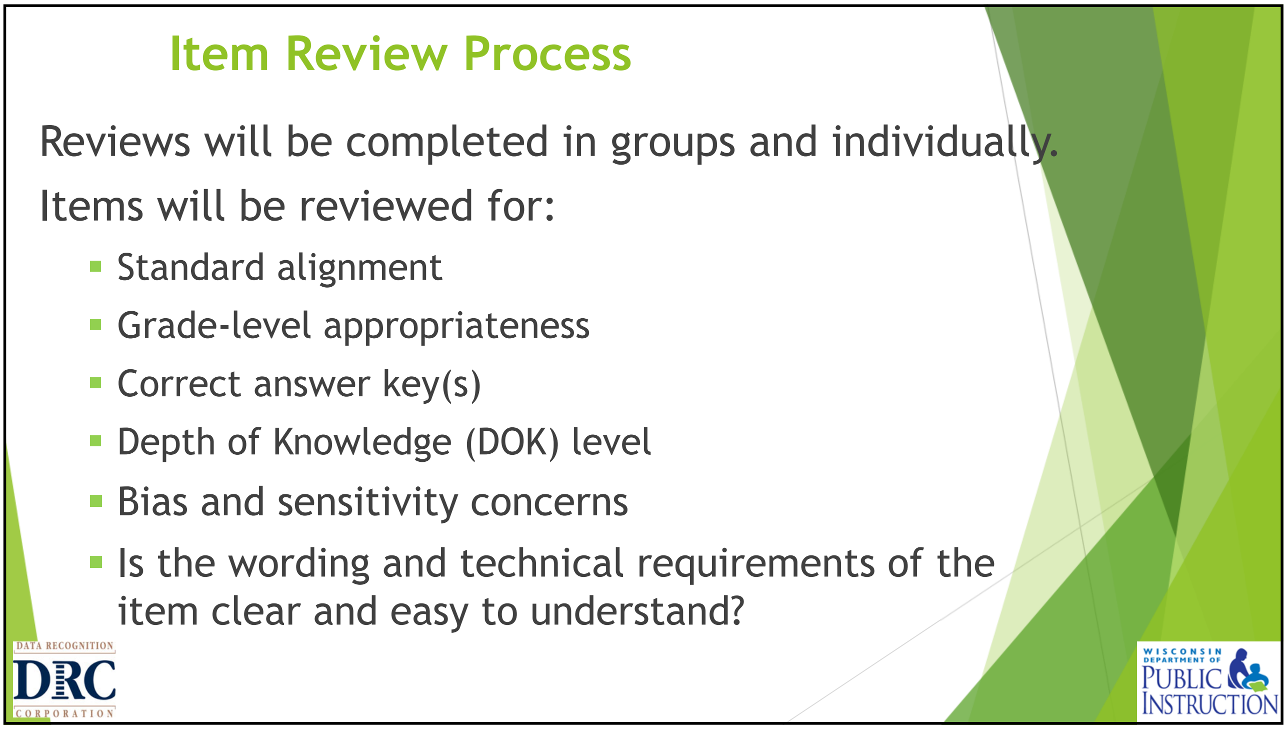
Item Review Process

Participants will view items online using the same testing engine students use-INSIGHT

- Allows interaction with item functionality, particularly useful for technology-enhanced items
- Facilitator will provide specific directions for logging in to begin reviews

DATA RECOGNITION
DRC
CORPORATION
WISCONSIN DEPARTMENT OF
PUBLIC INSTRUCTION

23



Item Review Process

Reviews will be completed in groups and individually.

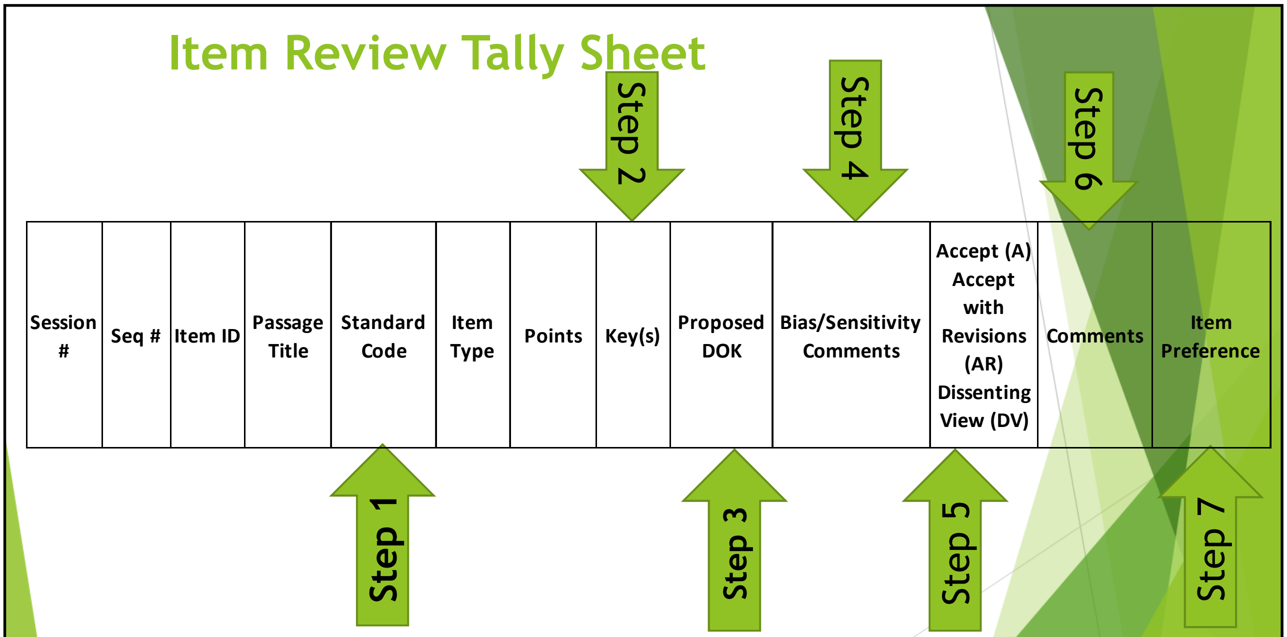
Items will be reviewed for:

- Standard alignment
- Grade-level appropriateness
- Correct answer key(s)
- Depth of Knowledge (DOK) level
- Bias and sensitivity concerns
- Is the wording and technical requirements of the item clear and easy to understand?

DATA RECOGNITION
DRC
CORPORATION
WISCONSIN DEPARTMENT OF
PUBLIC INSTRUCTION

24

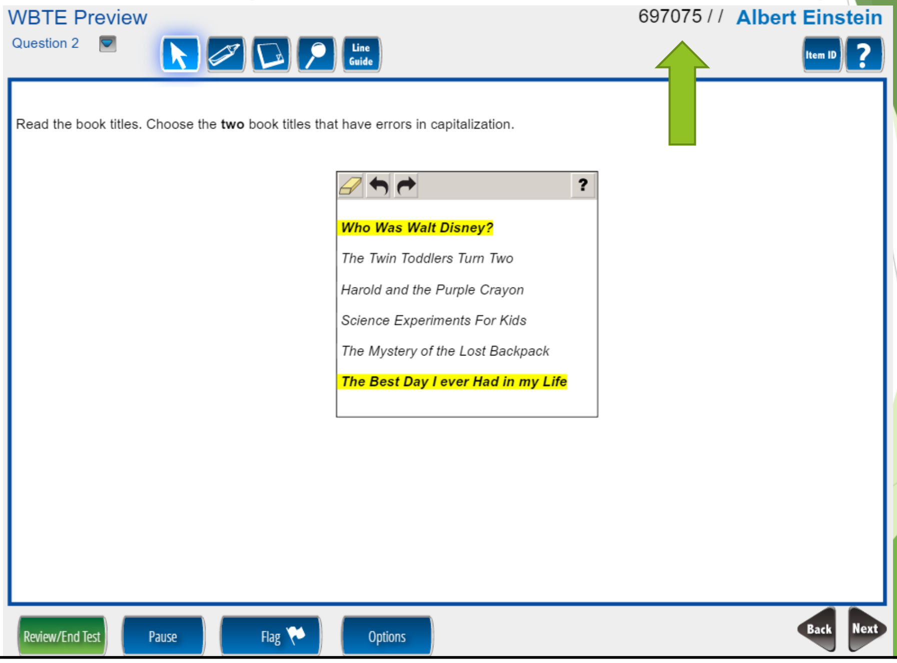
Item Review Tally Sheet



Session #	Seq #	Item ID	Passage Title	Standard Code	Item Type	Points	Key(s)	Proposed DOK	Bias/Sensitivity Comments	Accept (A) Accept with Revisions (AR) Dissenting View (DV)	Comments	Item Preference

25

Evaluating an Item: Grade 3 Writing



Read the book titles. Choose the **two** book titles that have errors in capitalization.

Who Was Walt Disney?

The Twin Toddlers Turn Two

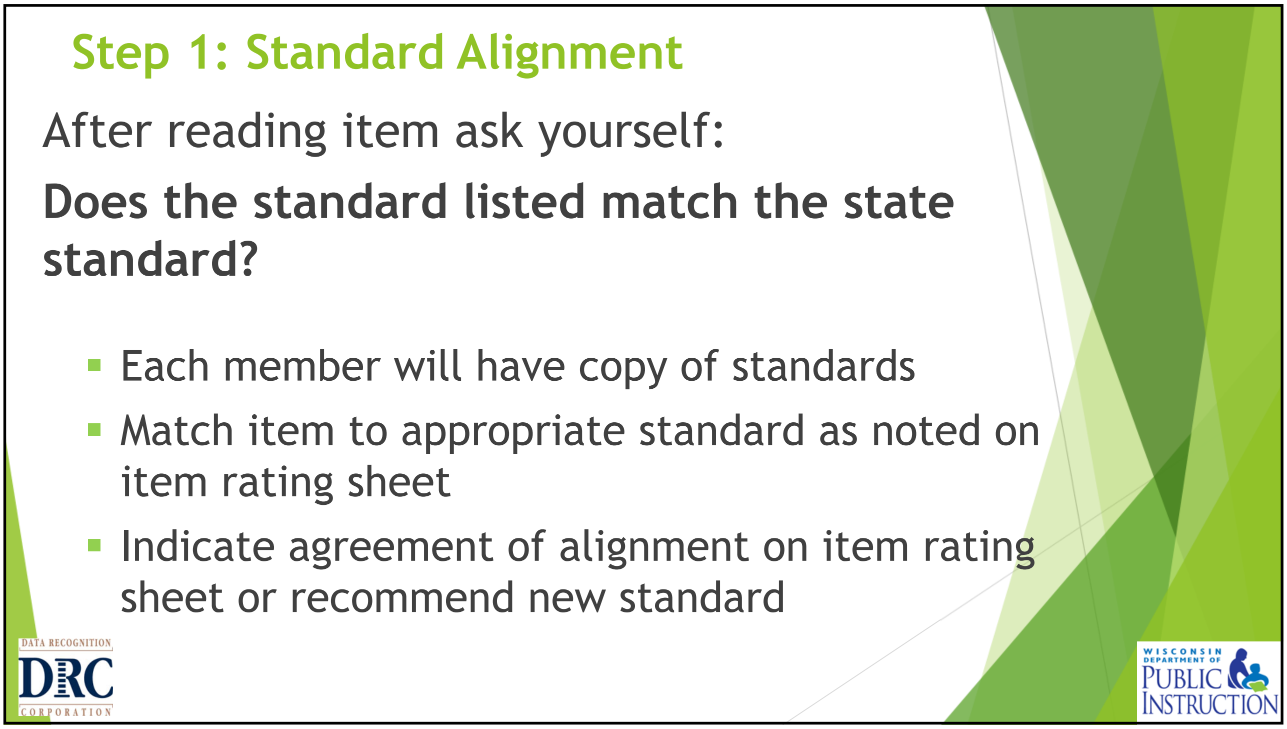
Harold and the Purple Crayon

Science Experiments For Kids

The Mystery of the Lost Backpack

The Best Day I ever Had in my Life

26



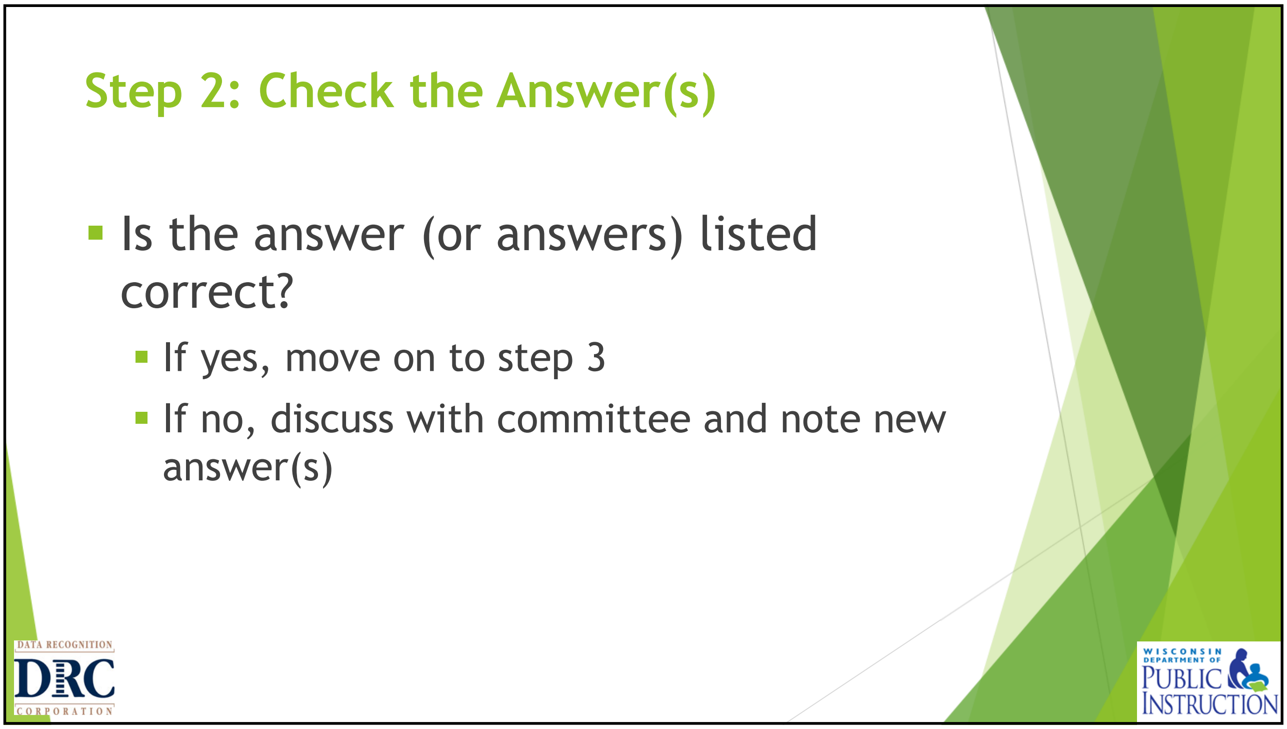
Step 1: Standard Alignment

After reading item ask yourself:

Does the standard listed match the state standard?

- Each member will have copy of standards
- Match item to appropriate standard as noted on item rating sheet
- Indicate agreement of alignment on item rating sheet or recommend new standard

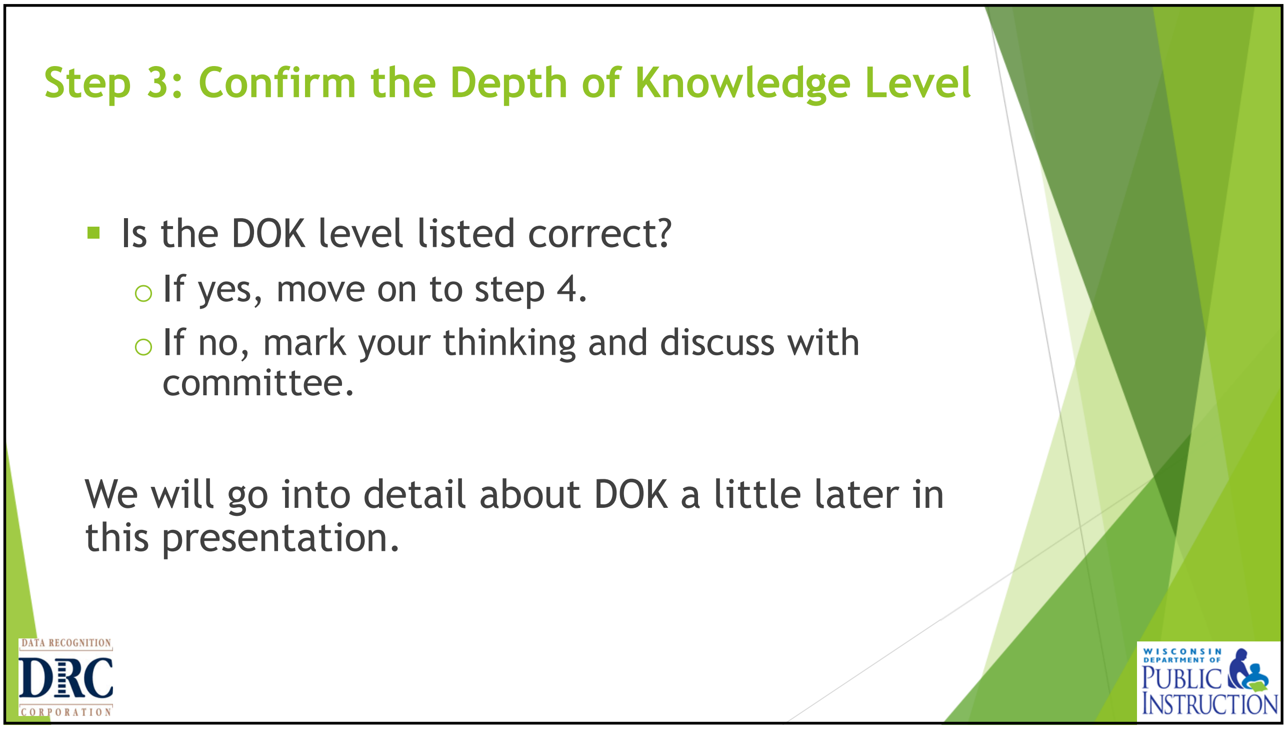
27



Step 2: Check the Answer(s)

- Is the answer (or answers) listed correct?
  - If yes, move on to step 3
  - If no, discuss with committee and note new answer(s)

28

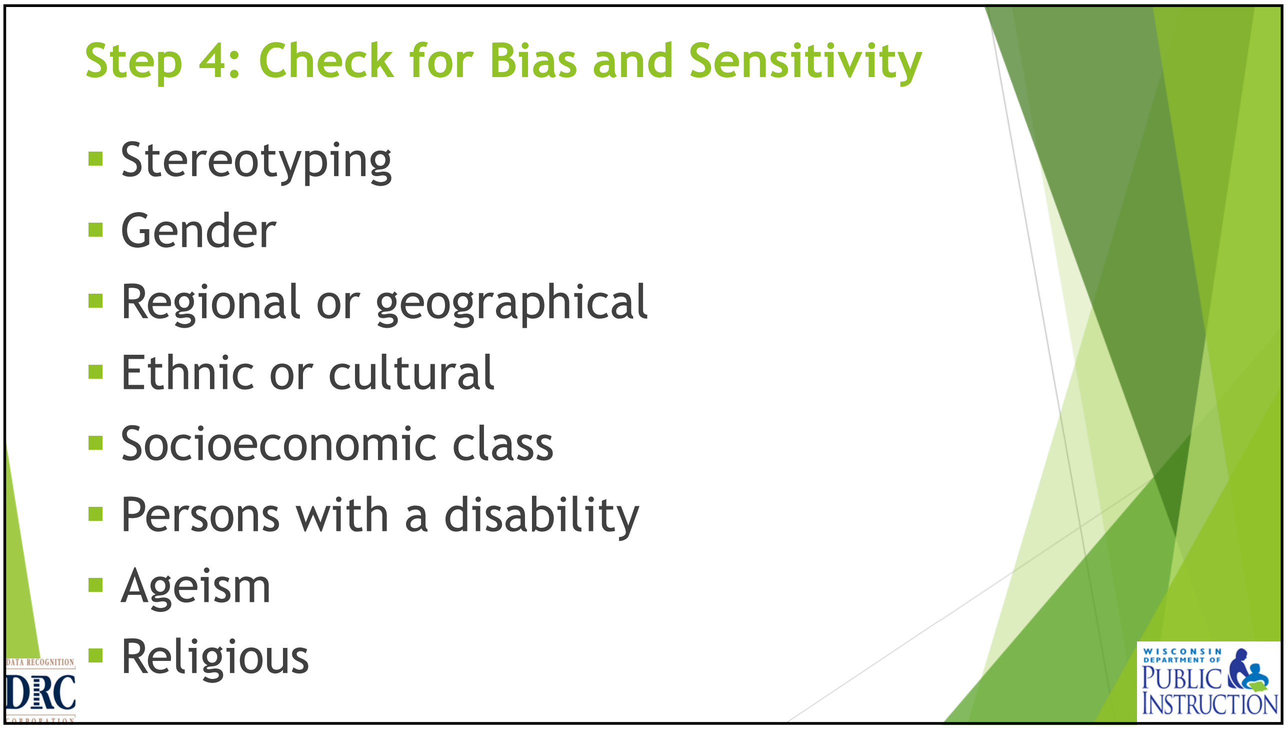


Step 3: Confirm the Depth of Knowledge Level

- Is the DOK level listed correct?
  - If yes, move on to step 4.
  - If no, mark your thinking and discuss with committee.

We will go into detail about DOK a little later in this presentation.

