



Ohio Assessments for Educators

For candidates seeking **initial licensure** in a subject area

Test Administration Technical Report

2014–2015 Program Year

December 2015

Prepared for:
Center for the Teaching Profession
Ohio Department of Education
25 South Front Street, MS 505
Columbus, OH 43215-4183

Contents

Preface	ii
Purpose of the Ohio Assessments for Educators (OAE) Program	1
Introduction	1
Composition of the OAE Program.....	1
Technical Properties of the OAE Program	3
Scoring	3
Item Analyses.....	5
Test Equating	5
Scaled Scores	6
Reliability	8
Validity	10
Score Reporting	12
Candidate Score Reports	12
Other Score Recipients.....	12
Test Statistics Reports	13
Total Scaled Score Distribution by Test Field.....	14
Title II Reporting.....	14
References	16
Appendices	17

Preface

This test administration technical report provides information on the technical characteristics of the tests and score reporting for the Ohio Assessments for Educators (OAE) for the 2014–2015 program year.

Purpose of the Ohio Assessments for Educators (OAE) Program

Introduction

The Ohio Assessments for Educators (OAE): For candidates seeking initial licensure in a subject area (OAE: Initial Licensure) program assesses the content-area and professional (pedagogical) knowledge of candidates who are either seeking initial Ohio educator licensure or adding a new licensure area. The OAE program, administered by Pearson, includes 38 content-area assessments and four professional (pedagogical) knowledge assessments. Five OAE assessments include two separate tests each (i.e., Subtest I and Subtest II) for a total of 47 unique tests. The OAE tests are aligned with Ohio Educational Preparation Standards, Ohio Standards for the Teaching profession, and Ohio Student Standards.

Each test was validated for use in Ohio in accordance with the practices recommended by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The Standards require a clear definition of content domain and a rationale to support a claim that the knowledge and skills being assessed in a licensure test are required for credential-worthy performance. Educators, educator preparation faculty, and administrators from across Ohio were involved in reviewing the test materials for content, job-relatedness, and prevention of bias; validating their appropriateness for use in Ohio; and making recommendations for the passing score for each test. In addition, in accordance with State of Ohio requirements, assessment materials, where available, were to have been previously administered to educator licensure candidates in states other than Ohio.

The OAE tests are computer-based and delivered through a national network of Pearson computer-based testing centers. Most tests are available year round by appointment.

The OAE program offers several web-based resources to help candidates prepare for the tests. These resources include online study guides, practice tests, detailed score reports, and computer-based testing tutorials. In addition, a suite of faculty resources and interactive worksheets are available to assist in candidate preparation. The Ohio Department of Education and educator preparation programs have access to an interactive, electronic database that allows them to create customized reports of candidate test results, institution performance, or perform customized data queries.

Composition of the OAE Program

Currently, 46 OAE tests¹ are available for test administration. The OAE program includes four professional (pedagogy) knowledge tests that are matched to Ohio licensure grade bands (Early Childhood, Middle Childhood, Adolescence to Young Adult, and Multi-Age). Content-area tests match Ohio license types. Thirty-two OAE tests have been operational since September 3, 2013; 12 OAE tests have been operational since January 21, 2014; two OAE tests have been operational since September 2, 2014.

¹ Validation activities are in progress for the Dance test.

Ohio Assessments for Educators (OAE)

Pedagogical Knowledge Assessments

- 001 Assessment of Professional Knowledge: Early Childhood (PK–3)
- 002 Assessment of Professional Knowledge: Middle Childhood (4–9)
- 003 Assessment of Professional Knowledge: Adolescence to Young Adult (7–12)
- 004 Assessment of Professional Knowledge: Multi-Age (PK–12)