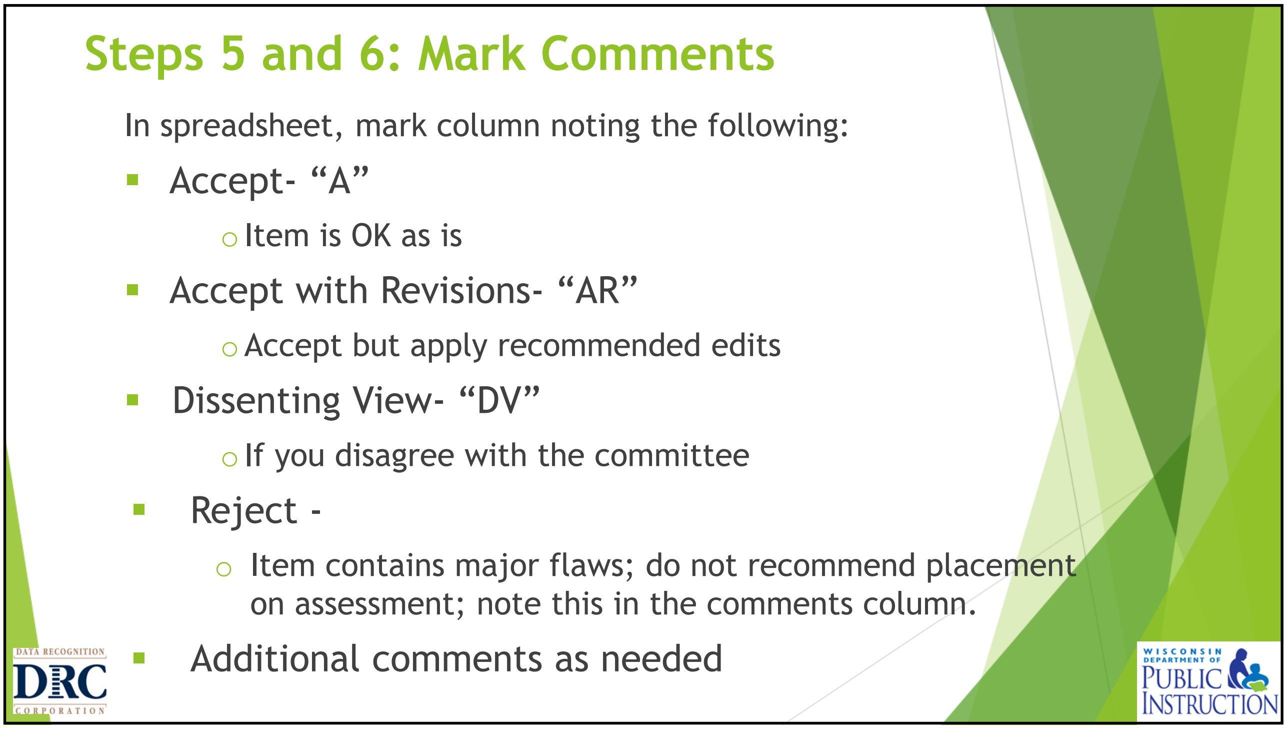
29



Step 4: Check for Bias and Sensitivity

- Stereotyping
- Gender
- Regional or geographical
- Ethnic or cultural
- Socioeconomic class
- Persons with a disability
- Ageism
- Religious

30



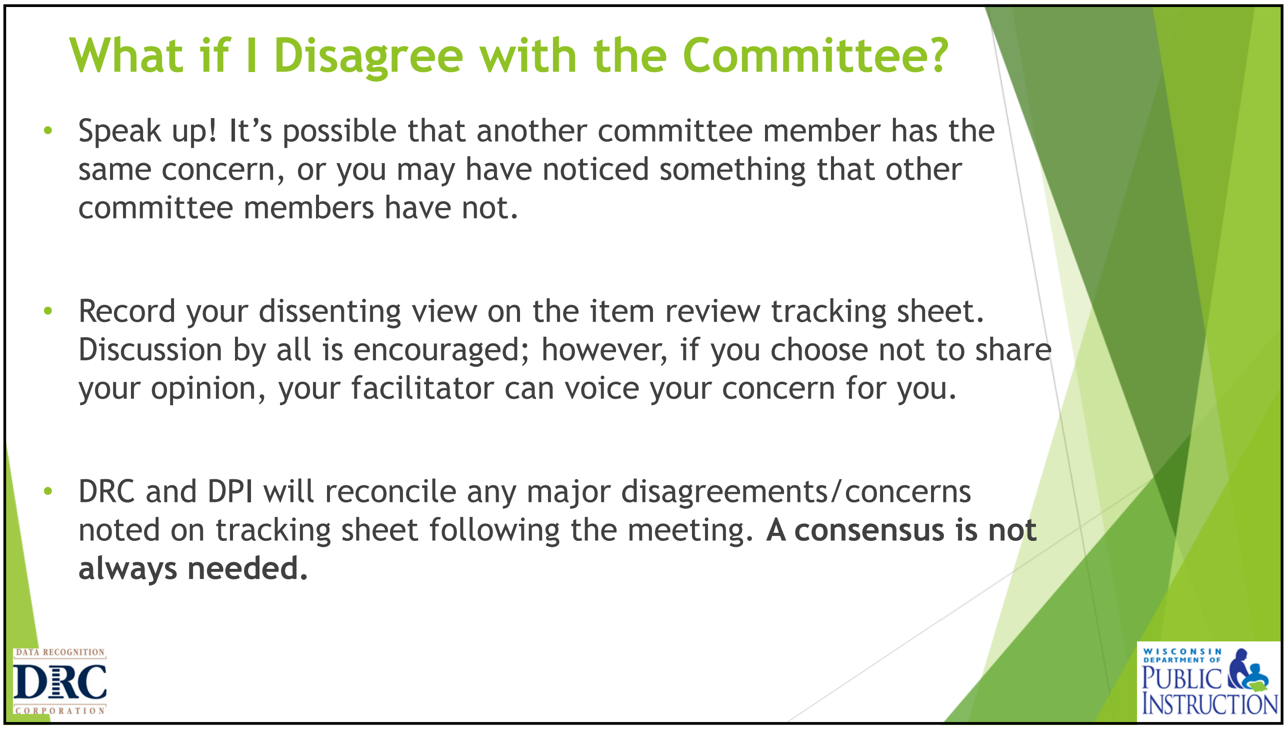
Steps 5 and 6: Mark Comments

In spreadsheet, mark column noting the following:

- Accept- “A”
  - Item is OK as is
- Accept with Revisions- “AR”
  - Accept but apply recommended edits
- Dissenting View- “DV”
  - If you disagree with the committee
- Reject -
  - Item contains major flaws; do not recommend placement on assessment; note this in the comments column.
- Additional comments as needed

DATA RECOGNITION
DRC
CORPORATION
WISCONSIN DEPARTMENT OF
PUBLIC INSTRUCTION

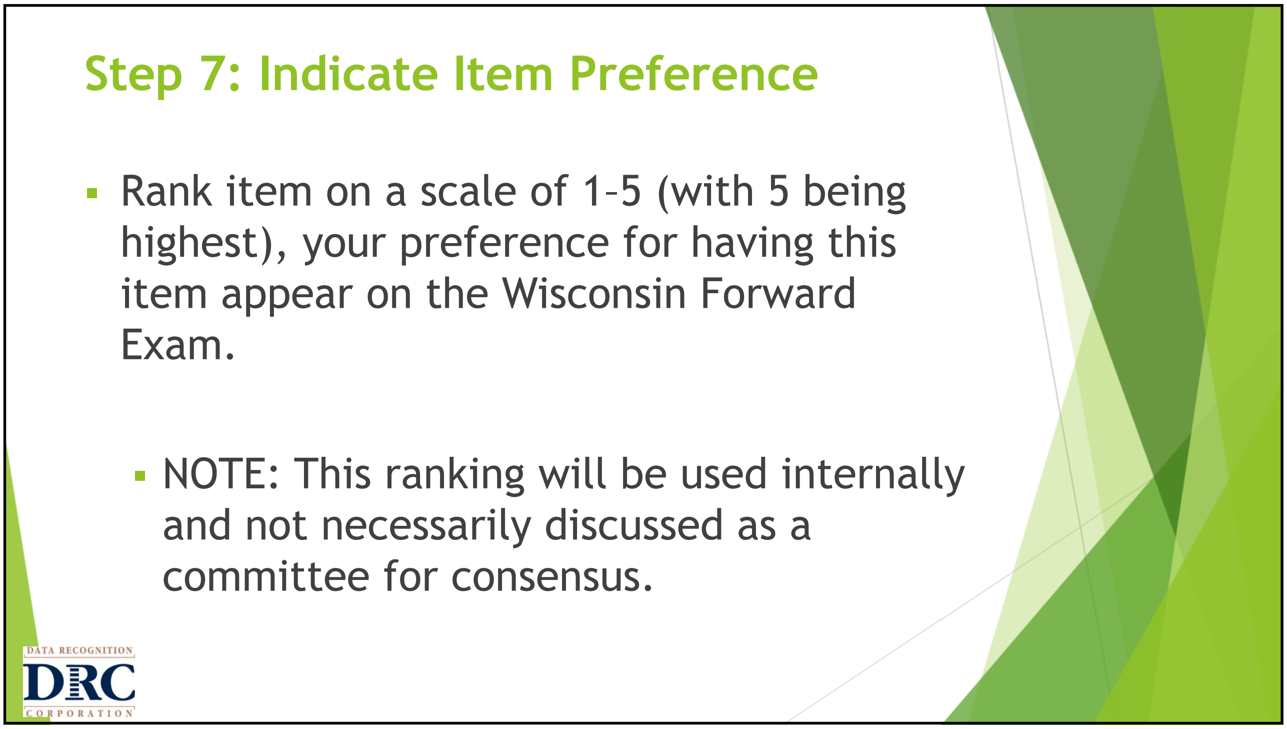
31



What if I Disagree with the Committee?

- Speak up! It’s possible that another committee member has the same concern, or you may have noticed something that other committee members have not.
- Record your dissenting view on the item review tracking sheet. Discussion by all is encouraged; however, if you choose not to share your opinion, your facilitator can voice your concern for you.
- DRC and DPI will reconcile any major disagreements/concerns noted on tracking sheet following the meeting. **A consensus is not always needed.**

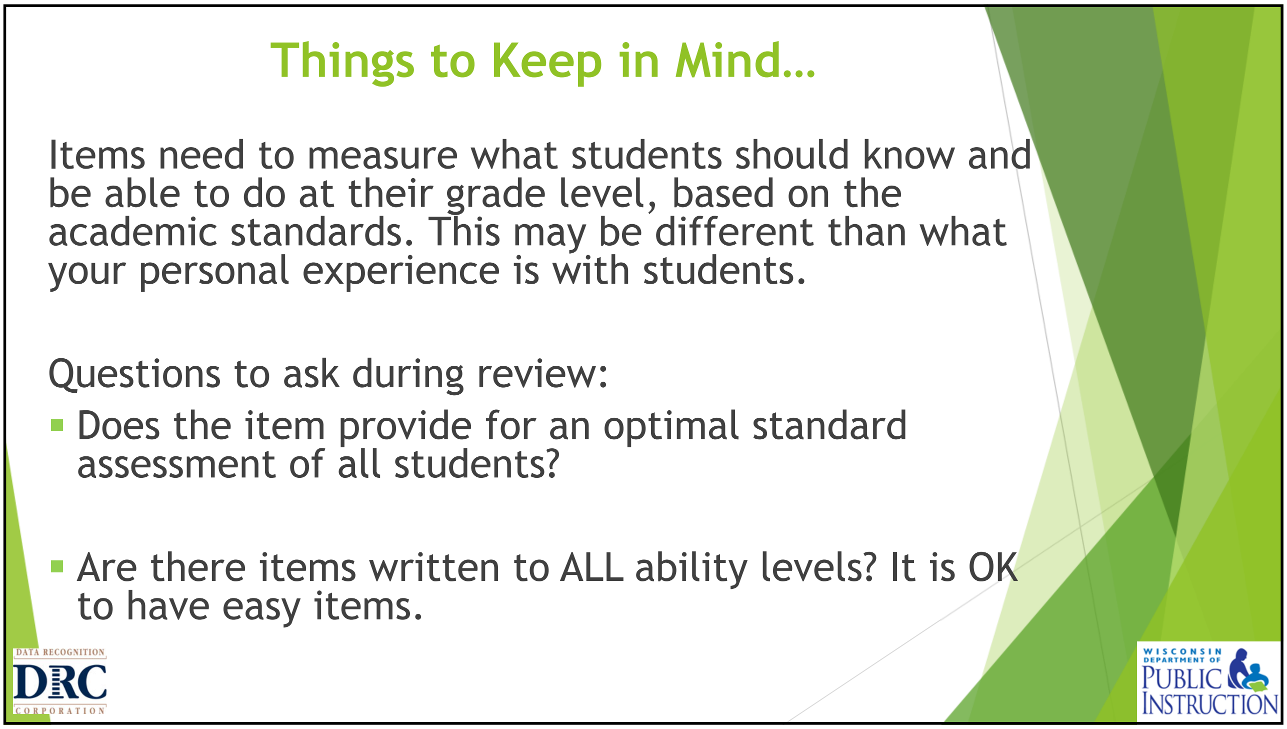
32



Step 7: Indicate Item Preference

- Rank item on a scale of 1–5 (with 5 being highest), your preference for having this item appear on the Wisconsin Forward Exam.
  - NOTE: This ranking will be used internally and not necessarily discussed as a committee for consensus.

33



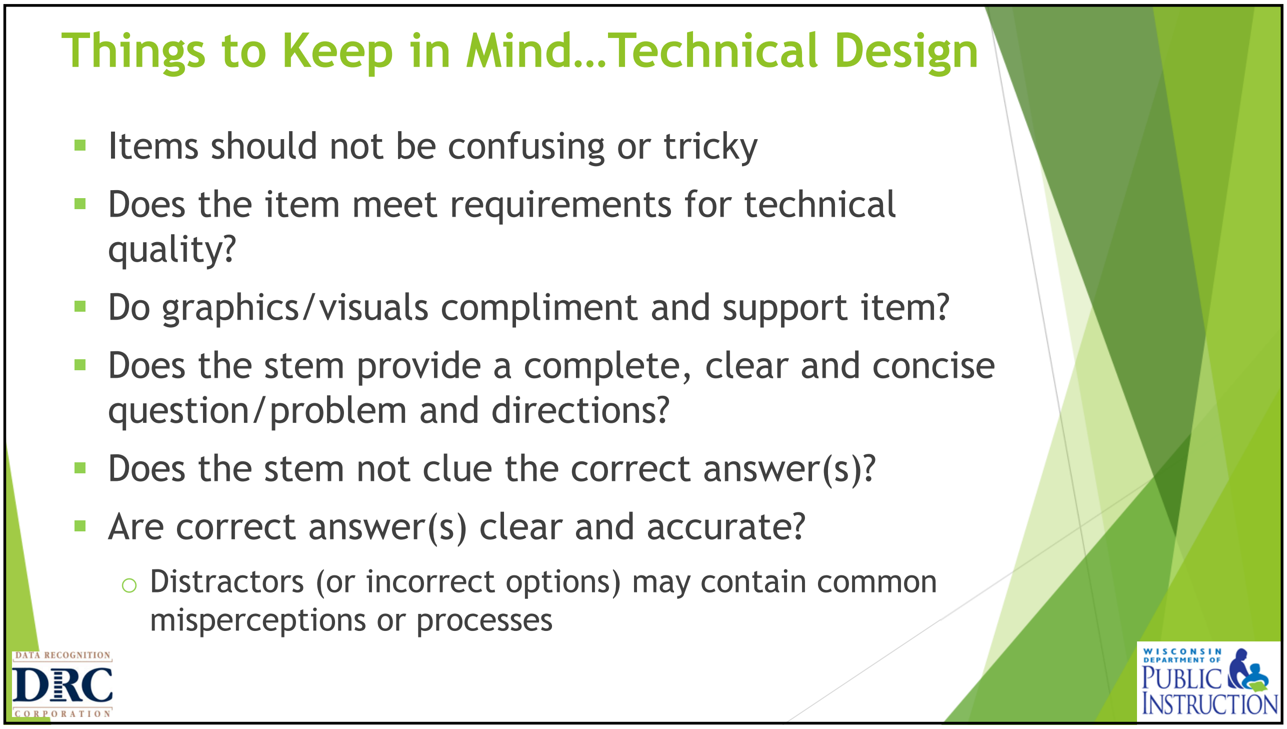
Things to Keep in Mind…

Items need to measure what students should know and be able to do at their grade level, based on the academic standards. This may be different than what your personal experience is with students.

Questions to ask during review:

- Does the item provide for an optimal standard assessment of all students?
- Are there items written to ALL ability levels? It is OK to have easy items.

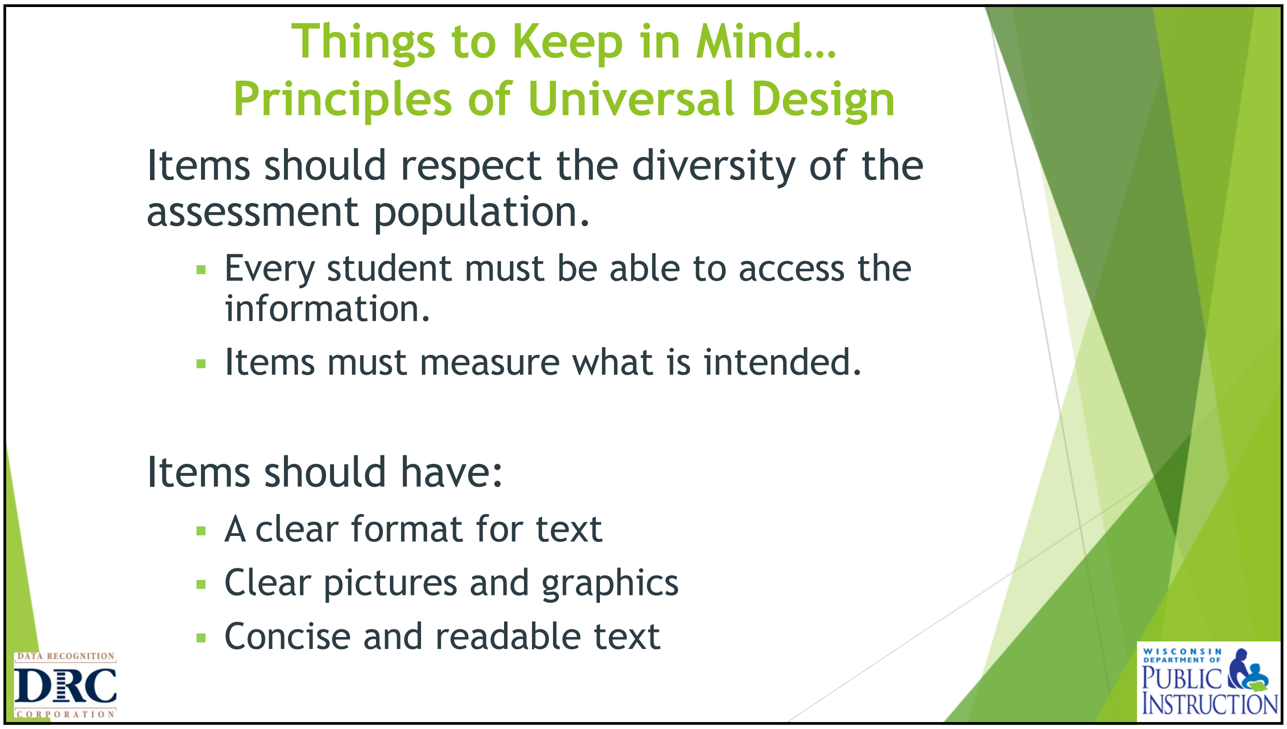
34



Things to Keep in Mind…Technical Design

- Items should not be confusing or tricky
- Does the item meet requirements for technical quality?
- Do graphics/visuals compliment and support item?
- Does the stem provide a complete, clear and concise question/problem and directions?
- Does the stem not clue the correct answer(s)?
- Are correct answer(s) clear and accurate?
  - Distractors (or incorrect options) may contain common misperceptions or processes

35



Things to Keep in Mind… Principles of Universal Design

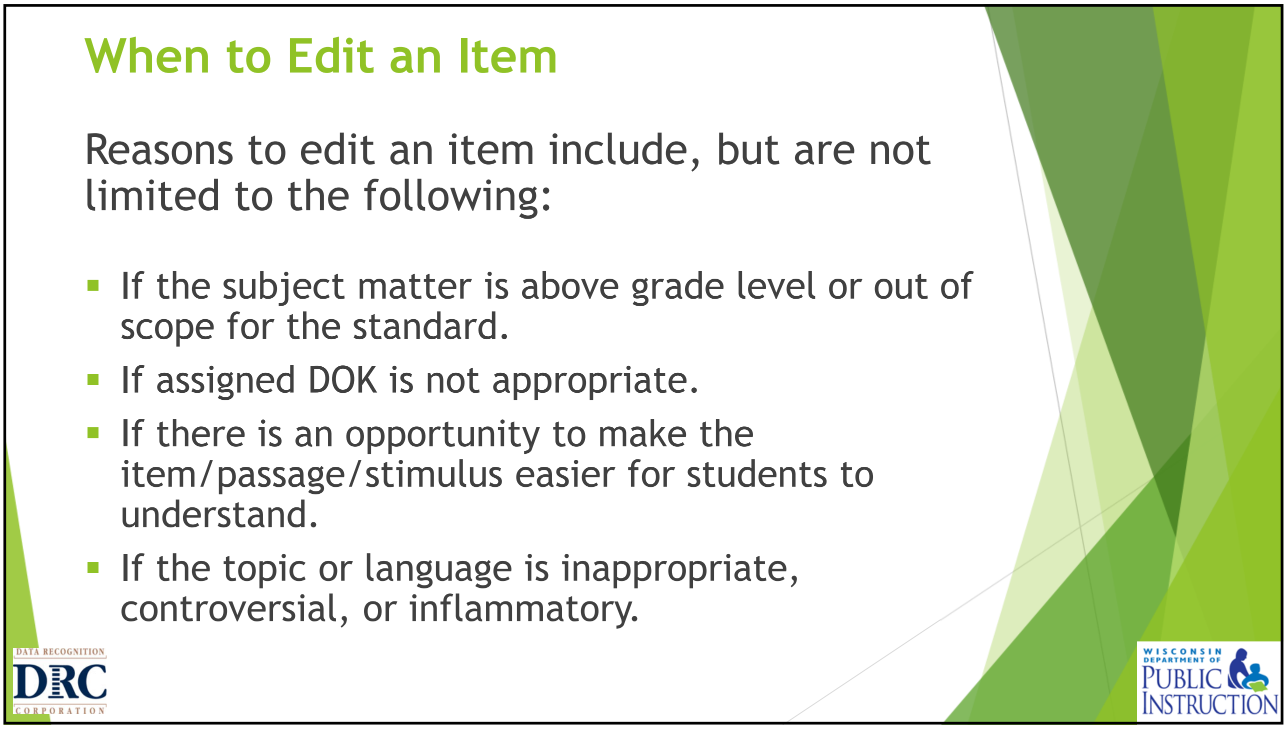
Items should respect the diversity of the assessment population.

- Every student must be able to access the information.
- Items must measure what is intended.

Items should have:

- A clear format for text
- Clear pictures and graphics
- Concise and readable text

36

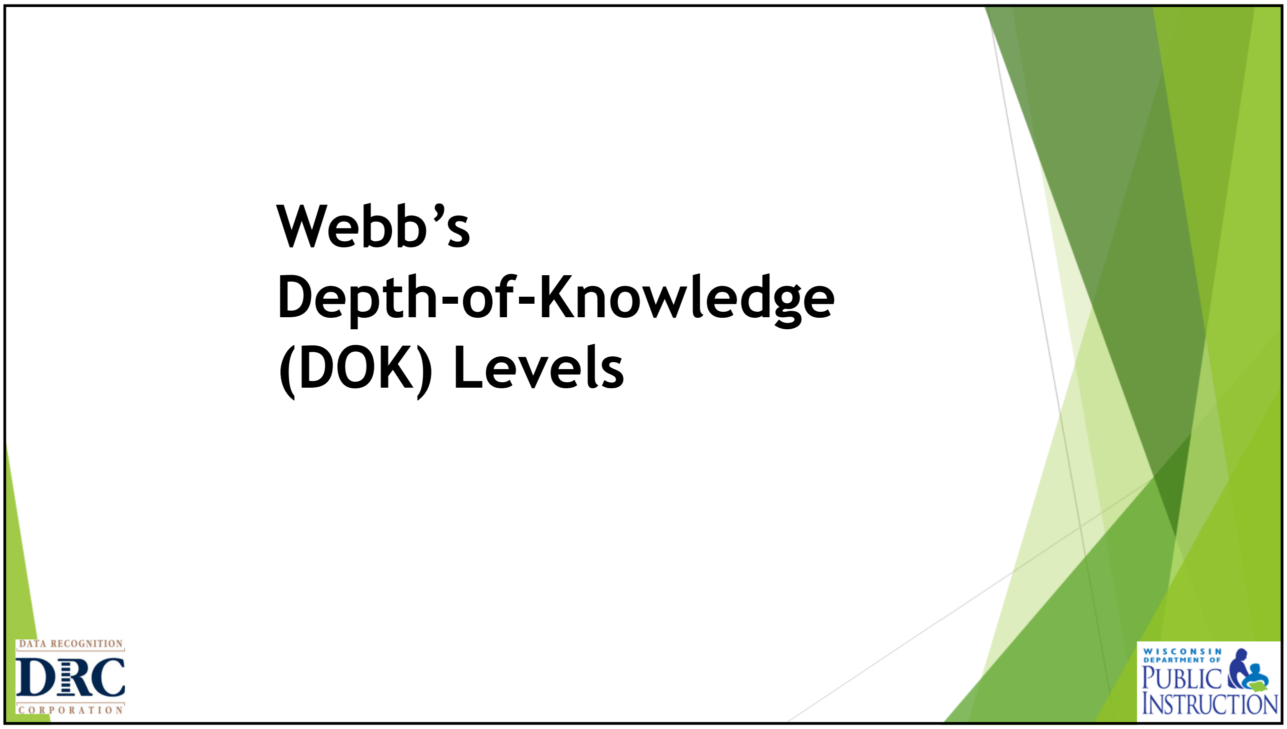


When to Edit an Item

Reasons to edit an item include, but are not limited to the following:

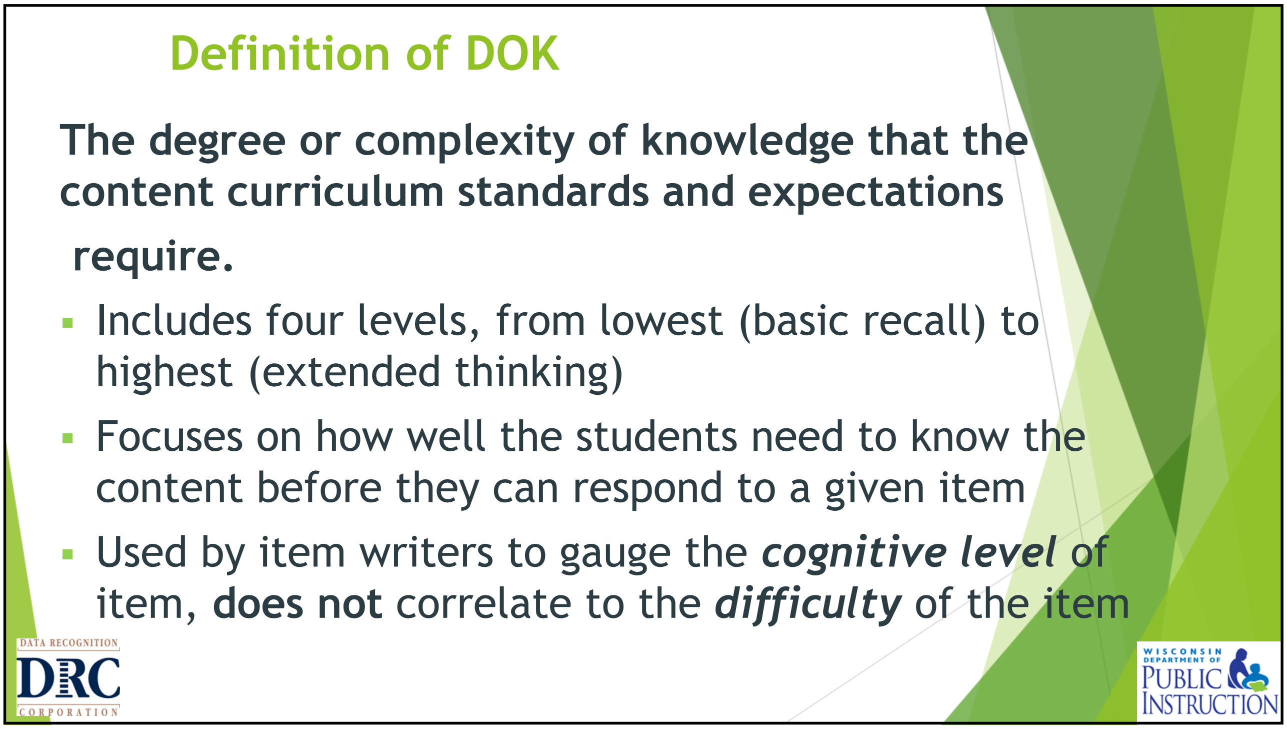
- If the subject matter is above grade level or out of scope for the standard.
- If assigned DOK is not appropriate.
- If there is an opportunity to make the item/passage/stimulus easier for students to understand.
- If the topic or language is inappropriate, controversial, or inflammatory.

37



Webb's Depth-of-Knowledge (DOK) Levels

38

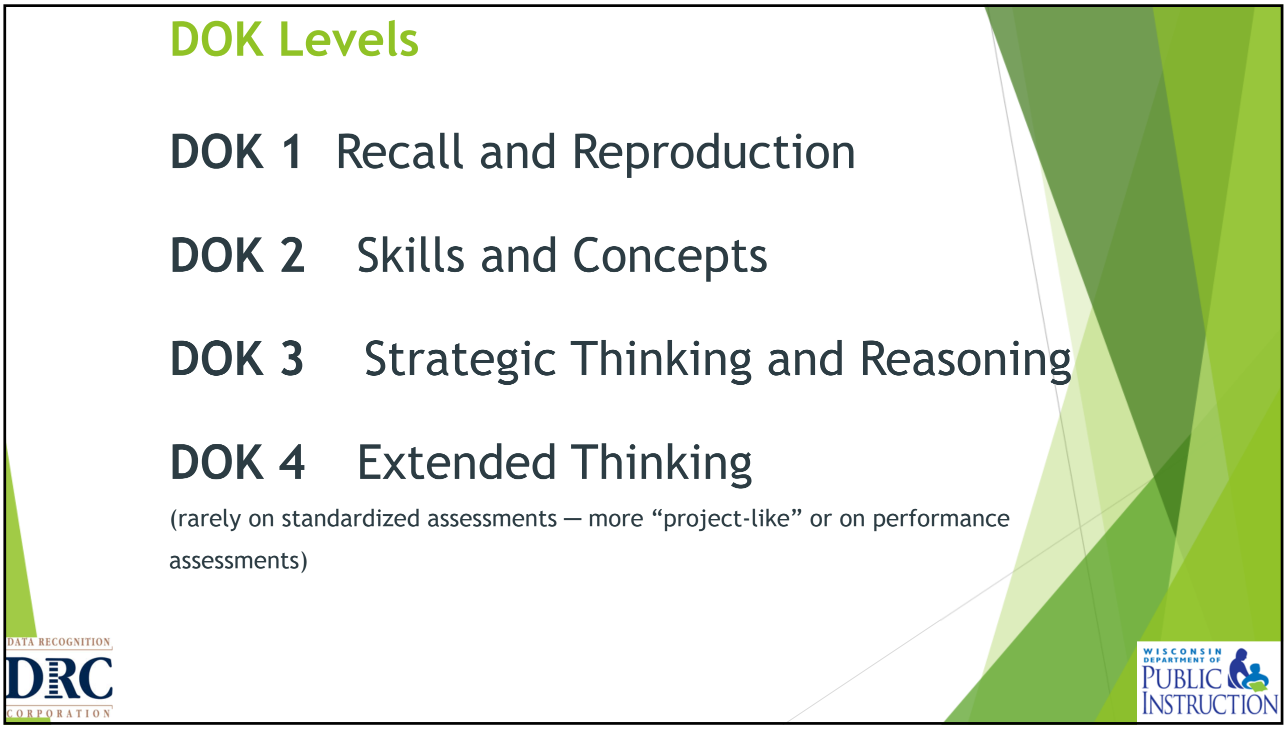


Definition of DOK

The degree or complexity of knowledge that the content curriculum standards and expectations require.

- Includes four levels, from lowest (basic recall) to highest (extended thinking)
- Focuses on how well the students need to know the content before they can respond to a given item
- Used by item writers to gauge the ***cognitive level*** of item, **does not** correlate to the ***difficulty*** of the item

39



DOK Levels

DOK 1 Recall and Reproduction

DOK 2 Skills and Concepts

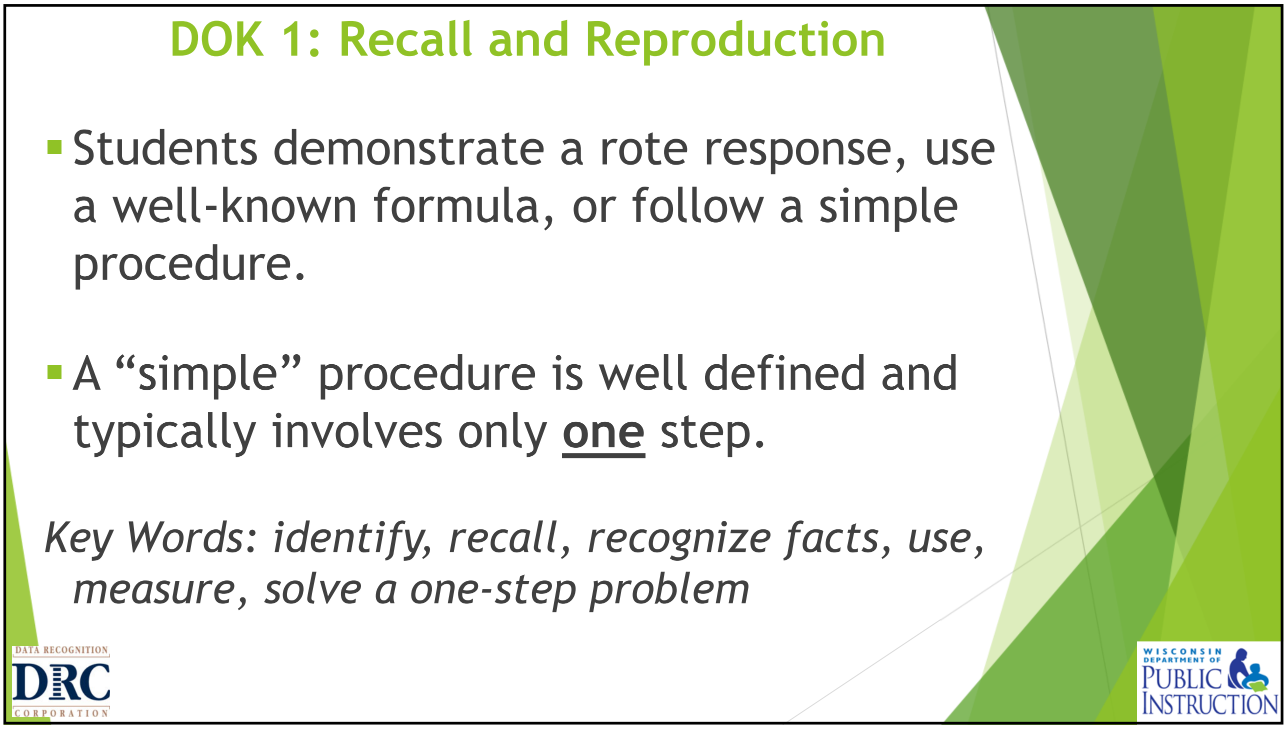
DOK 3 Strategic Thinking and Reasoning

DOK 4 Extended Thinking

(rarely on standardized assessments — more "project-like" or on performance assessments)

DATA RECOGNITION
DRC
CORPORATION
WISCONSIN DEPARTMENT OF
PUBLIC
INSTRUCTION

40

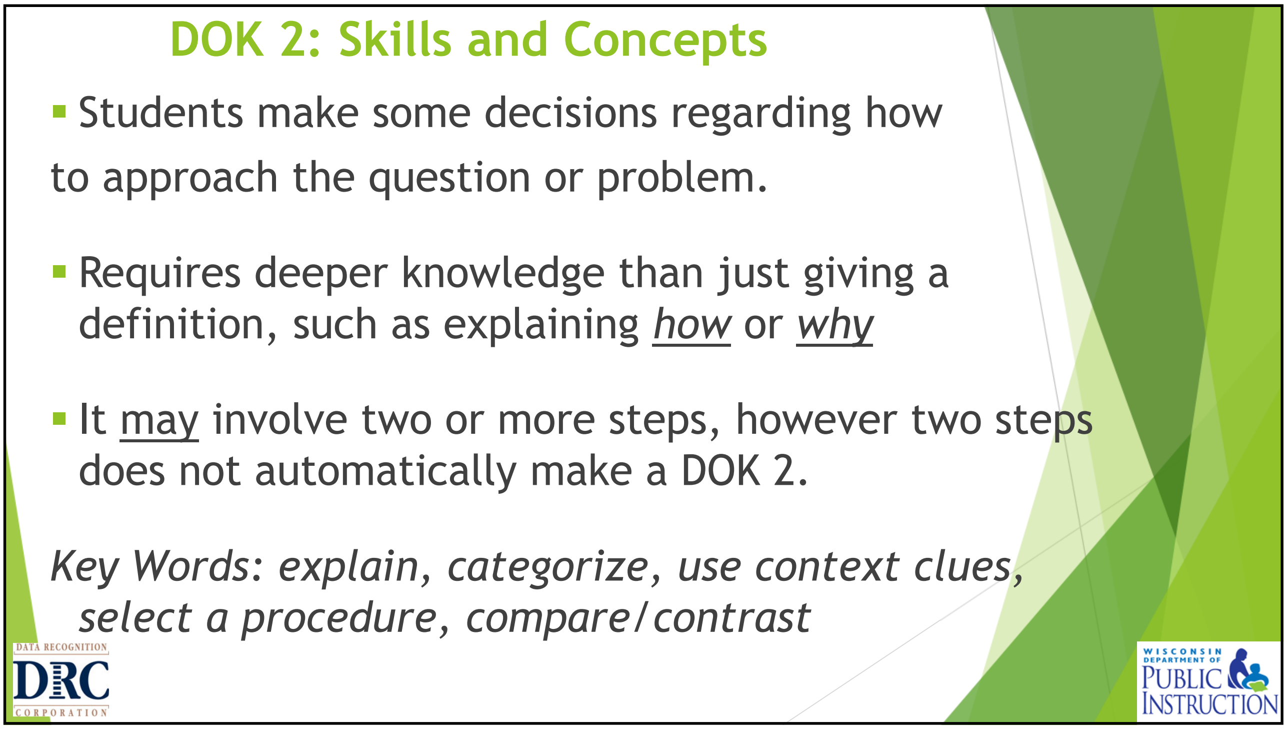


DOK 1: Recall and Reproduction

- Students demonstrate a rote response, use a well-known formula, or follow a simple procedure.
- A “simple” procedure is well defined and typically involves only **one** step.

Key Words: identify, recall, recognize facts, use, measure, solve a one-step problem

41

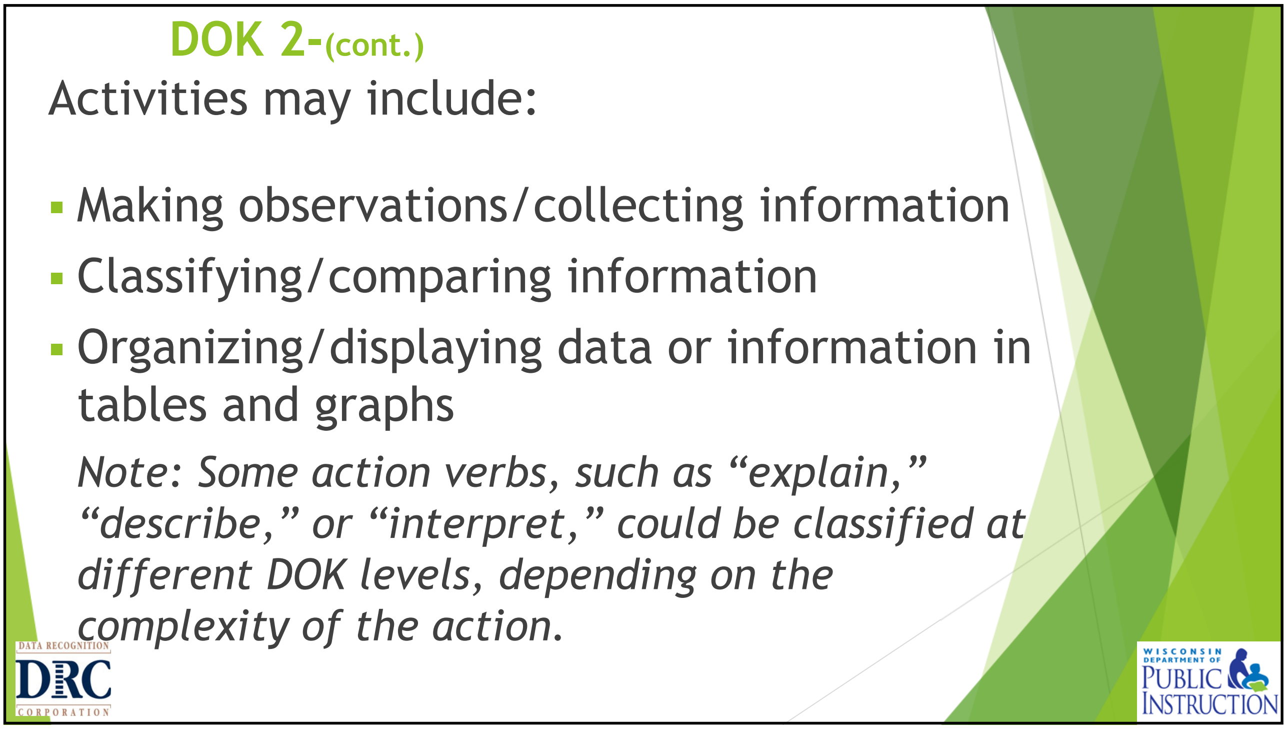


DOK 2: Skills and Concepts

- Students make some decisions regarding how to approach the question or problem.
- Requires deeper knowledge than just giving a definition, such as explaining *how* or *why*
- It may involve two or more steps, however two steps does not automatically make a DOK 2.

Key Words: explain, categorize, use context clues, select a procedure, compare/contrast

42



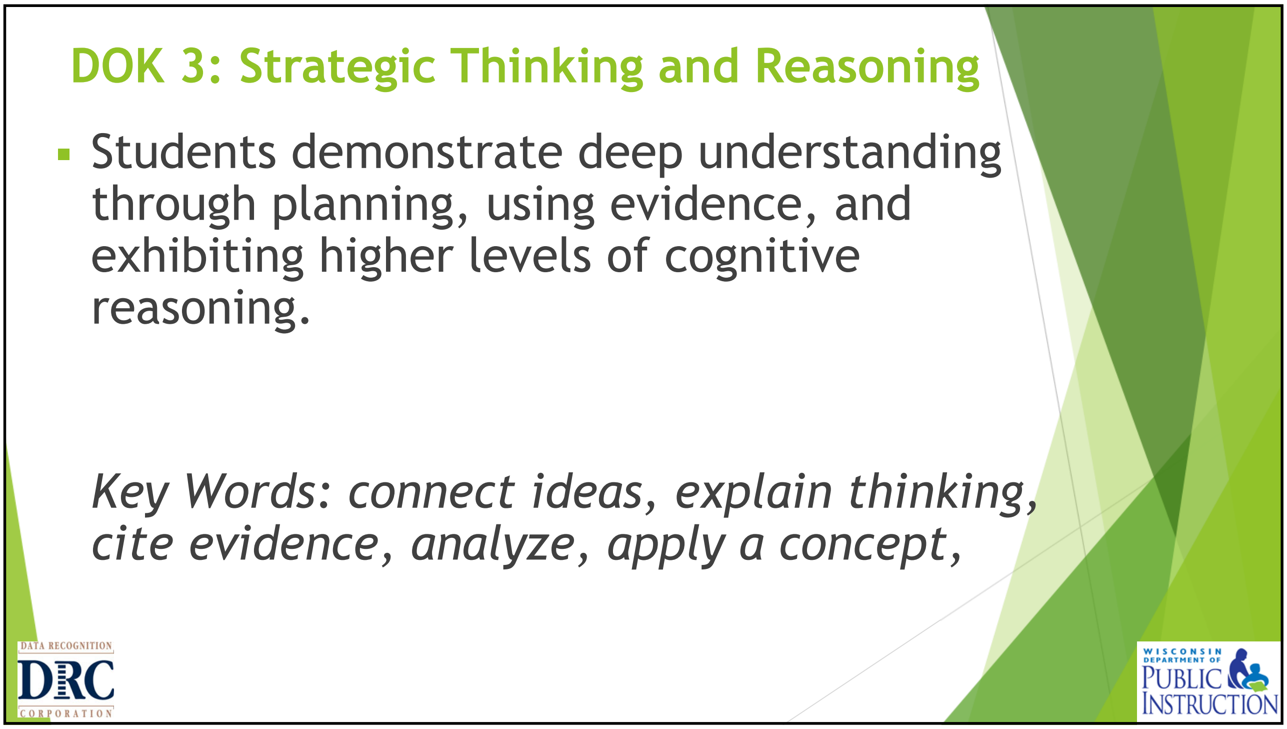
DOK 2-(cont.)

Activities may include:

- Making observations/collecting information
- Classifying/comparing information
- Organizing/displaying data or information in tables and graphs

Note: Some action verbs, such as "explain," "describe," or "interpret," could be classified at different DOK levels, depending on the complexity of the action.

43

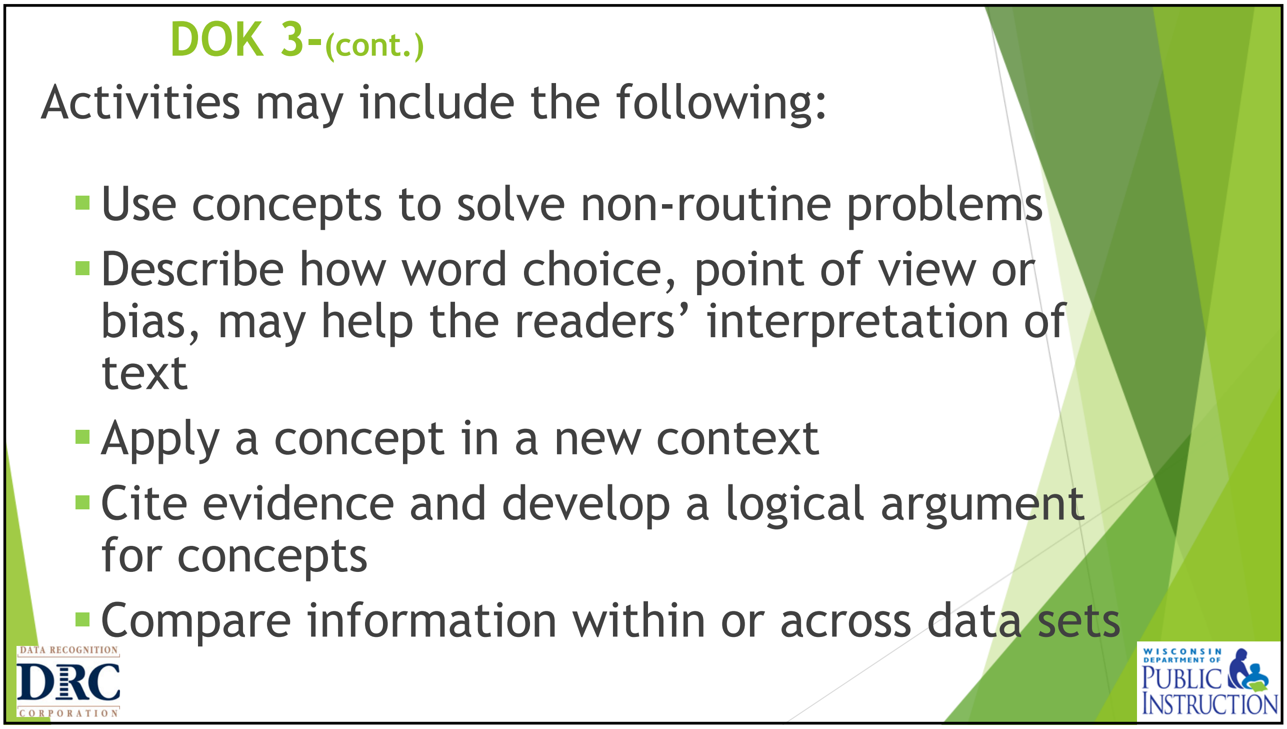


DOK 3: Strategic Thinking and Reasoning

- Students demonstrate deep understanding through planning, using evidence, and exhibiting higher levels of cognitive reasoning.

Key Words: connect ideas, explain thinking, cite evidence, analyze, apply a concept,

44

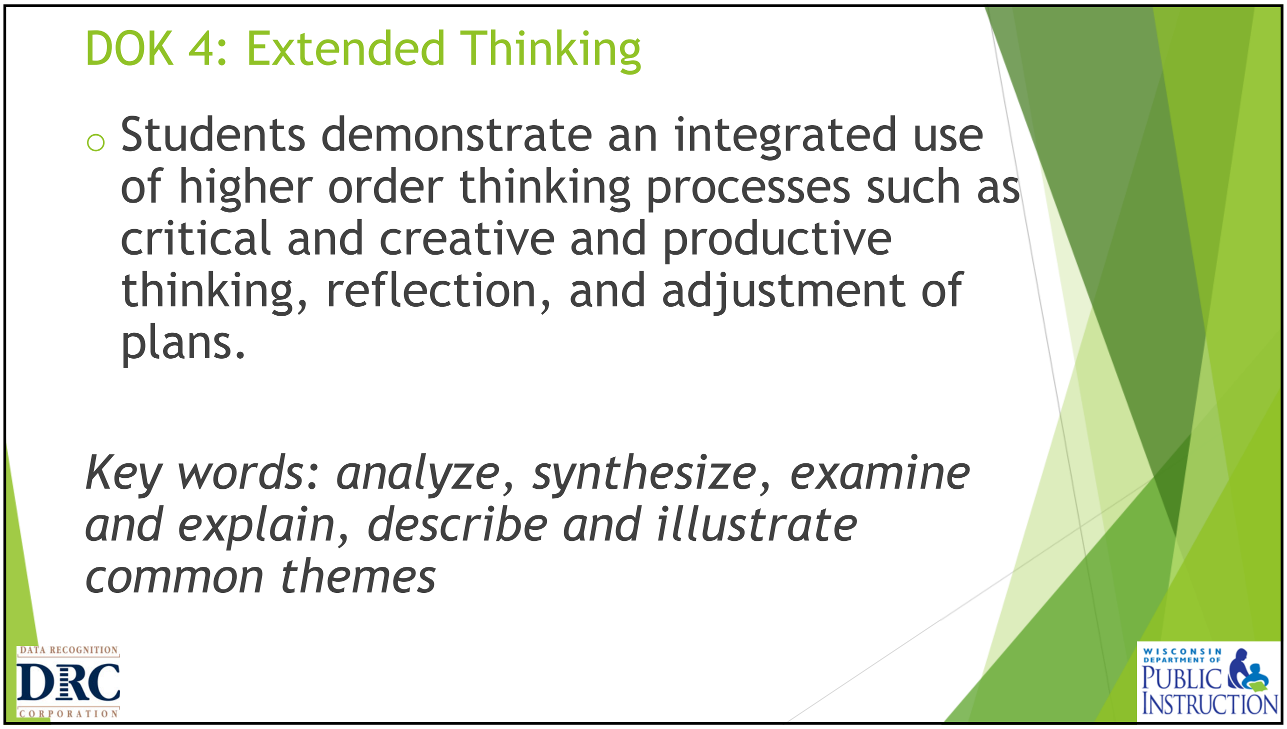


DOK 3-(cont.)