Content Knowledge Assessments

- 005 Agriscience
- 006 Art
- 007 Biology
- 008 Business Education
- 009 Chemistry
- 010 Computer Information Science
- 011 Dance
- 012 Early Childhood Education
- 013 Early Childhood Special Education
- 014 Earth and Space Science
- 015 Educational Leadership
- 016 Computer/Technology (Subtest I)
- 017 Computer/Technology (Subtest II)
- 018 Elementary Education (Subtest I)
- 019 Elementary Education (Subtest II)
- 020 English Language Arts
- 021 English to Speakers of Other Languages
- 022 Family and Consumer Sciences
- 023 Health
- 024 Integrated Science
- 025 Integrated Social Studies
- 026 Marketing
- 027 Mathematics
- 028 Middle Grades English Language Arts
- 029 Middle Grades Science
- 030 Middle Grades Mathematics
- 031 Middle Grades Social Studies
- 032 Music
- 034 Physical Education
- 035 Physics
- 036 Prekindergarten (Subtest I)
- 037 Prekindergarten (Subtest II)
- 038 Reading (Subtest I)

- 039 Reading (Subtest II)
- 040 School Counselor
- 041 School Library Media Specialist
- 042 School Psychologist
- 043 Special Education
- 044 Special Education Specialist: Deaf/Hard of Hearing
- 045 Special Education Specialist: Visually Impaired
- 046 Technology Education (Subtest I)
- 047 Technology Education (Subtest II)
- 048 Theater

Technical Properties of the OAE Program

The *Standards for Educational and Psychological Testing* require that testing agencies provide relevant technical information about the assessments so that test users and reviewers have sufficient information to make judgments about the quality of the test, the resulting scores, and interpretations based on test scores (AERA, APA, & NCME, 2014). This information can ultimately assist test users and reviewers in determining the appropriateness of the test for its intended purpose (AERA, APA, & NCME, 2014).

Scoring

The OAE includes tests consisting of multiple-choice items only as well as tests consisting of both multiple-choice and constructed-response items. The scoring procedures for the program are carefully documented for both multiple-choice and constructed-response items. Additionally, performance monitoring is conducted to check the accuracy of scoring and reporting for the constructed-response items.

Scoring Multiple-Choice Items

Answer keys for multiple-choice items are prepared during the construction of test forms. These keys are reviewed and checked at several points during development and production. Automated technology at computer-based testing centers compare these answer keys to candidate responses immediately following the conclusion of testing. This technology allows for on-site unofficial test results to be generated at the testing centers, which can then be provided to candidates. These results are provided to the candidates only for the OAE assessments that do not include constructed-response items.

Multiple-choice items are dichotomously scored, meaning a single point is awarded for each correct response, and no points are awarded for an incorrect response. For tests composed of multiple-choice items only, the final raw score is the total number of correct responses on the test. The raw scores are transformed and reported on a scale ranging from 100 to 300 with a scaled passing score of 220.

Each test form for all OAE fields includes both scorable and nonscorable multiple-choice items. Scorable items are those that are used to compute candidates' scores. Nonscorable items are those that are included on a test form to collect additional psychometric information (to support pilot testing) and to support test form equating methodologies but do not contribute to candidates' scores.

Scoring Constructed-Response Items

Some OAE tests include constructed-response items in addition to multiple-choice items. Candidate responses to constructed-response items are scored using a focused holistic scoring methodology. In this method, scorers judge the overall effectiveness of each response using a set of characteristics that have been defined as important to help inform the overall score. The score is holistic in that each score is based on the overall effectiveness of these characteristics working together, focusing on the response as a whole.

The Special Education Specialist: Visually Impaired constructed-response item is scored differently. This item requires candidates to transcribe a passage into braille, and the score provided is based on the number of errors made in the transcription.

Scoring Components. Candidate responses are scored on a scale. Some tests use a four-point score scale; other tests use a three-point score scale. A score of 1 represents little or no command of the characteristics, and 3 or 4 represents a strong command of the characteristics. Each constructed response is independently scored by two scorers, and these scores are summed for a total possible score range of 2 to 6 for three-point score scales or 2 to 8 for four-point score scales.

To participate in the holistic scoring process, scorers must meet specific qualifications such as:

- A state teaching certificate/license;
- Teaching experience in public schools; and/or
- Experience as a college educator responsible for preparing prospective teachers.