Activities may include the following:

- Use concepts to solve non-routine problems
- Describe how word choice, point of view or bias, may help the readers' interpretation of text
- Apply a concept in a new context
- Cite evidence and develop a logical argument for concepts
- Compare information within or across data sets

45

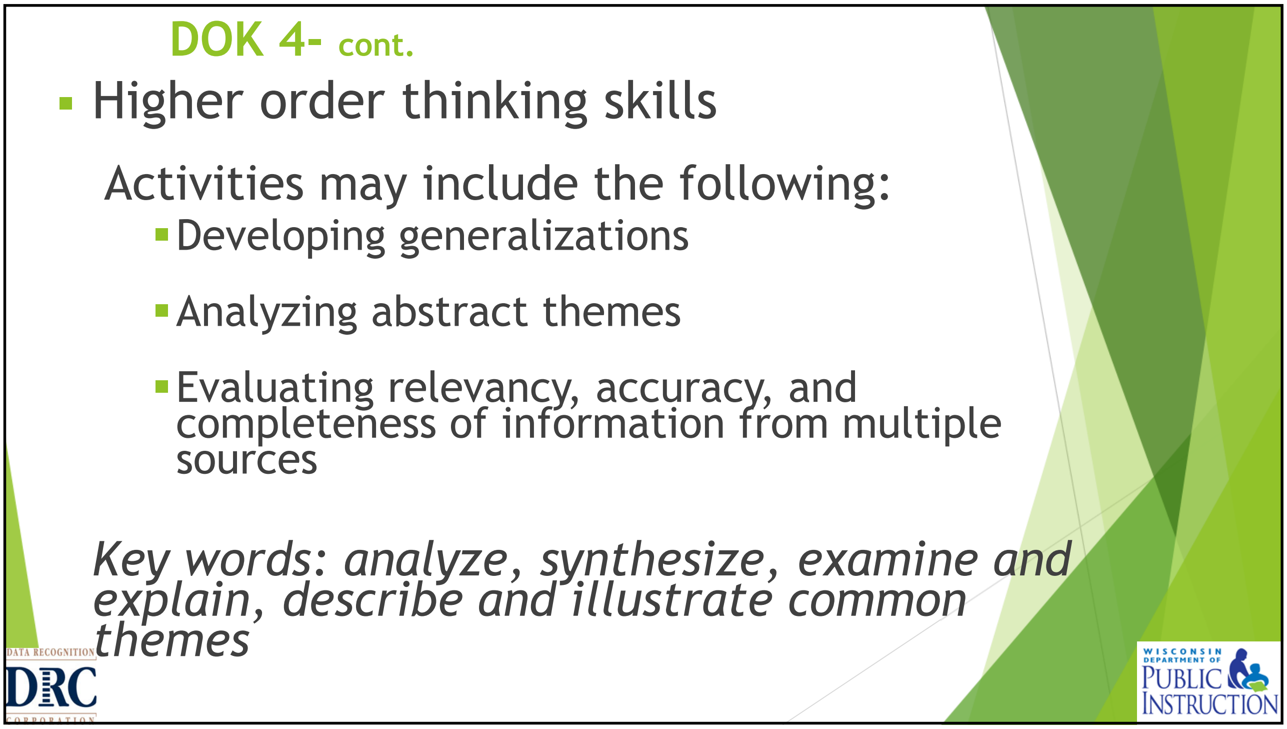


DOK 4: Extended Thinking

- Students demonstrate an integrated use of higher order thinking processes such as critical and creative and productive thinking, reflection, and adjustment of plans.

Key words: analyze, synthesize, examine and explain, describe and illustrate common themes

46



DOK 4- cont.

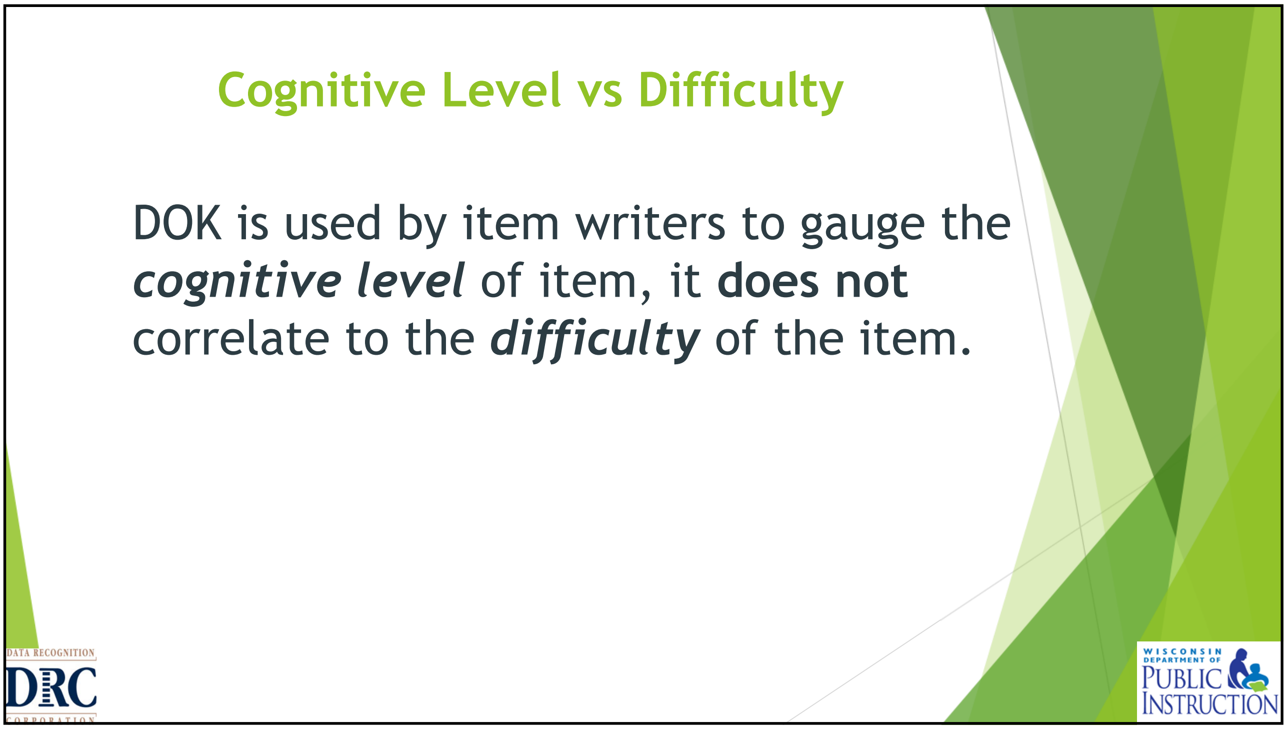
- Higher order thinking skills

Activities may include the following:

- Developing generalizations
- Analyzing abstract themes
- Evaluating relevancy, accuracy, and completeness of information from multiple sources

Key words: analyze, synthesize, examine and explain, describe and illustrate common themes

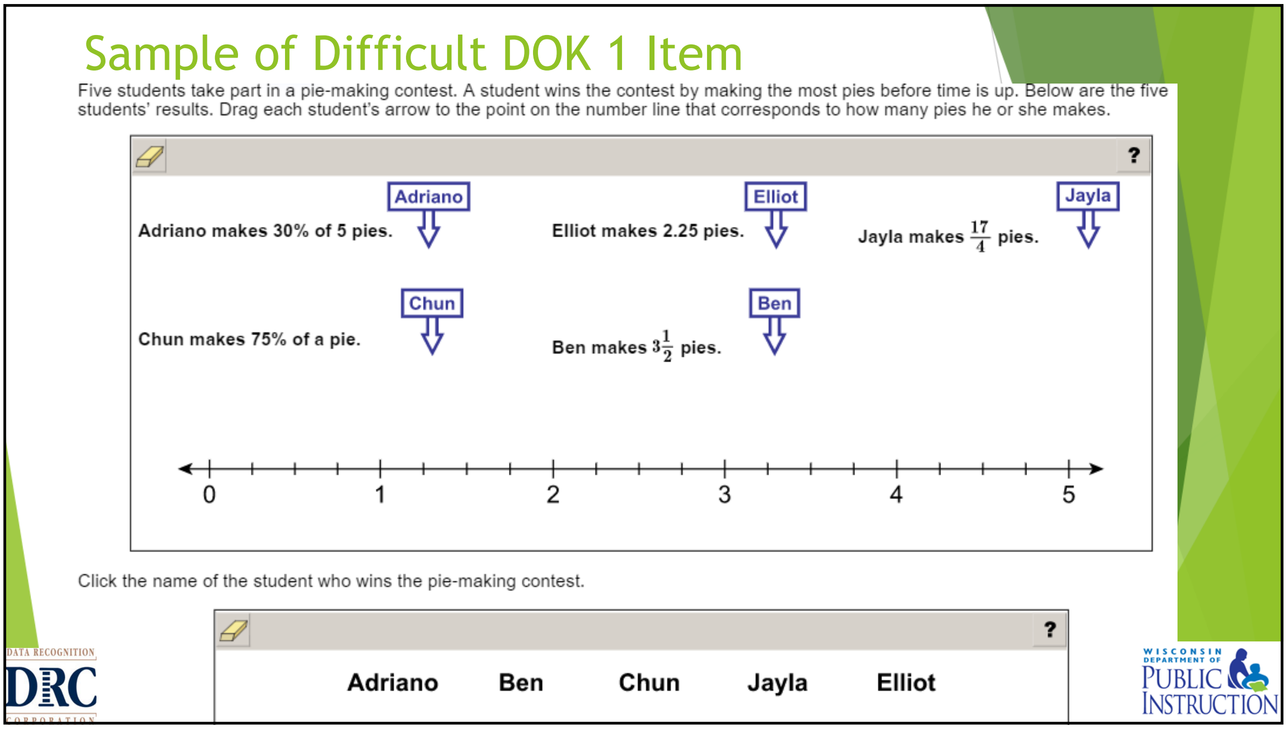
47



Cognitive Level vs Difficulty

DOK is used by item writers to gauge the ***cognitive level*** of item, it **does not** correlate to the ***difficulty*** of the item.

48



Sample of Difficult DOK 1 Item

Five students take part in a pie-making contest. A student wins the contest by making the most pies before time is up. Below are the five students' results. Drag each student's arrow to the point on the number line that corresponds to how many pies he or she makes.

?

Adriano makes 30% of 5 pies.

Elliot makes 2.25 pies.

Jayla makes $\frac{17}{4}$ pies.

Chun makes 75% of a pie.

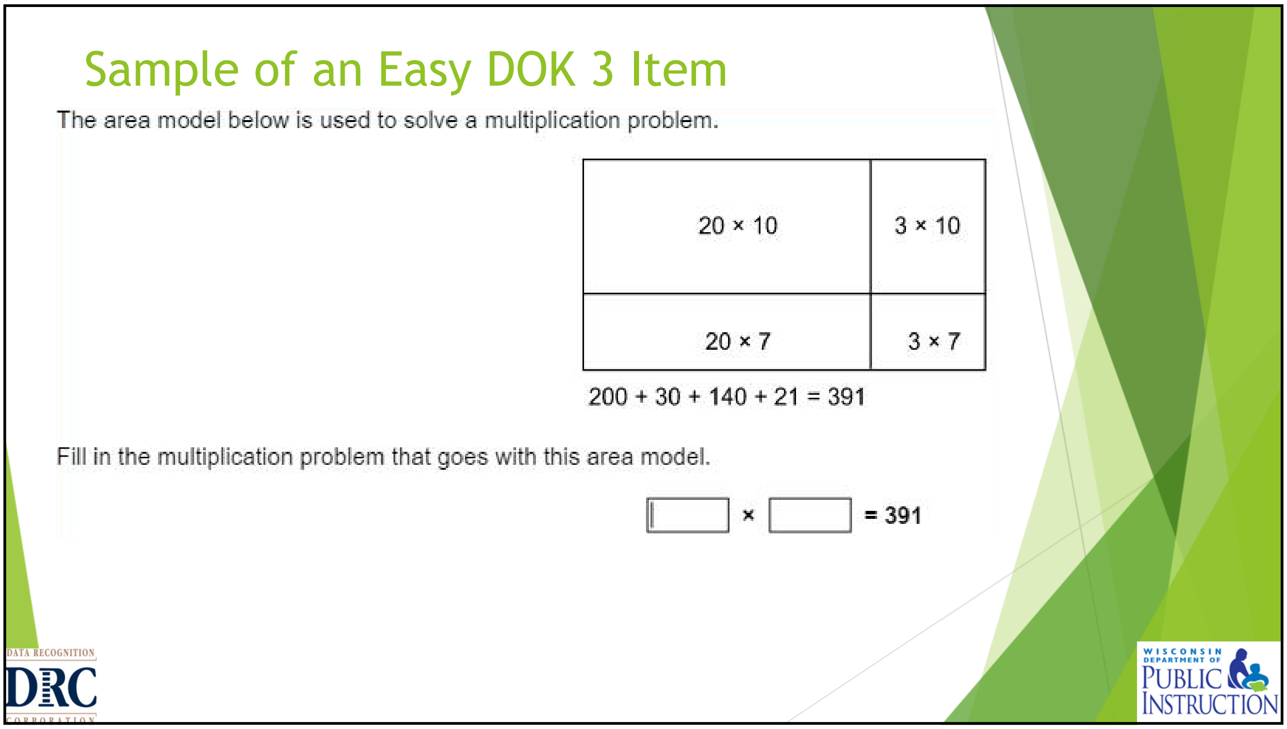
Ben makes $3\frac{1}{2}$ pies.

Click the name of the student who wins the pie-making contest.

?

Adriano Ben Chun Jayla Elliot

49



Sample of an Easy DOK 3 Item

The area model below is used to solve a multiplication problem.

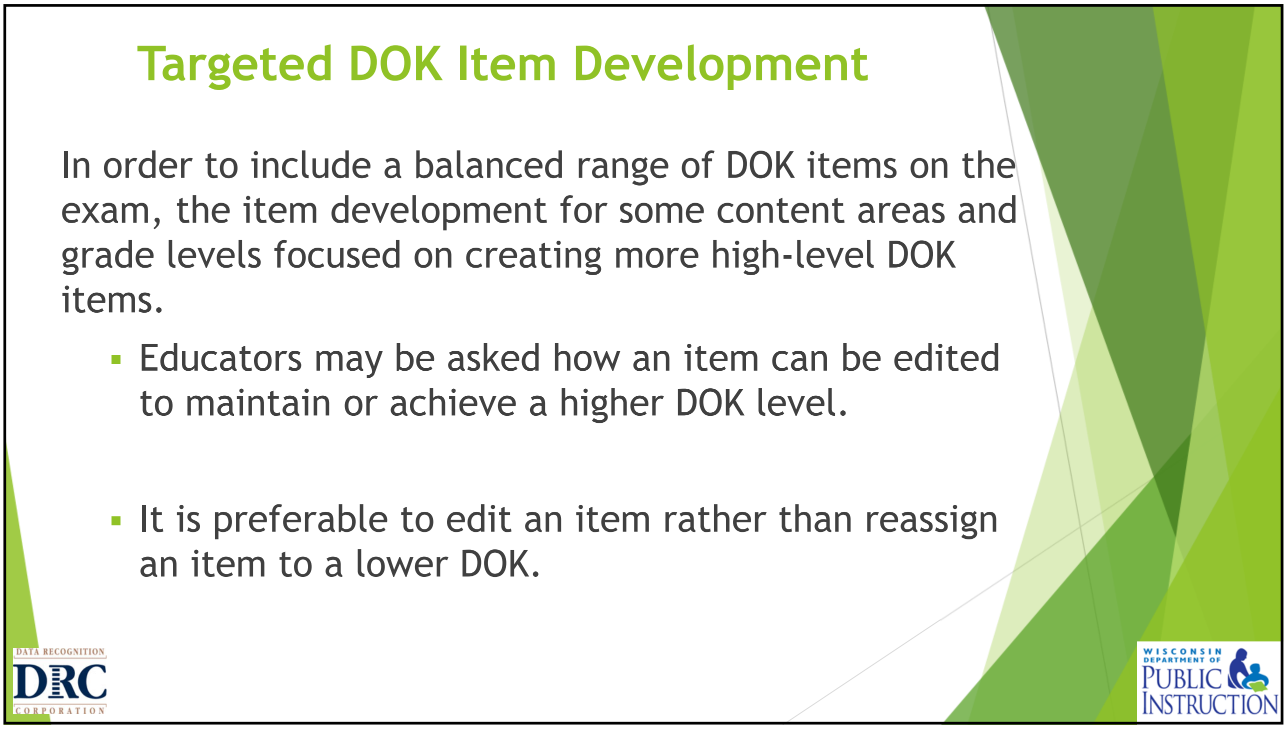
20 × 10	3 × 10
20 × 7	3 × 7

200 + 30 + 140 + 21 = 391

Fill in the multiplication problem that goes with this area model.

□ × □ = 391

50

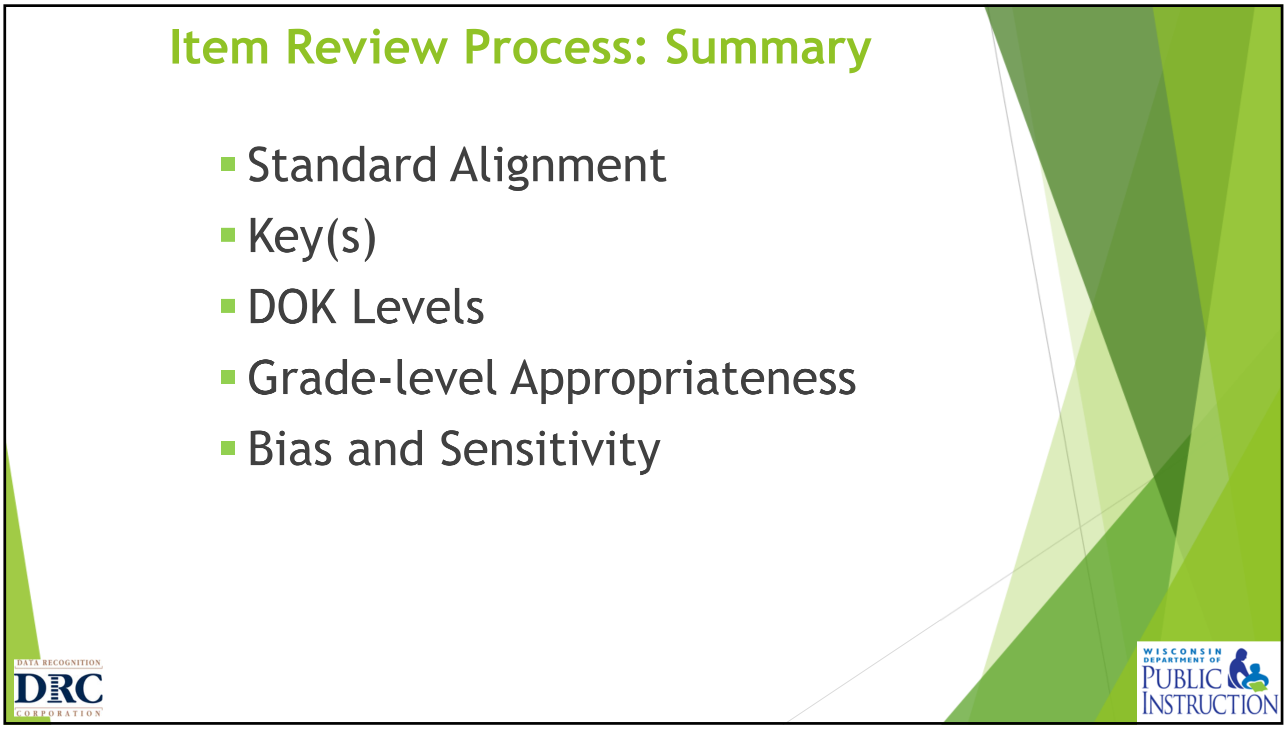


Targeted DOK Item Development

In order to include a balanced range of DOK items on the exam, the item development for some content areas and grade levels focused on creating more high-level DOK items.

- Educators may be asked how an item can be edited to maintain or achieve a higher DOK level.
- It is preferable to edit an item rather than reassign an item to a lower DOK.

51

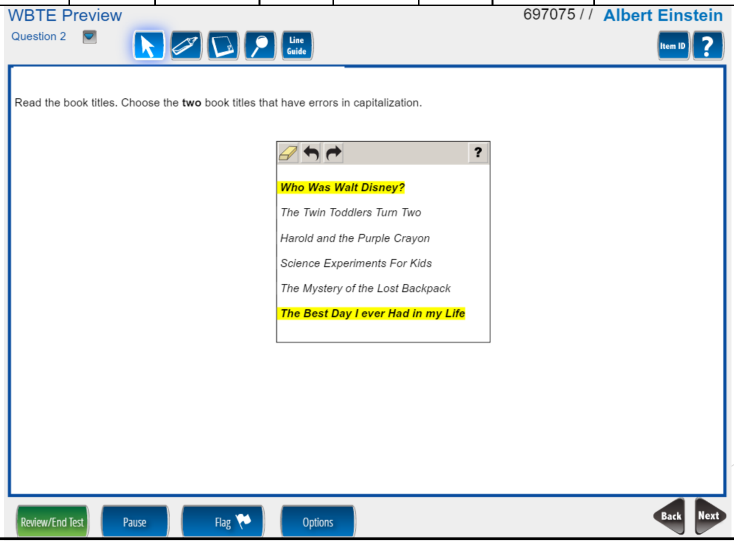


Item Review Process: Summary

- Standard Alignment
- Key(s)
- DOK Levels
- Grade-level Appropriateness
- Bias and Sensitivity

52

Session #	Seq #	Item ID	Passage Title	Standard Code	Item Type	Points	Key(s)	Proposed DOK	Bias/Sensitivity Comments	Accept (A) Accept with Revisions (AR) Dissenting View (DV)	Comments	Item Preference



697075 / / Albert Einstein

Read the book titles. Choose the **two** book titles that have errors in capitalization.

Who Was Walt Disney?