Scorer Training. Prior to scoring, each scorer receives training and orientation by a Chief Reader who trains each scorer to evaluate responses to a variety of teacher licensure examinations. The Chief Reader provides each scorer with the background of the OAE tests, the context of the scoring task, the tasks they will perform, the procedures they will follow, the scoring scale, and the characteristics that will guide their scoring judgments. The scorers are also oriented to the prompt-specific rubrics that apply the characteristics and scoring scale to the constructed-response item(s) used on test forms for the current test administration. The Chief Reader leads scoring sessions and identifies and develops orientation materials, conducts calibration orientation and assessments, and monitors the progress of the scoring session.

Scorers receive practice in scoring training sets of responses to which scores have already been assigned, including marker responses (i.e., exemplar responses of each score point on the score scale). The training emphasizes equity and fair application of the score scale. Once scorers have been trained, they have to prove their ability to score accurately by completing a calibration exercise. Each response is read independently and scored by two scorers. If the two scores are not identical or adjacent, additional scoring is conducted to resolve the discrepancy.

Performance Monitoring of Scorers

Pearson monitors the performance of scorers throughout the scoring process. Specific areas monitored include a scorers' ability to understand and apply the established scoring scale, the consistency of the scores assigned in comparison with those assigned by the second scorer, and the scorers' consistency over time. At points in the holistic scoring process, scorers are recalibrated to the scale, typically through discussions of specific questions. Scorers must demonstrate continued scoring accuracy on the

responses. If scorers fail to demonstrate accuracy, they receive additional, individual reorientation before proceeding with scoring.

Item Analyses

Item analyses are conducted on multiple-choice items to assess the accuracy and psychometric quality of the items. Additionally, data from constructed-response items are reviewed to confirm that items in the item bank for each field are comparable in terms of difficulty and score distribution.

Item Analysis for Multiple-Choice Items

The purpose of item analysis for multiple-choice items is to verify the accuracy of the answer key for each administered test form and as an additional quality assurance check before providing final results and before official candidate score reports are produced. Data are collected on each item, allowing for the empirical consideration of item difficulty, item discrimination, content accuracy, and the plausibility of distractors. These item statistics are calculated and evaluated for the current administration and cumulatively (i.e., using combined statistics for previous operational test administrations).

The item statistics calculated and evaluated for each multiple-choice item include:

- Item difficulty (p-value);
- Distribution of responses (percentages of participants selecting each response option);
- Item-to-total test point biserial correlation (correlation of performance on the item to performance on the total test); and
- Mean score by response choice (average score on the total multiple-choice set achieved by all participants selecting each response option).

Those items that do not perform within defined statistical parameters are flagged and reviewed. Flagged items are reviewed by content specialists, test development specialists, psychometricians, and/or editors and are deleted or revised and subjected to additional pilot testing. Specifically, item review includes: a review to confirm that the wording on the test form is the same as the wording of the item as validated by the CACs, a check of content, topicality, and correct answer with documented sources, and/or review by a content expert. Based on the results of the review process, a decision is made regarding whether or not to include those items as scorable, and Pearson implements any necessary scoring actions.

Item Review for Constructed-Response Items

Constructed-response item data are regularly reviewed to monitor the difficulty of the items across administrations. Throughout the scoring process, monitor reports containing mean scores and standard deviations are reviewed. The monitor reports also contain distribution of scores assigned by the first two scorers, distribution of score differences, and distribution of score combinations assigned by the first two scorers.

Test Equating

Each OAE test consists of multiple forms. Multiple forms are utilized within and across test administrations to address issues of item exposure and security. Statistical

adjustment (equating) is implemented to adjust for small differences in difficulty across forms.

Purpose of Equating

According to the *Standards for Educational and Psychological Testing*, equating refers to the process of placing scores from alternate, parallel, or equivalent forms of a test on a common scale (AERA, APA, & NCME, 2014). The central purpose of statistical equating methods is to compensate statistically for possible variability in the characteristics of test forms that may affect candidates' scores (e.g., differences in the overall difficulty of a new test form compared to a previous test form). Statistical equating methods ensure that a candidate's scaled score is adjusted for the relative difficulty of the particular test form that was taken. Equating thus allows test developers to attribute differences in scores across test forms to differences in knowledge or skills, and not differences in the tests. Equating helps to ensure that different test forms are comparable and that test scores from multiple test forms can be used interchangeably.