The Twin Toddlers Turn Two

Harold and the Purple Crayon

Science Experiments For Kids

The Mystery of the Lost Backpack

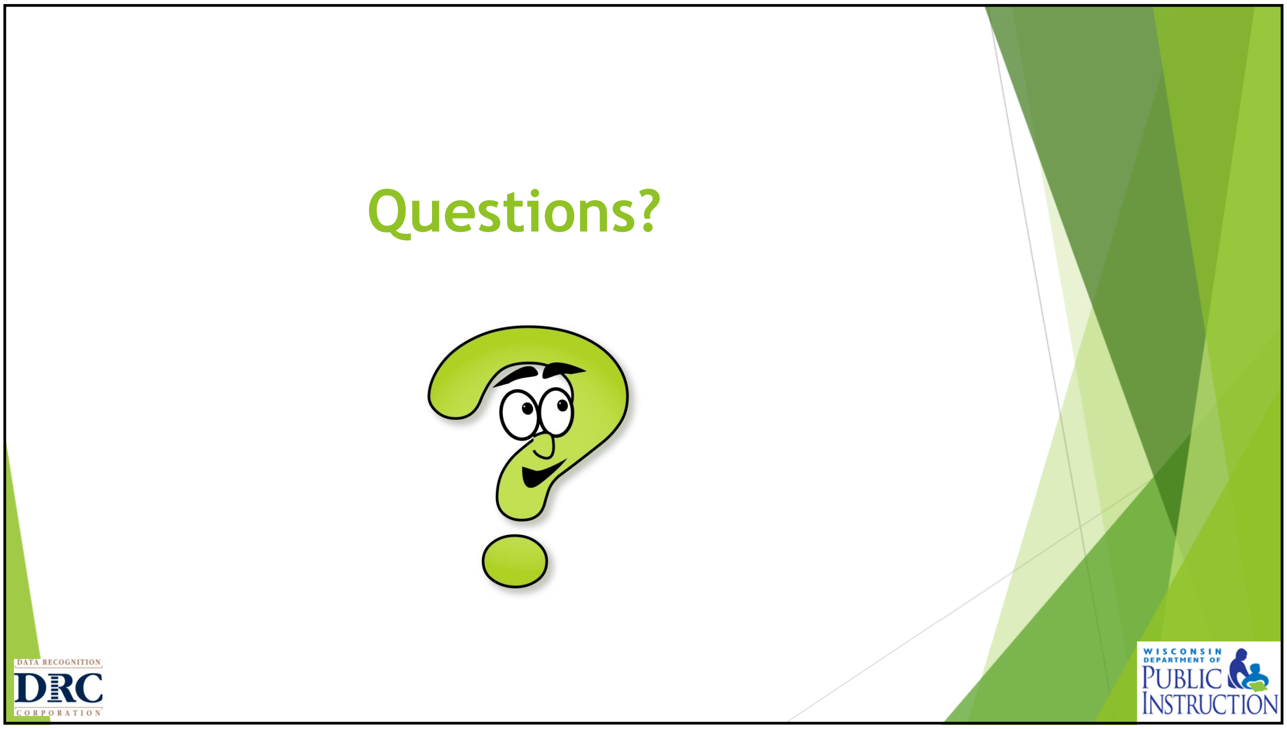
The Best Day I ever Had in my Life

53

Roles & Responsibilities: Summary

- Invest yourself in the process
- Share your opinions
- Listen to your colleagues

54

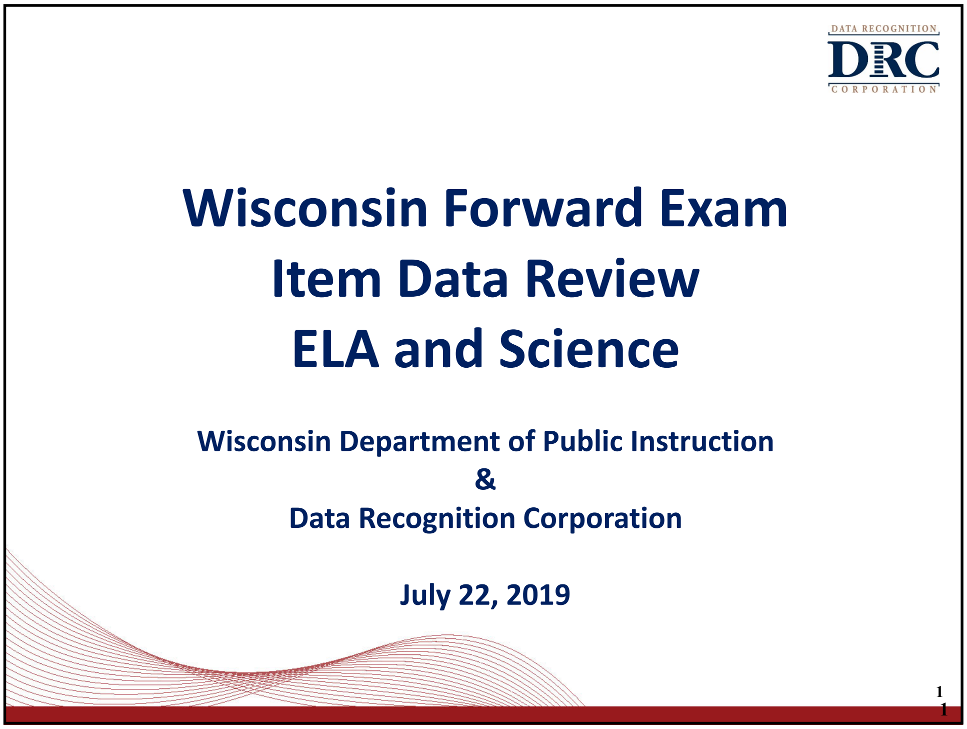


Questions?

55

Appendix B

Summer 2019 Field Test Data Review Training Slides



DATA RECOGNITION DRC CORPORATION

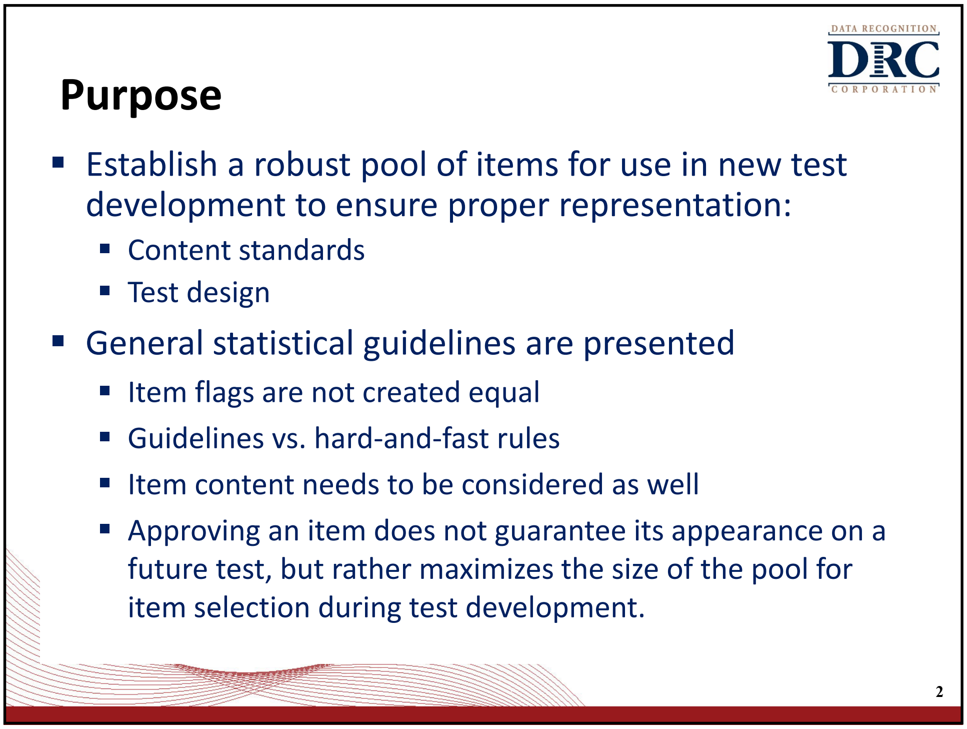
Wisconsin Forward Exam Item Data Review ELA and Science

Wisconsin Department of Public Instruction
&
Data Recognition Corporation

July 22, 2019

1
1

1



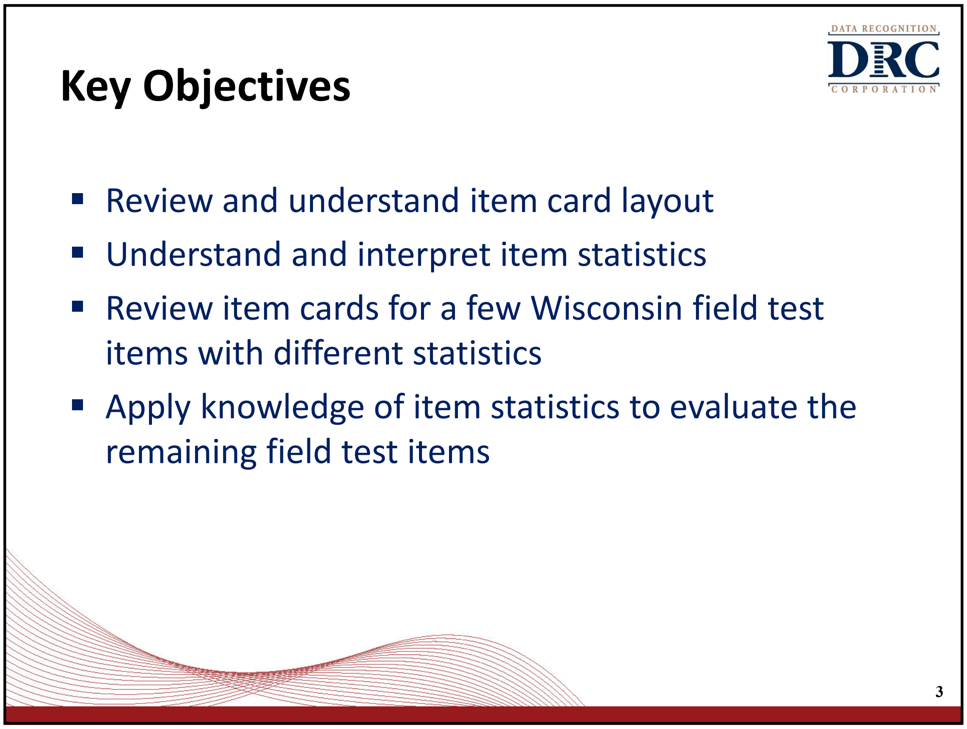
DATA RECOGNITION DRC CORPORATION

Purpose

- Establish a robust pool of items for use in new test development to ensure proper representation:
  - Content standards
  - Test design
- General statistical guidelines are presented
  - Item flags are not created equal
  - Guidelines vs. hard-and-fast rules
  - Item content needs to be considered as well
  - Approving an item does not guarantee its appearance on a future test, but rather maximizes the size of the pool for item selection during test development.

2

2

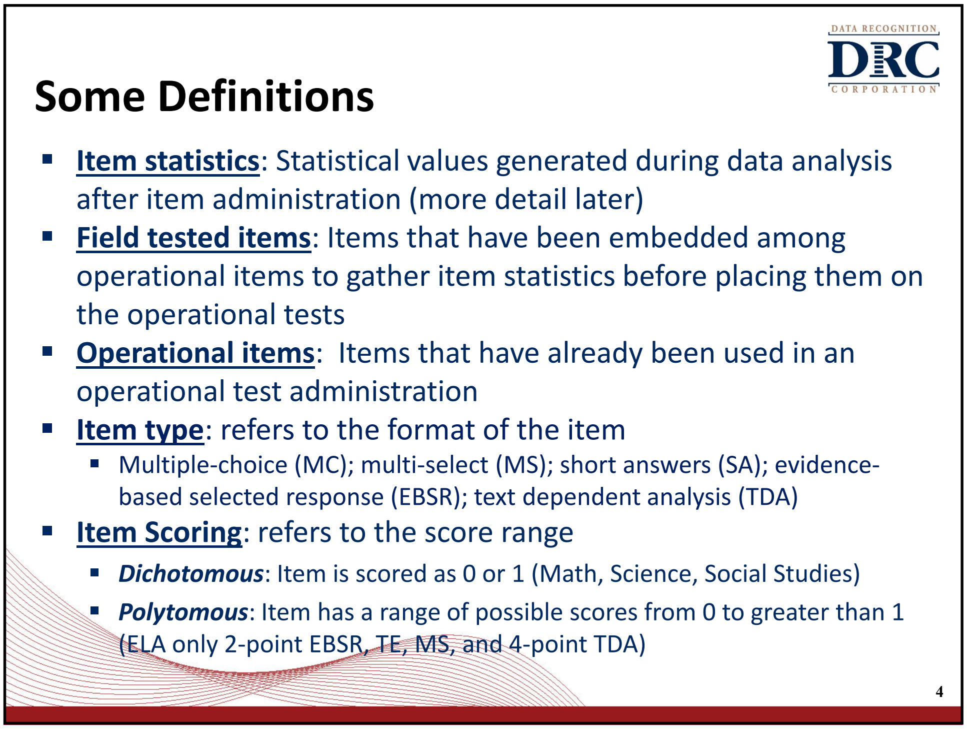


Key Objectives

- Review and understand item card layout
- Understand and interpret item statistics
- Review item cards for a few Wisconsin field test items with different statistics
- Apply knowledge of item statistics to evaluate the remaining field test items

3

3

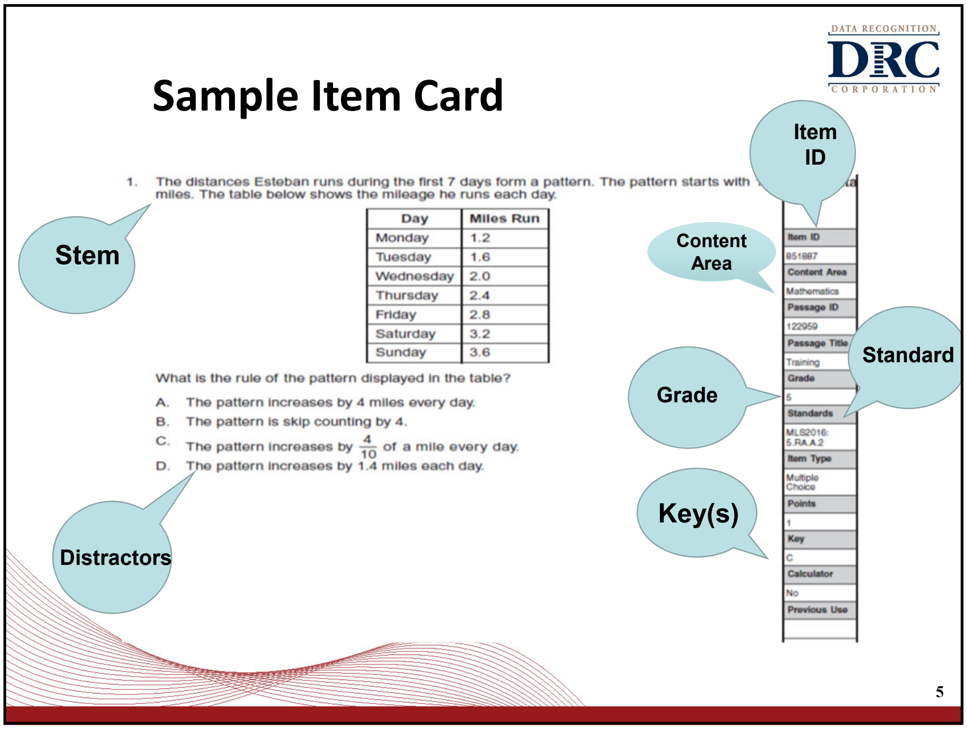


Some Definitions

- **Item statistics**: Statistical values generated during data analysis after item administration (more detail later)
- **Field tested items**: Items that have been embedded among operational items to gather item statistics before placing them on the operational tests
- **Operational items**: Items that have already been used in an operational test administration
- **Item type**: refers to the format of the item
  - Multiple-choice (MC); multi-select (MS); short answers (SA); evidence-based selected response (EBSR); text dependent analysis (TDA)
- **Item Scoring**: refers to the score range
  - ***Dichotomous***: Item is scored as 0 or 1 (Math, Science, Social Studies)
  - ***Polytomous***: Item has a range of possible scores from 0 to greater than 1 (ELA only 2-point EBSR, TE, MS, and 4-point TDA)

4

4



DATA RECOGNITION DRC CORPORATION

Sample Item Card

1. The distances Esteban runs during the first 7 days form a pattern. The pattern starts with
miles. The table below shows the mileage he runs each day.

Day	Miles Run
Monday	1.2
Tuesday	1.6
Wednesday	2.0
Thursday	2.4
Friday	2.8
Saturday	3.2
Sunday	3.6

What is the rule of the pattern displayed in the table?

A. The pattern increases by 4 miles every day.
B. The pattern is skip counting by 4.
C. The pattern increases by $\frac{4}{10}$ of a mile every day.
D. The pattern increases by 1.4 miles each day.

Item ID	851887
Content Area	Mathematics
Passage ID	122959
Passage Title	Training
Grade	5
Standards	MLS2016: 5.RA.A.2
Item Type	Multiple Choice
Points	1
Key	C
Calculator	No
Previous Use	

Stem

Distractors

Item ID

Content Area

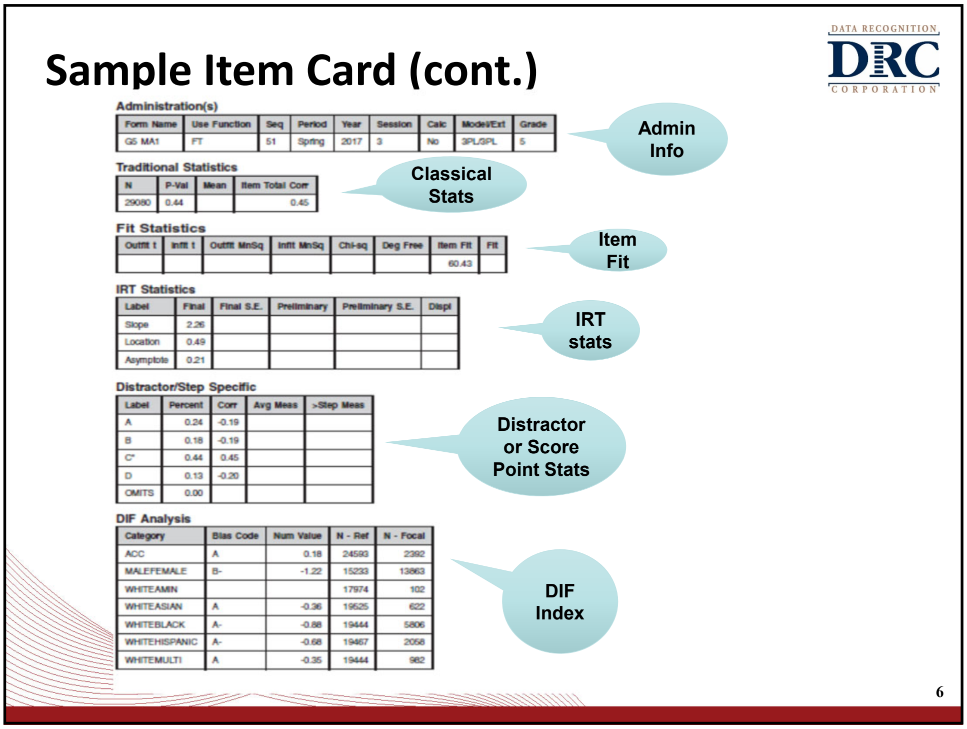
Grade

Standard

Key(s)

5

5



DATA RECOGNITION DRC CORPORATION

Sample Item Card (cont.)

Administration(s)

Form Name	Use Function	Seq	Period	Year	Session	Calc	Model/Ext	Grade
G5 MA1	FT	51	Spring	2017	3	No	3PL/3PL	5

Traditional Statistics

N	P-Val	Mean	Item Total Corr
29080	0.44		0.45

Fit Statistics

Outfit t	Infit t	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Item Fit	Fit
						60.43	

IRT Statistics

Label	Final	Final S.E.	Preliminary	Preliminary S.E.	Displ
Slope	2.26				
Location	0.49				
Asymptote	0.21				

Distractor/Step Specific

Label	Percent	Corr	Avg Meas	>Step Meas
A	0.24	-0.19		
B	0.18	-0.19		
C*	0.44	0.45		
D	0.13	-0.20		
OMITS	0.00			

DIF Analysis

Category	Bias Code	Num Value	N - Ref	N - Focal
ACC	A	0.18	24593	2392
MALEFEMALE	B-	-1.22	15233	13863
WHITEAMIN			17974	102
WHITEASIAN	A	-0.36	19525	622
WHITEBLACK	A-	-0.88	19444	5806
WHITEHISPANIC	A-	-0.68	19467	2058
WHITEMULTI	A	-0.35	19444	982

Admin Info

Classical Stats

Item Fit

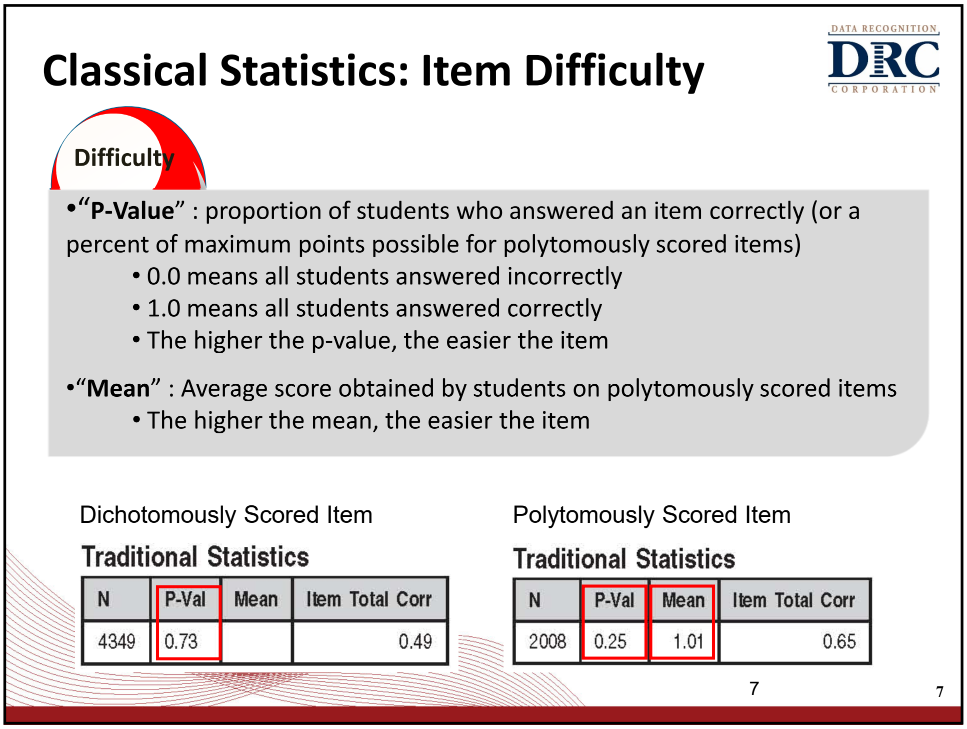
IRT stats

Distractor or Score Point Stats

DIF Index

6

6



Classical Statistics: Item Difficulty

Difficulty

- “**P-Value**” : proportion of students who answered an item correctly (or a percent of maximum points possible for polytomously scored items)
  - 0.0 means all students answered incorrectly
  - 1.0 means all students answered correctly
  - The higher the p-value, the easier the item
- “**Mean**” : Average score obtained by students on polytomously scored items
  - The higher the mean, the easier the item

Dichotomously Scored Item

Traditional Statistics

N	P-Val	Mean	Item Total Corr
4349	0.73		0.49

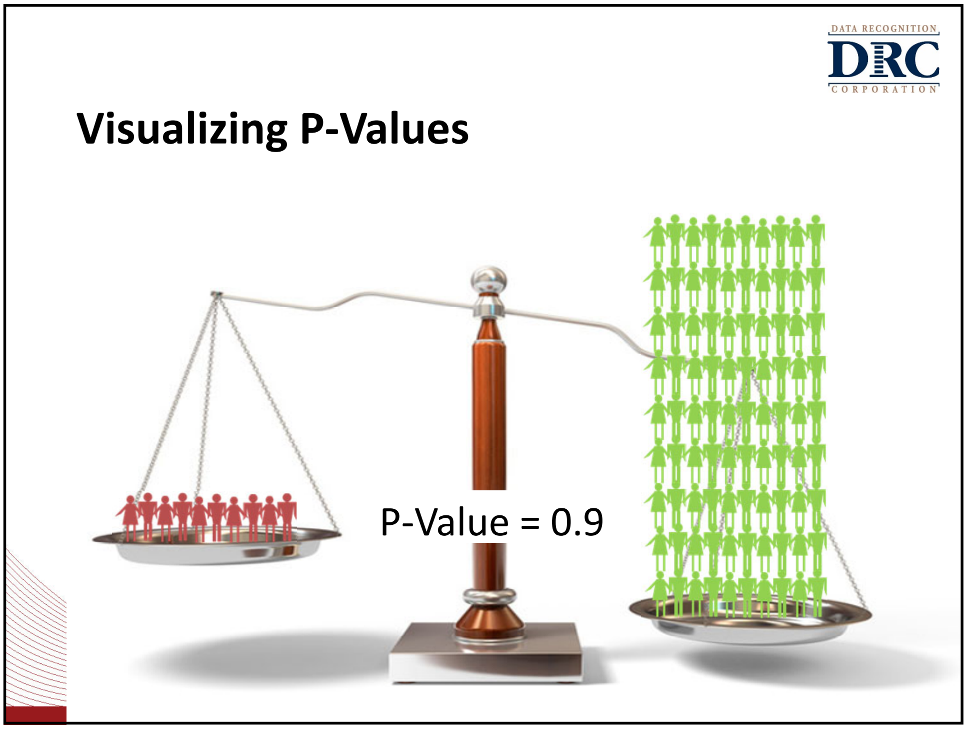
Polytomously Scored Item

Traditional Statistics

N	P-Val	Mean	Item Total Corr
2008	0.25	1.01	0.65

7 7

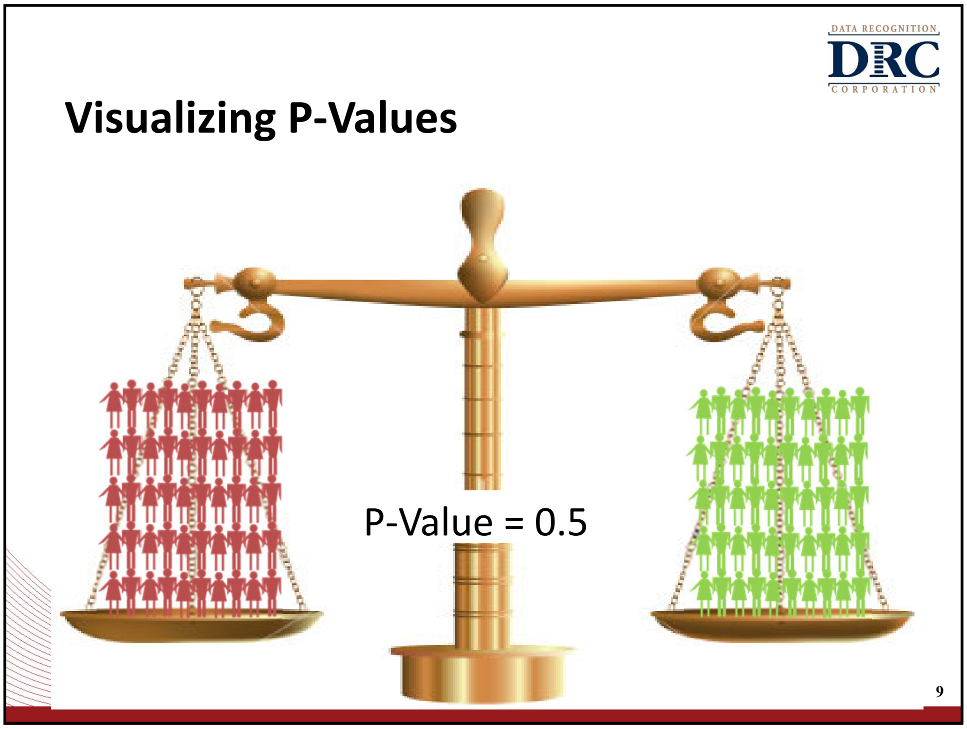
7



Visualizing P-Values

P-Value = 0.9

8

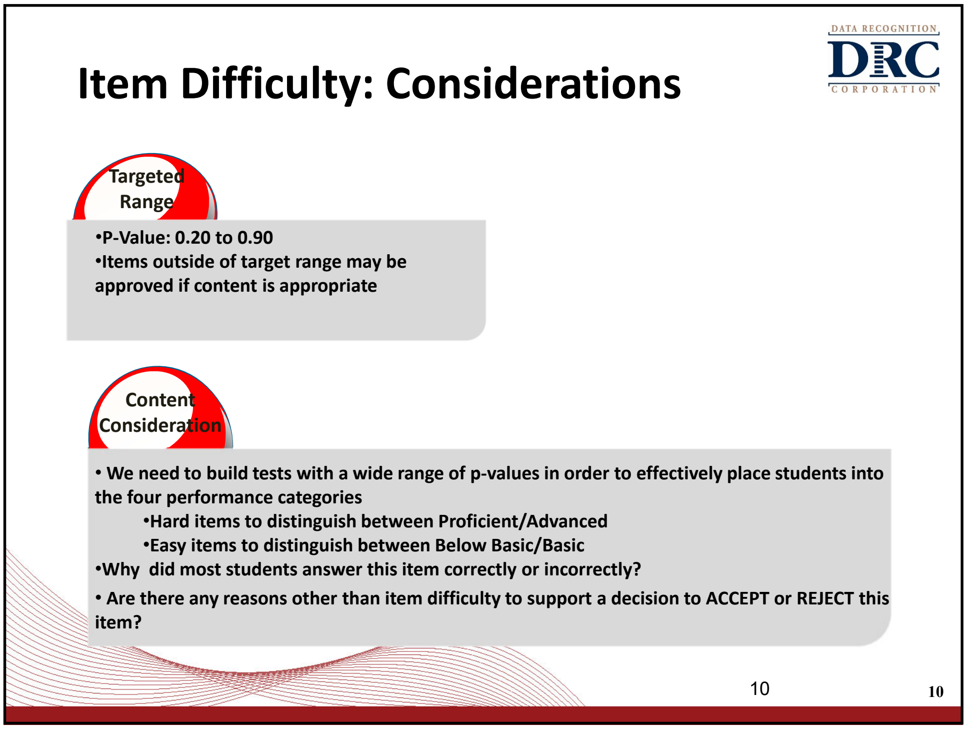


Visualizing P-Values

P-Value = 0.5

9

9



Item Difficulty: Considerations

Targeted Range

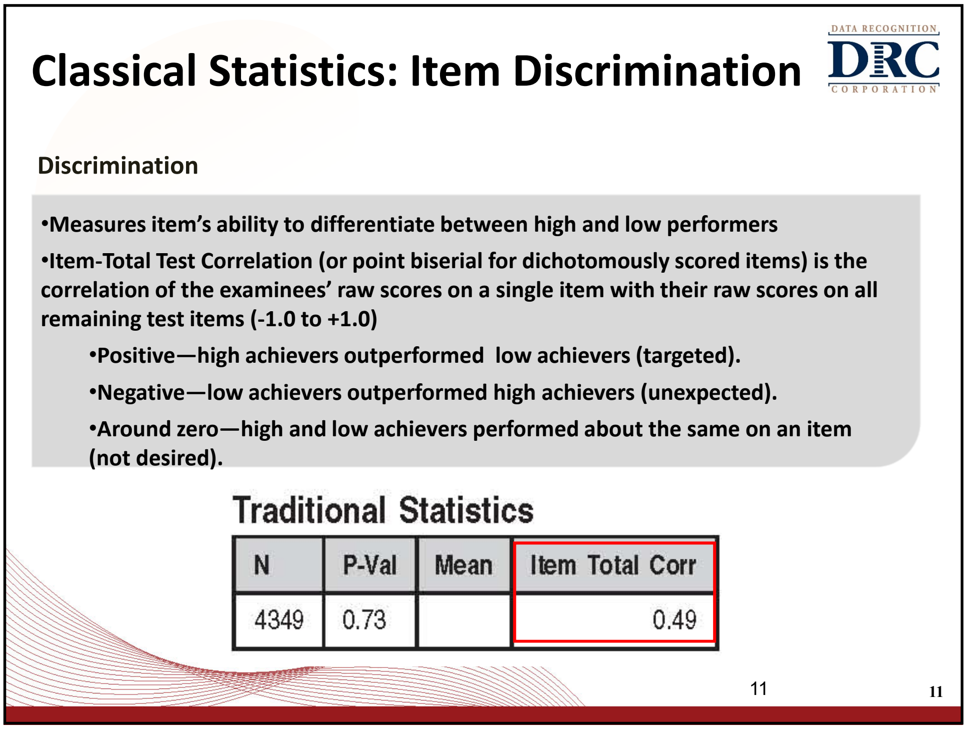
- **P-Value: 0.20 to 0.90**
- **Items outside of target range may be approved if content is appropriate**

Content Consideration

- **We need to build tests with a wide range of p-values in order to effectively place students into the four performance categories**
  - **Hard items to distinguish between Proficient/Advanced**
  - **Easy items to distinguish between Below Basic/Basic**
- **Why did most students answer this item correctly or incorrectly?**
- **Are there any reasons other than item difficulty to support a decision to ACCEPT or REJECT this item?**

10 10

10



Classical Statistics: Item Discrimination

Discrimination

- **Measures item's ability to differentiate between high and low performers**
- **Item-Total Test Correlation (or point biserial for dichotomously scored items) is the correlation of the examinees' raw scores on a single item with their raw scores on all remaining test items (-1.0 to +1.0)**
  - **Positive—high achievers outperformed low achievers (targeted).**
  - **Negative—low achievers outperformed high achievers (unexpected).**
  - **Around zero—high and low achievers performed about the same on an item (not desired).**

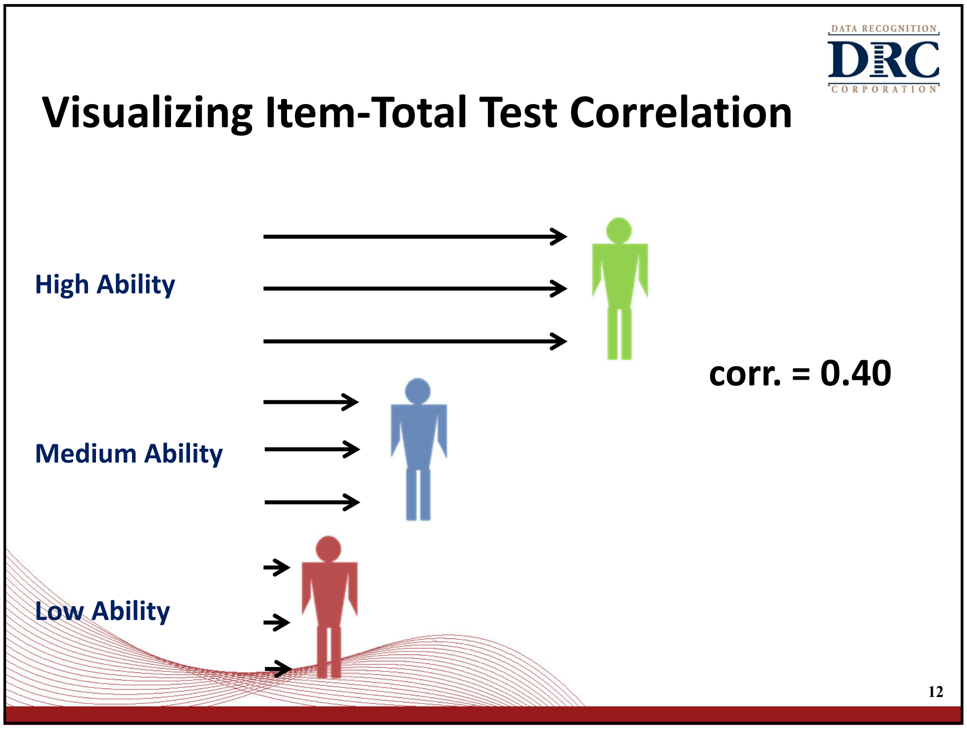
Traditional Statistics

N	P-Val	Mean	Item Total Corr
4349	0.73		0.49

11

11

11



Visualizing Item-Total Test Correlation

High Ability

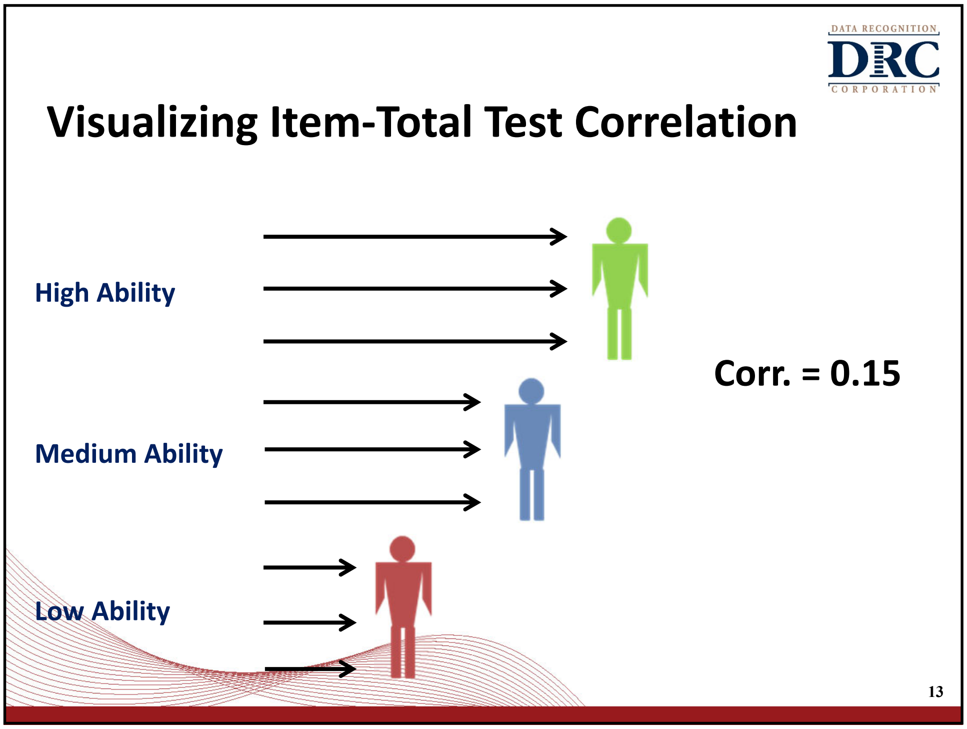
Medium Ability

Low Ability

corr. = 0.40

12

12



Visualizing Item-Total Test Correlation

High Ability

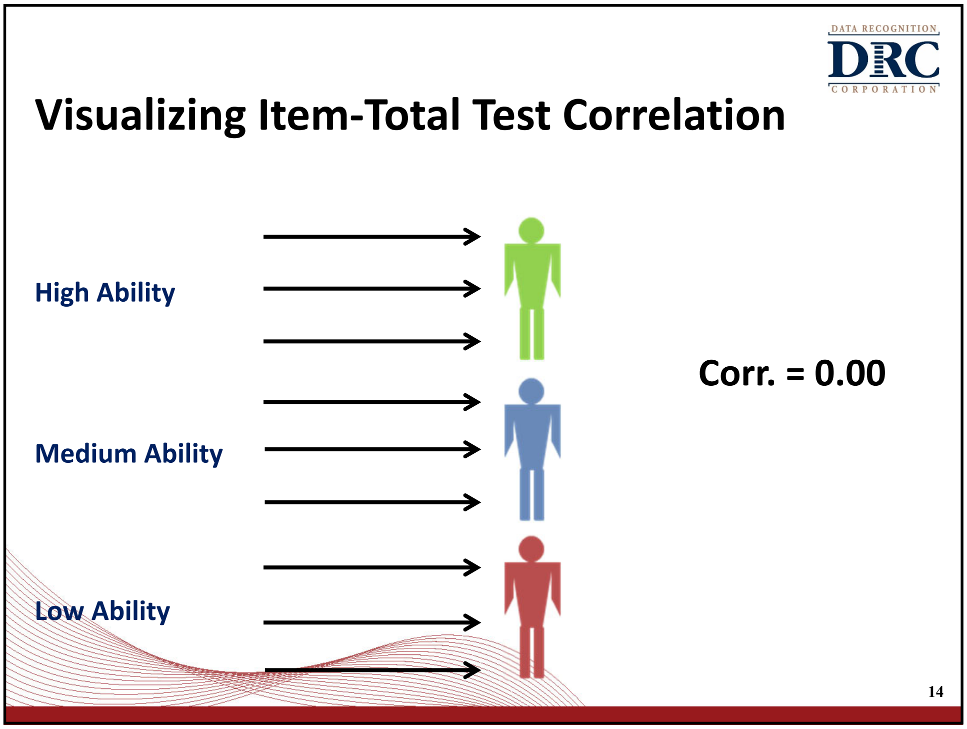
Medium Ability

Low Ability

Corr. = 0.15

13

13



Visualizing Item-Total Test Correlation

High Ability

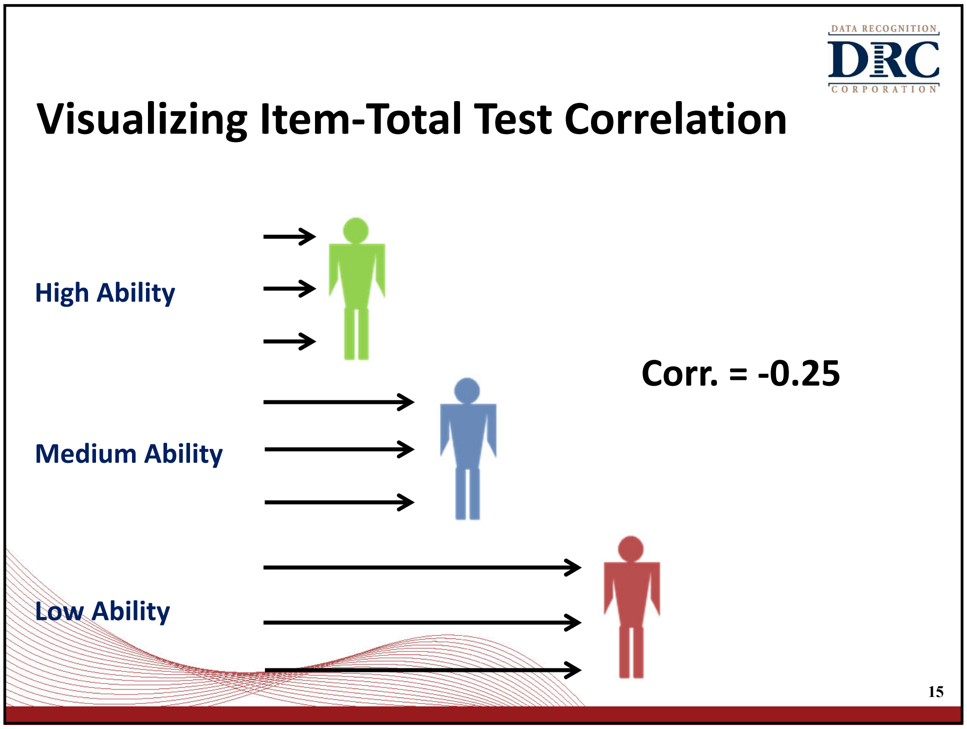
Medium Ability

Low Ability

Corr. = 0.00

14

14



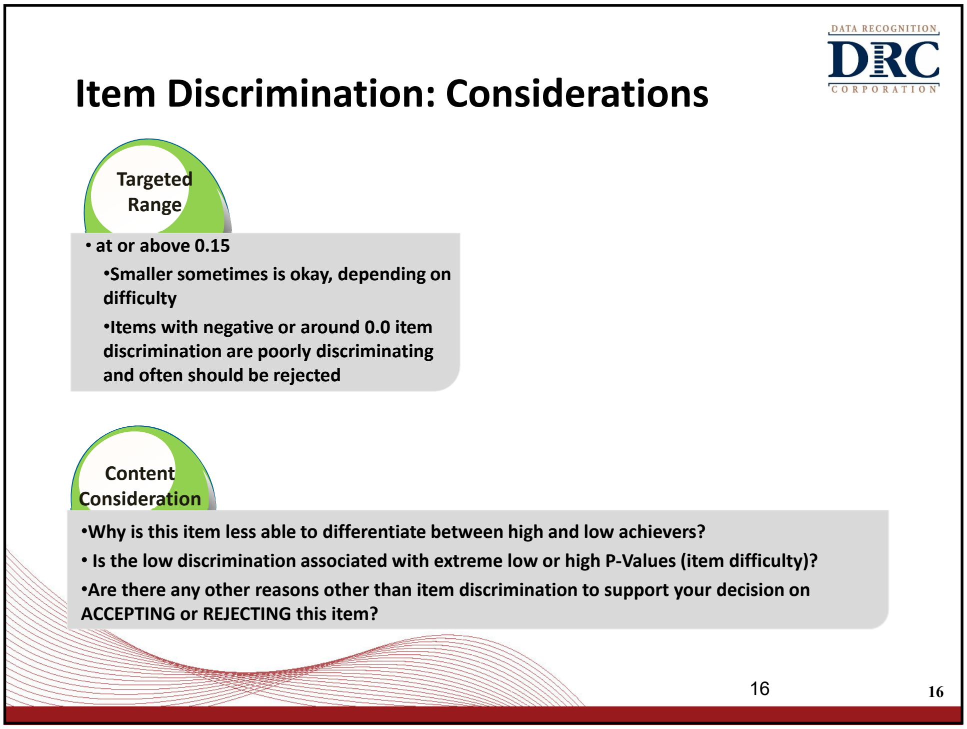
DATA RECOGNITION DRC CORPORATION

Visualizing Item-Total Test Correlation

Corr. = -0.25

15

15



DATA RECOGNITION DRC CORPORATION

Item Discrimination: Considerations

Targeted Range

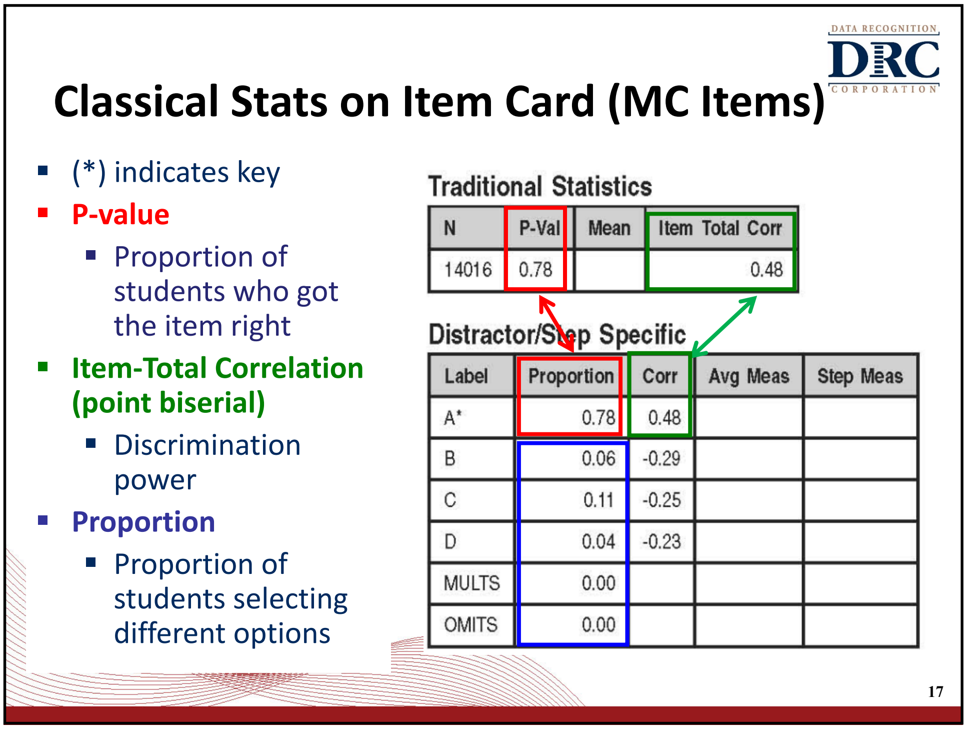
- at or above 0.15
  - •Smaller sometimes is okay, depending on difficulty
  - •Items with negative or around 0.0 item discrimination are poorly discriminating and often should be rejected

Content Consideration

- •Why is this item less able to differentiate between high and low achievers?
- Is the low discrimination associated with extreme low or high P-Values (item difficulty)?
- •Are there any other reasons other than item discrimination to support your decision on ACCEPTING or REJECTING this item?

16 16

16



Classical Stats on Item Card (MC Items)

- (*) indicates key
- **P-value**
  - Proportion of students who got the item right
- **Item-Total Correlation (point biserial)**
  - Discrimination power
- **Proportion**
  - Proportion of students selecting different options

Traditional Statistics

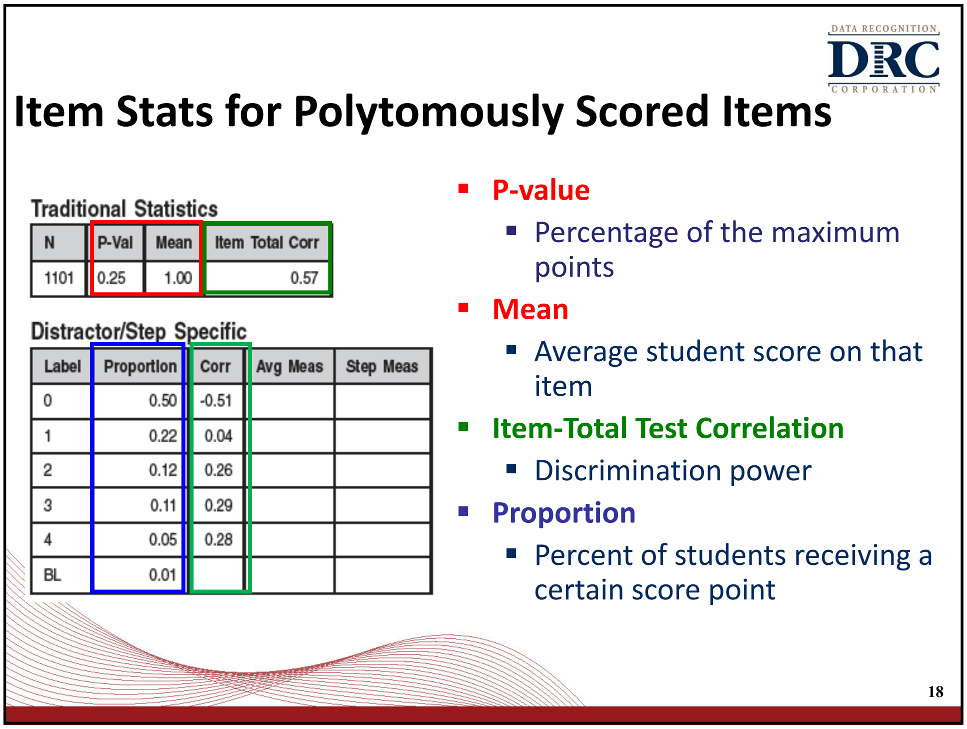
N	P-Val	Mean	Item Total Corr
14016	0.78		0.48

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Step Meas
A*	0.78	0.48		
B	0.06	-0.29		
C	0.11	-0.25		
D	0.04	-0.23		
MULTS	0.00			
OMITS	0.00			

17

17



Item Stats for Polytomously Scored Items

Traditional Statistics

N	P-Val	Mean	Item Total Corr
1101	0.25	1.00	0.57

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Step Meas
0	0.50	-0.51		
1	0.22	0.04		
2	0.12	0.26		
3	0.11	0.29		
4	0.05	0.28		
BL	0.01			

- **P-value**
  - Percentage of the maximum points
- **Mean**
  - Average student score on that item
- **Item-Total Test Correlation**
  - Discrimination power
- **Proportion**
  - Percent of students receiving a certain score point

18

18

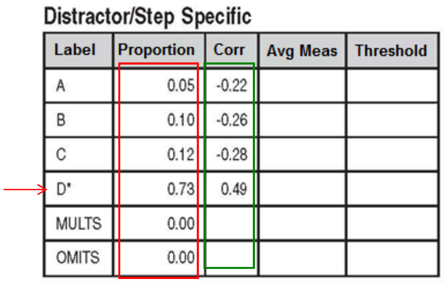
Polytomously Scored Items

- The mean gives a general idea of item difficulty but can sometimes be deceptive.

Proportion of students			
Score 0	Score 1	Score 2	Item Mean
0.40	0.20	0.40	1.0
0.15	0.70	0.15	1.0
0.33	0.33	0.33	1.0

- Use the score point proportions to determine if the distribution is reasonable.
- We want some students in all score-point categories.
  - item parameters cannot be estimated for the category with no or very few students.

Distractor Specifc Analysis (MC Items)



Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Threshold
A	0.05	-0.22		
B	0.10	-0.26		
C	0.12	-0.28		
D*	0.73	0.49		
MULTS	0.00			
OMITS	0.00			

Guideline

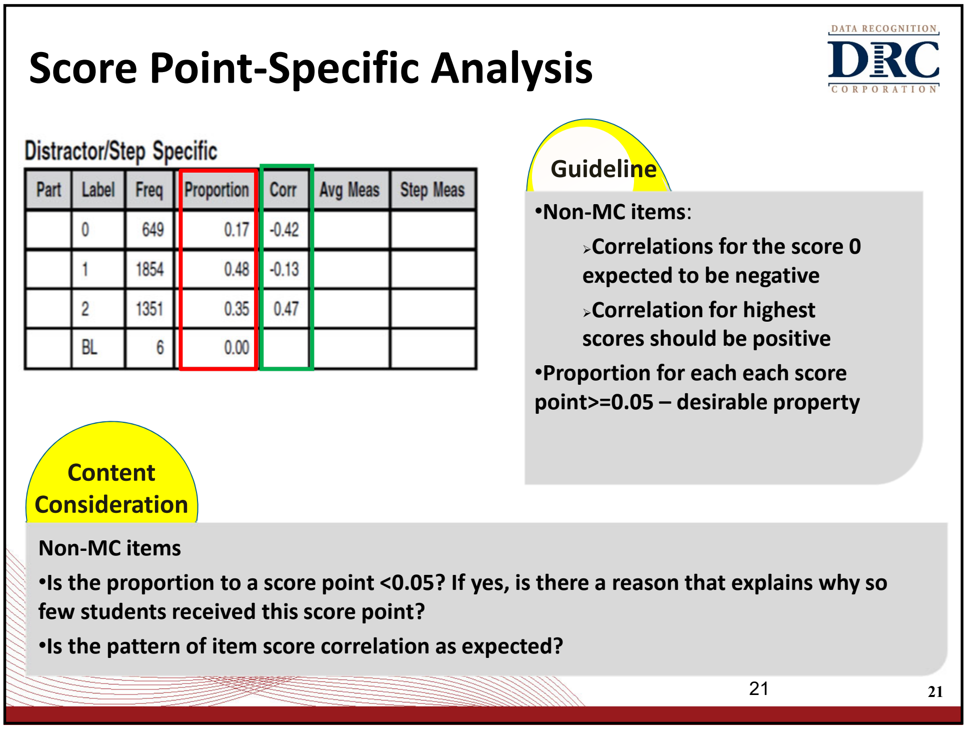
•MC items:

- Correlations for the distractors should be negative.
- Correlations for the distractors should never be higher than correlation for the correct answer
- Proportion of distractor < proportion of key

Content Consideraton

•Is the correlation of selecting any incorrect option greater than 0? If yes, why does this option distract more high achievers than low achievers?

•Is the proportion of selecting any incorrect option greater than the proportion of selecting the key? If yes, why?



Score Point-Specific Analysis

Distractor/Step Specific

Part	Label	Freq	Proportion	Corr	Avg Meas	Step Meas
	0	649	0.17	-0.42		
	1	1854	0.48	-0.13		
	2	1351	0.35	0.47		
	BL	6	0.00			

Guideline

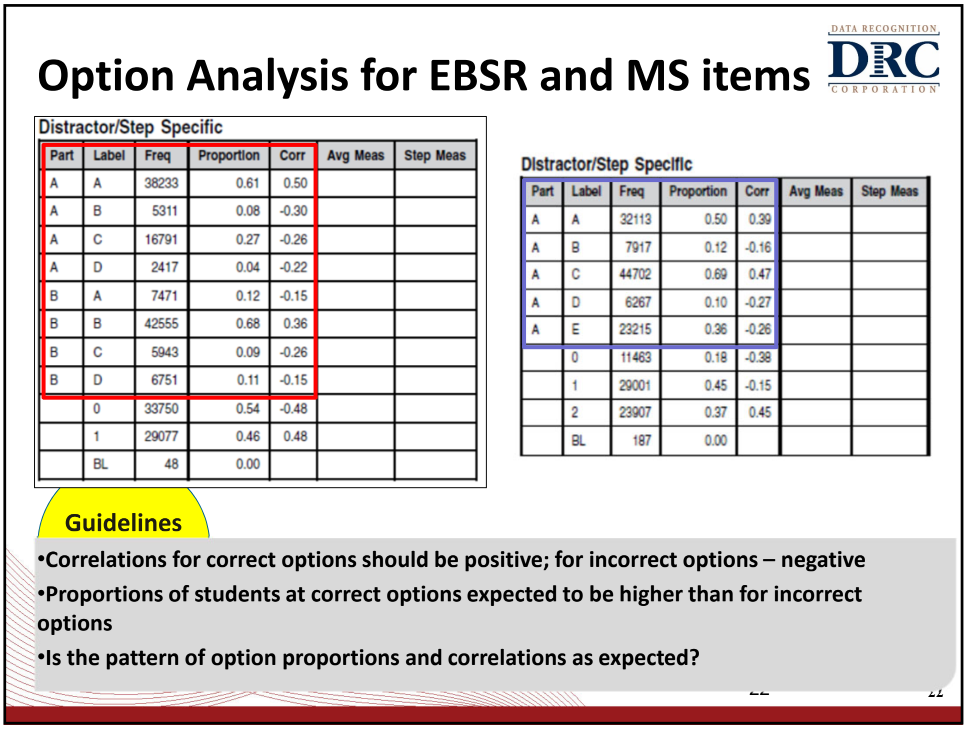
- •Non-MC items:
  - ›Correlations for the score 0 expected to be negative
  - ›Correlation for highest scores should be positive
- •Proportion for each each score point>=0.05 – desirable property

Content Consideration

Non-MC items

•Is the proportion to a score point <0.05? If yes, is there a reason that explains why so few students received this score point?

•Is the pattern of item score correlation as expected?



Option Analysis for EBSR and MS items

Distractor/Step Specific

Part	Label	Freq	Proportion	Corr	Avg Meas	Step Meas
A	A	38233	0.61	0.50		
A	B	5311	0.08	-0.30		
A	C	16791	0.27	-0.26		
A	D	2417	0.04	-0.22		
B	A	7471	0.12	-0.15		
B	B	42555	0.68	0.36		
B	C	5943	0.09	-0.26		
B	D	6751	0.11	-0.15		
	0	33750	0.54	-0.48		
	1	29077	0.46	0.48		
	BL	48	0.00			

Distractor/Step Specific

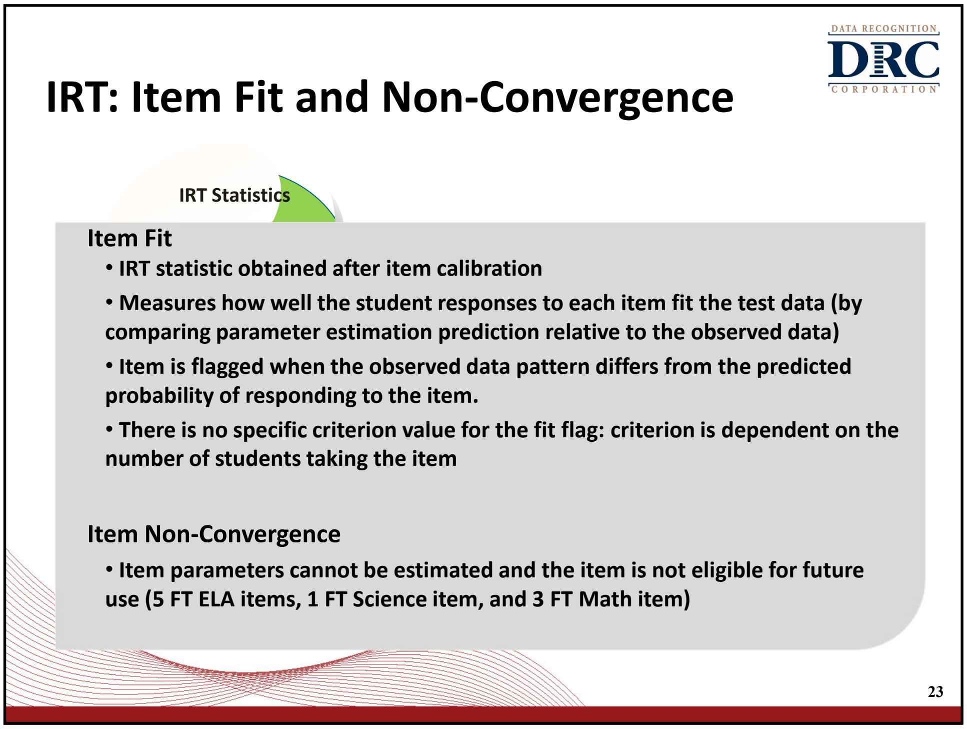
Part	Label	Freq	Proportion	Corr	Avg Meas	Step Meas
A	A	32113	0.50	0.39		
A	B	7917	0.12	-0.16		
A	C	44702	0.69	0.47		
A	D	6267	0.10	-0.27		
A	E	23215	0.36	-0.26		
	0	11463	0.18	-0.38		
	1	29001	0.45	-0.15		
	2	23907	0.37	0.45		
	BL	187	0.00			

Guidelines

•Correlations for correct options should be positive; for incorrect options – negative

•Proportions of students at correct options expected to be higher than for incorrect options

•Is the pattern of option proportions and correlations as expected?



IRT: Item Fit and Non-Convergence

IRT Statistics

Item Fit

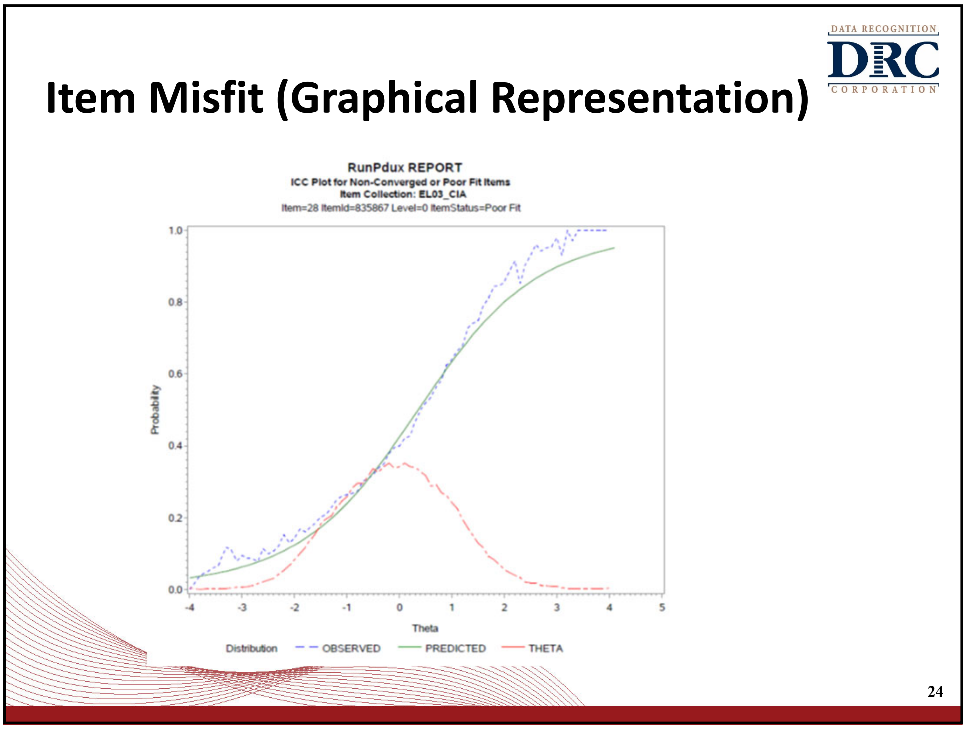
- IRT statistic obtained after item calibration
- Measures how well the student responses to each item fit the test data (by comparing parameter estimation prediction relative to the observed data)
- Item is flagged when the observed data pattern differs from the predicted probability of responding to the item.
- There is no specific criterion value for the fit flag: criterion is dependent on the number of students taking the item

Item Non-Convergence

- Item parameters cannot be estimated and the item is not eligible for future use (5 FT ELA items, 1 FT Science item, and 3 FT Math item)

23

23



Item Misfit (Graphical Representation)

RunPdux REPORT

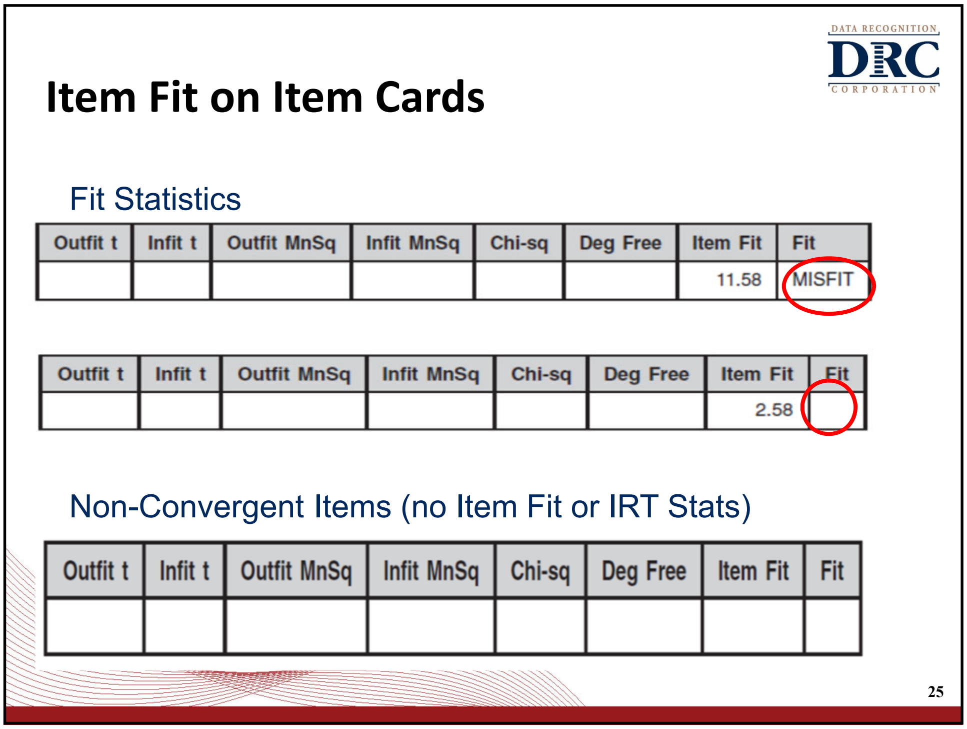
ICC Plot for Non-Converged or Poor Fit Items

Item Collection: EL03_CIA

Item=28 ItemId=835867 Level=0 ItemStatus=Poor Fit

24

24



Item Fit on Item Cards

DATA RECOGNITION DRC CORPORATION

Fit Statistics

Outfit t	Infit t	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Item Fit	Fit
						11.58	MISFIT

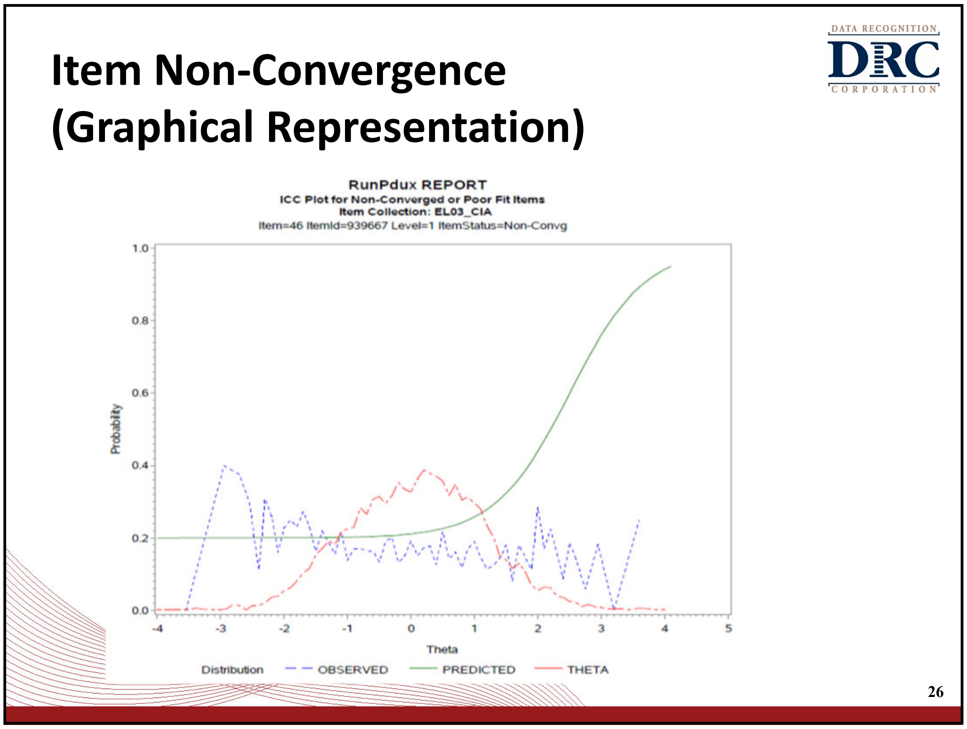
Outfit t	Infit t	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Item Fit	Fit
						2.58	

Non-Convergent Items (no Item Fit or IRT Stats)

Outfit t	Infit t	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Item Fit	Fit

25

25



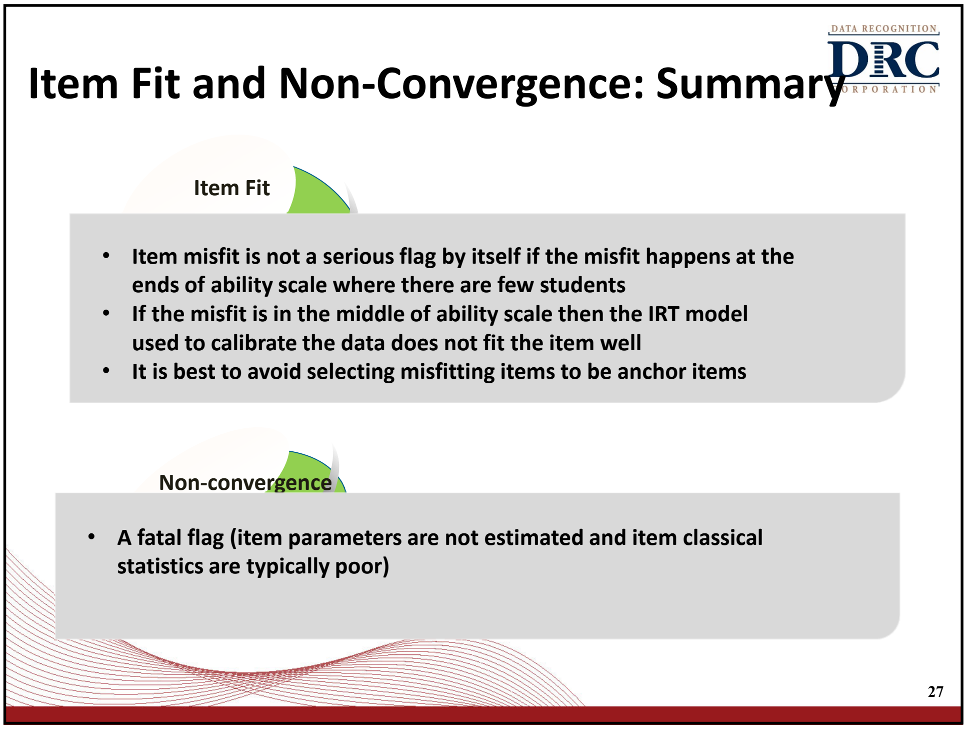
Item Non-Convergence (Graphical Representation)

DATA RECOGNITION DRC CORPORATION

RunPdux REPORT
ICC Plot for Non-Converged or Poor Fit Items
Item Collection: EL03_CIA
Item=46 ItemId=939667 Level=1 ItemStatus=Non-Convg

26

26



Item Fit and Non-Convergence: Summary

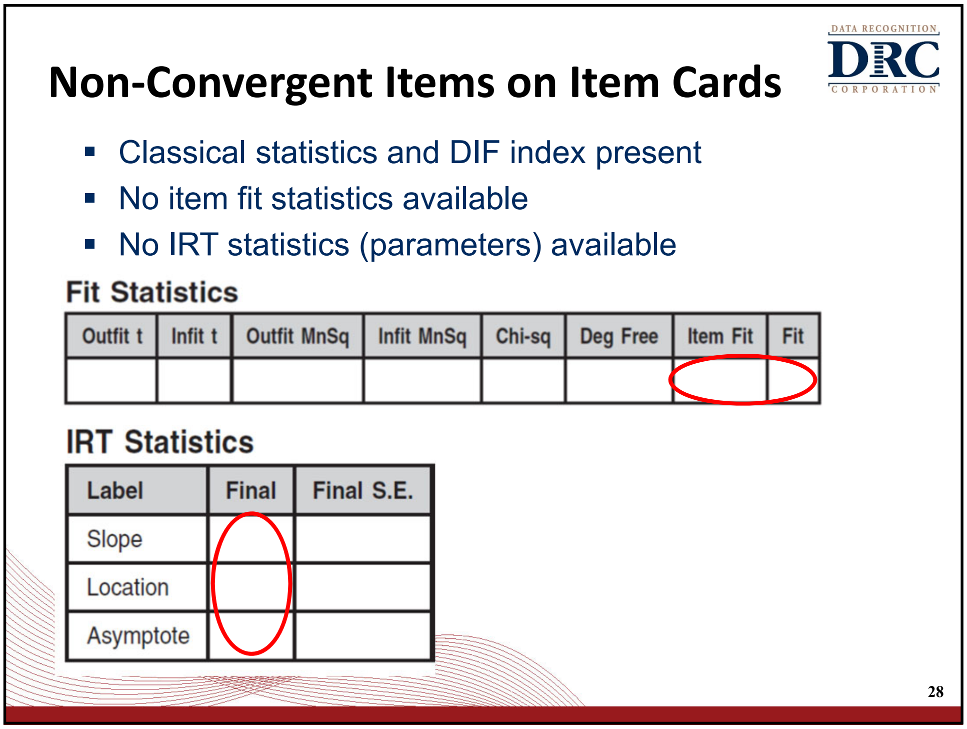
DATA RECOGNITION DRC CORPORATION

Item Fit

- Item misfit is not a serious flag by itself if the misfit happens at the ends of ability scale where there are few students
- If the misfit is in the middle of ability scale then the IRT model used to calibrate the data does not fit the item well
- It is best to avoid selecting misfitting items to be anchor items

Non-convergence

- A fatal flag (item parameters are not estimated and item classical statistics are typically poor)



Non-Convergent Items on Item Cards

DATA RECOGNITION DRC CORPORATION

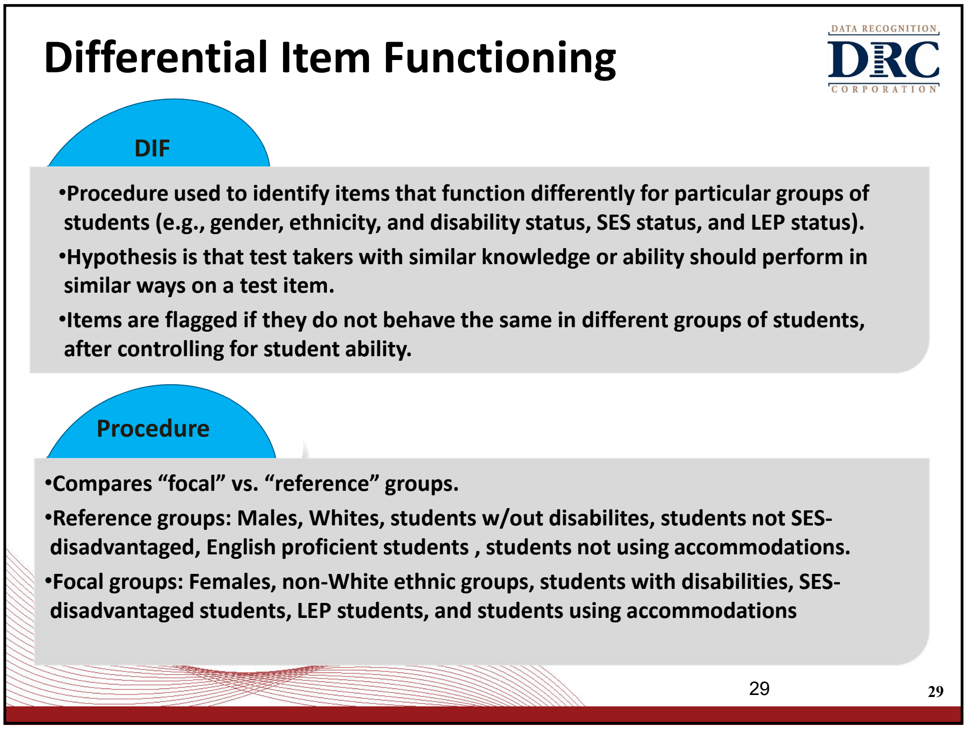
- Classical statistics and DIF index present
- No item fit statistics available
- No IRT statistics (parameters) available

Fit Statistics

Outfit t	Infit t	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Item Fit	Fit

IRT Statistics

Label	Final	Final S.E.
Slope		
Location		
Asymptote		



Differential Item Functioning

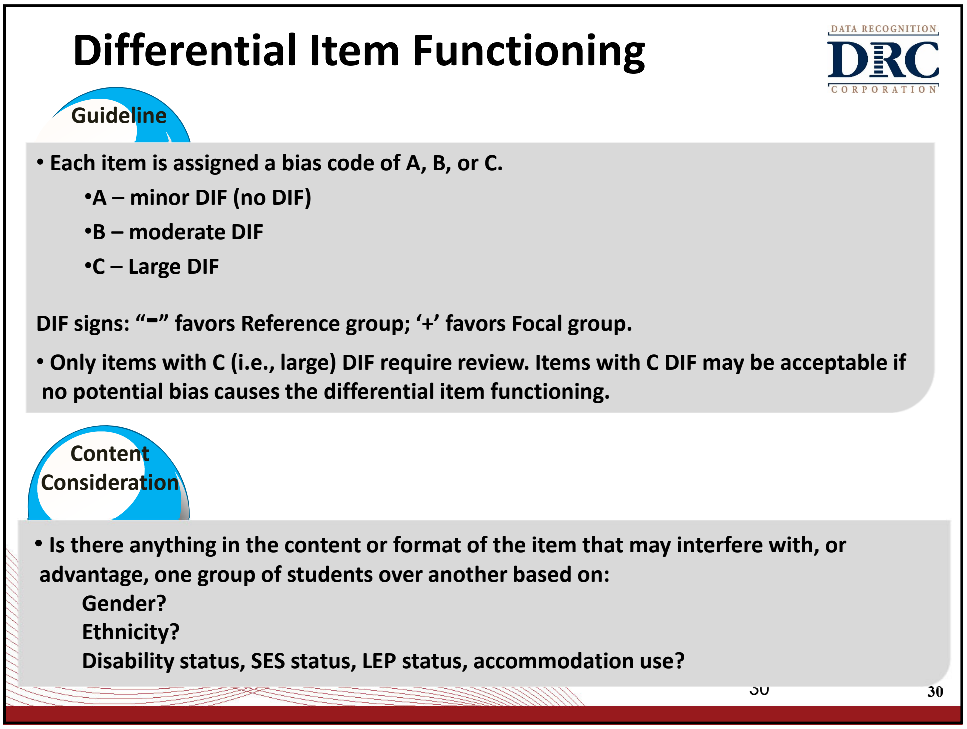
DATA RECOGNITION
DRC
CORPORATION

DIF

- Procedure used to identify items that function differently for particular groups of students (e.g., gender, ethnicity, and disability status, SES status, and LEP status).
- Hypothesis is that test takers with similar knowledge or ability should perform in similar ways on a test item.
- Items are flagged if they do not behave the same in different groups of students, after controlling for student ability.

Procedure

- Compares “focal” vs. “reference” groups.
- Reference groups: Males, Whites, students w/out disabilites, students not SES-disadvantaged, English proficient students , students not using accommodations.
- Focal groups: Females, non-White ethnic groups, students with disabilities, SES-disadvantaged students, LEP students, and students using accommodations



Differential Item Functioning

DATA RECOGNITION
DRC
CORPORATION

Guideline

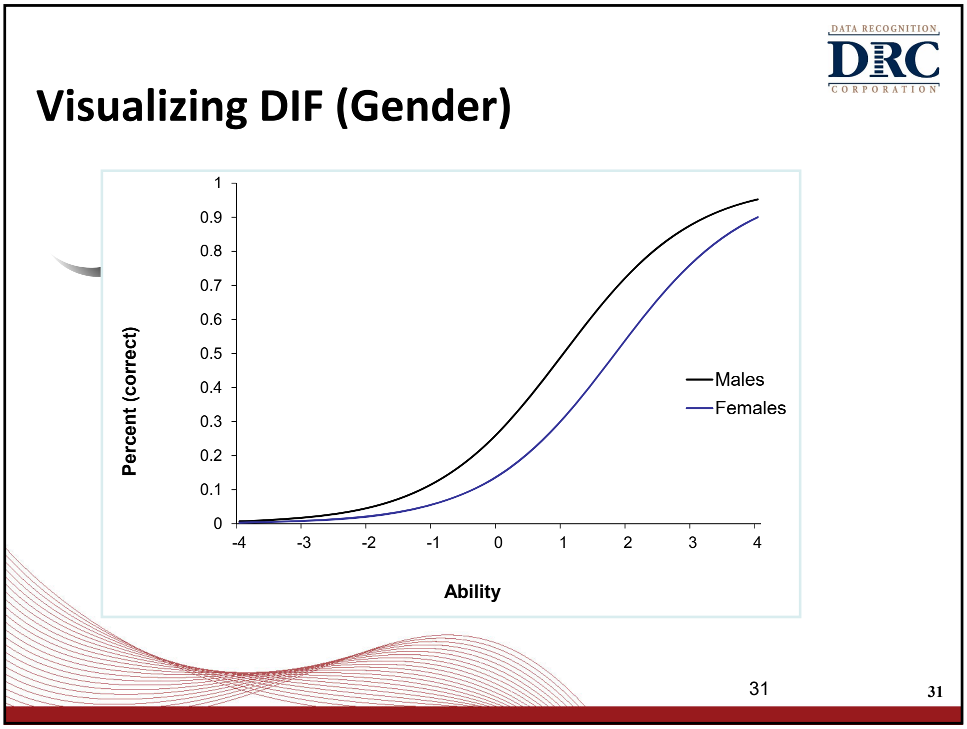
- Each item is assigned a bias code of A, B, or C.
    - A – minor DIF (no DIF)
    - B – moderate DIF
    - C – Large DIF

DIF signs: “-” favors Reference group; ‘+’ favors Focal group.

- Only items with C (i.e., large) DIF require review. Items with C DIF may be acceptable if no potential bias causes the differential item functioning.

Content Consideration

- Is there anything in the content or format of the item that may interfere with, or advantage, one group of students over another based on:
  Gender?
  Ethnicity?
  Disability status, SES status, LEP status, accommodation use?

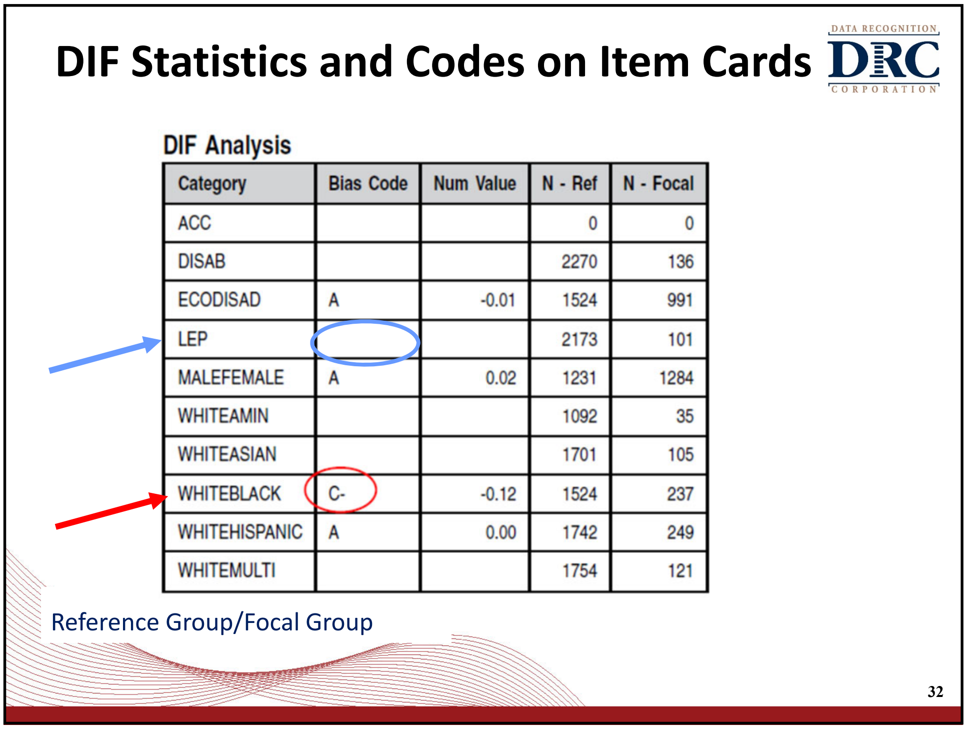


Visualizing DIF (Gender)

31

31

31



DIF Statistics and Codes on Item Cards

DATA RECOGNITION
DRC
CORPORATION

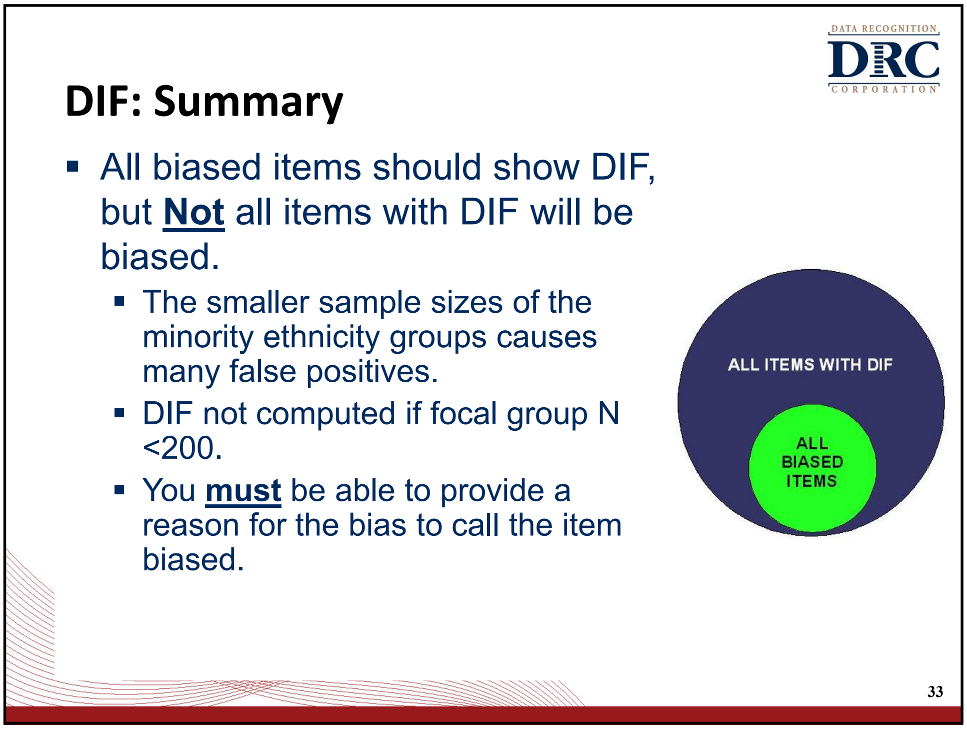
DIF Analysis

Category	Bias Code	Num Value	N - Ref	N - Focal
ACC			0	0
DISAB			2270	136
ECODISAD	A	-0.01	1524	991
LEP			2173	101
MALEFEMALE	A	0.02	1231	1284
WHITEAMIN			1092	35
WHITEASIAN			1701	105
WHITEBLACK	C-	-0.12	1524	237
WHITEHISPANIC	A	0.00	1742	249
WHITEMULTI			1754	121

Reference Group/Focal Group

32

32



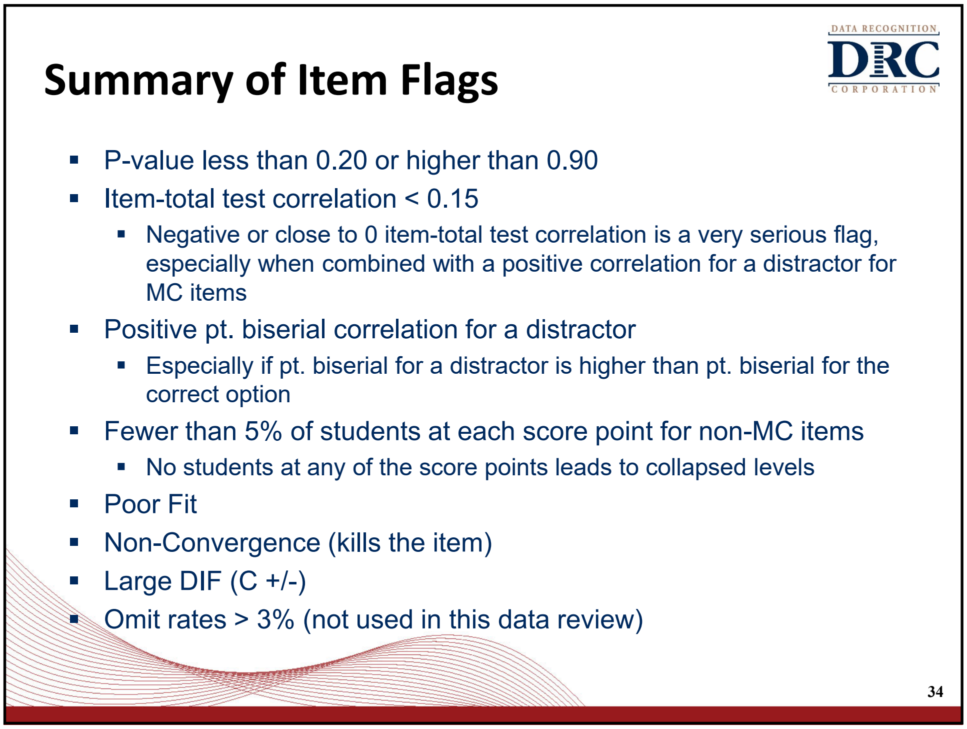
DIF: Summary

- All biased items should show DIF, but **Not** all items with DIF will be biased.
  - The smaller sample sizes of the minority ethnicity groups causes many false positives.
  - DIF not computed if focal group N <200.
  - You **must** be able to provide a reason for the bias to call the item biased.

ALL ITEMS WITH DIF
ALL BIASED ITEMS

33

33

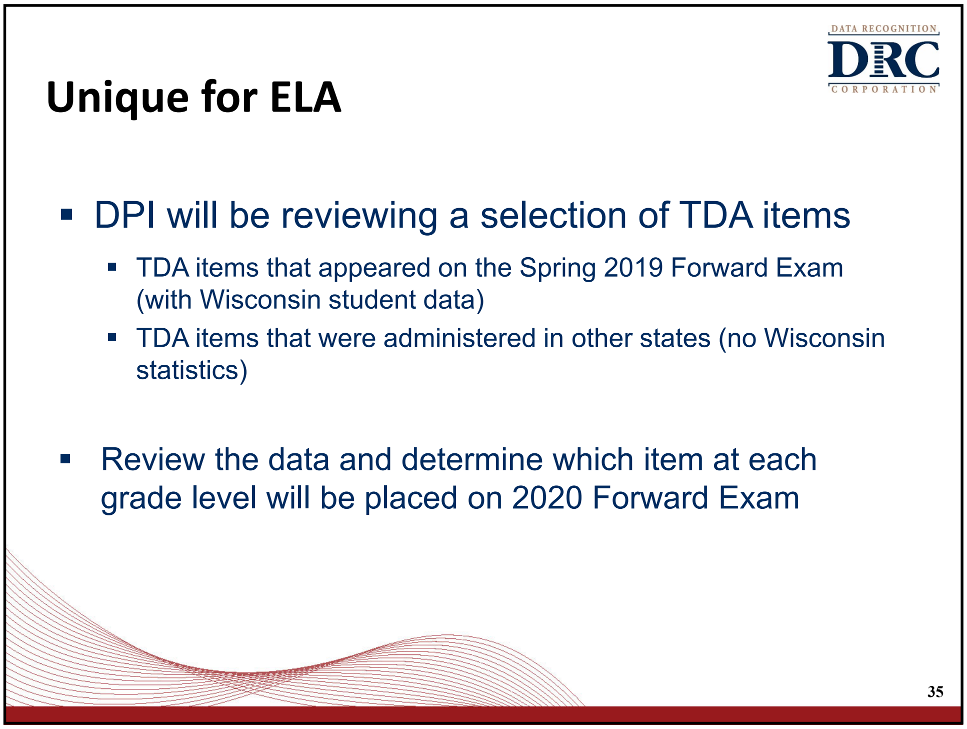


Summary of Item Flags

- P-value less than 0.20 or higher than 0.90
- Item-total test correlation < 0.15
  - Negative or close to 0 item-total test correlation is a very serious flag, especially when combined with a positive correlation for a distractor for MC items
- Positive pt. biserial correlation for a distractor
  - Especially if pt. biserial for a distractor is higher than pt. biserial for the correct option
- Fewer than 5% of students at each score point for non-MC items
  - No students at any of the score points leads to collapsed levels
- Poor Fit
- Non-Convergence (kills the item)
- Large DIF (C +/-)
- Omit rates > 3% (not used in this data review)

34

34



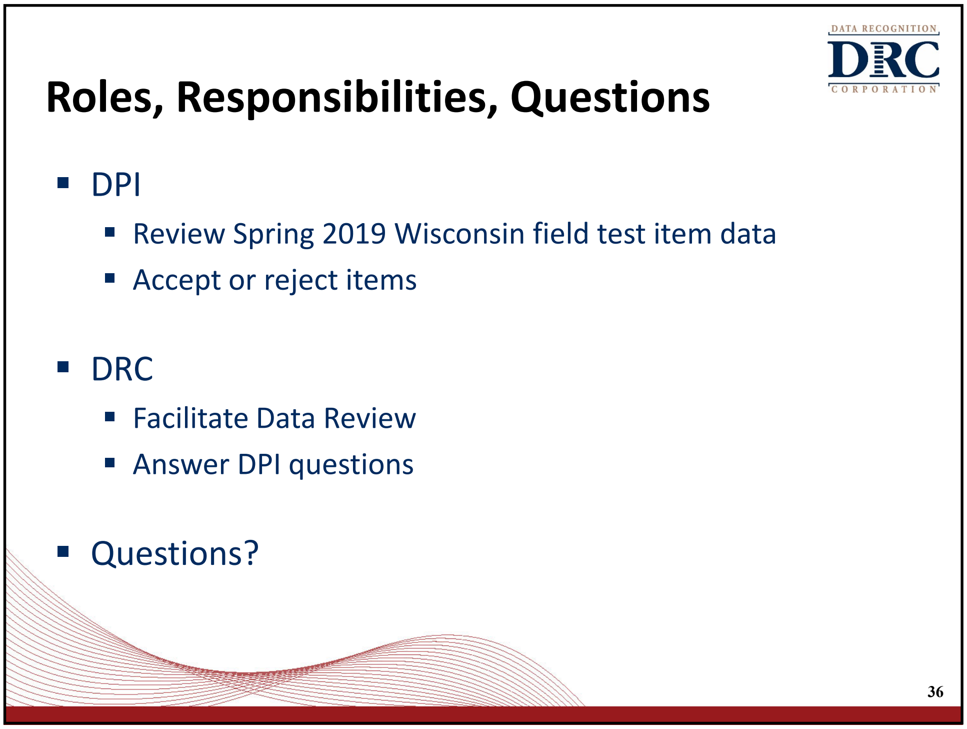
DATA RECOGNITION DRC CORPORATION

Unique for ELA

- DPI will be reviewing a selection of TDA items
  - TDA items that appeared on the Spring 2019 Forward Exam (with Wisconsin student data)
  - TDA items that were administered in other states (no Wisconsin statistics)
- Review the data and determine which item at each grade level will be placed on 2020 Forward Exam

35

35



DATA RECOGNITION DRC CORPORATION

Roles, Responsibilities, Questions

- DPI
  - Review Spring 2019 Wisconsin field test item data
  - Accept or reject items
- DRC
  - Facilitate Data Review
  - Answer DPI questions
- Questions?

36

36