Equating Design

A single-group equating design is utilized for the OAE tests. In a single-group design, the same group of candidates is scored on two alternative forms of the same test. The two forms are defined by designating two alternative subsets of items on the full test form as scorable, so that each alternative contains a set of common scorable items plus a set of scorable items unique to the form. The two alternative forms are then statistically equated.

To equate the two test forms, a linear equating method was implemented for the OAE. In linear equating, two scores are equivalent if they are the same number of standard deviation units above or below the mean for some group of candidates (Angoff, 1984). A linear equation is used to relate the scores from the two forms by setting standard deviation scores, or z-scores, to be equal on the two test forms (Kolen & Brennan, 2004).

This design is used for the OAE tests because of the need to pre-equate the test forms, enabling candidates to receive unofficial test results at the testing center immediately after the conclusion of their test administration (for multiple-choice only OAE tests). With pre-equating, the passing score for a new test form is established prior to operational administration.

Scaled Scores

The *Standards for Educational and Psychological Testing* state that scaled scores may aid in interpretation of the test. Specifically, scaled scores allow scores to be easily comparable regardless of test form or administration (AERA, APA, & NCME, 2014). Scaled score reporting is preferred to raw score reporting due to the confusion that may occur as a result of some changes in raw cut scores across test forms. Therefore, the use of scaled scores helps to support the communication of the OAE program results in the following ways:

- Candidates, institutions of higher education, and stakeholders will be able to interpret scores from the different OAE tests in a similar manner, regardless of the test taken.
- The meaning of the scaled passing scores will be consistent over time making it possible to compare performance from one test administration to the next.

Scaled Score Range for the OAE

Raw test scores, including the number correct for multiple-choice items, and holistic scores for constructed-response items are transformed to a standard scale ranging from 100 to 300. This scale is used to communicate OAE scores across all tests within the program. For each OAE test, 220 represents the passing score. By using a standard scale across all tests, test scores and feedback are provided in an easy-to-understand format for candidates, institutions of higher education, the ODE, and other stakeholders.

Calculation of Scaled Scores

For tests containing only multiple-choice items, a simple linear transformation is applied to the raw scores to compute the associated scaled score. The scaled score is derived from the candidate's raw score, the raw cut score, and the maximum possible raw score. Candidates who perform at the raw score that is equivalent to the passing score achieve a scaled score of 220, while those who achieve the maximum possible raw score will receive a scaled score of 300.

For the OAE tests that contain both multiple-choice and constructed-response items, the multiple-choice and constructed-response test sections are each scaled separately. These multiple-choice and constructed-response scaled scores are then combined in accordance with the following weightings.

OAE Test	Multiple-Choice Weight	Constructed-Response Weight
Assessment of Professional Knowledge: Early Childhood (PK-3)	80%	20%
Assessment of Professional Knowledge: Middle Childhood (4-9)	80%	20%
Assessment of Professional Knowledge: Adolescence to Young Adult (7-12)	80%	20%
Assessment of Professional Knowledge: Multi-Age (PK-12)	80%	20%
Computer Information Science	85%	15%
Educational Leadership	60%	40%
Computer Technology (Subtest I)	70%	30%
Computer Technology (Subtest II)	70%	30%
Pre-Kindergarten (Subtest I)	80%	20%
Pre-Kindergarten (Subtest II)	80%	20%
Reading (Subtest I)	80%	20%
Reading (Subtest II)	80%	20%
School Psychologist	85%	15%
Special Education: Visually Impaired	80%	20%
Technology Education (Subtest I)	70%	30%
Technology Education (Subtest II)	70%	30%

OAE Test	Multiple-Choice Weight	Constructed-Response Weight
Theater	80%	20%

Reliability

The *Standards for Educational and Psychological Testing* refers to reliability as the consistency of test scores for a group of candidates across administrations (AERA, APA & NCME, 2014). There are many common reasons for individual scores to fluctuate over time. Score fluctuation from one testing occasion to another has an impact on reliability. Some factors that affect reliability include:

- **Number of candidates.** The number of candidates whose test scores contribute to a statistical estimate of reliability affects the stability of the estimate. In general, reliability estimates based on larger numbers of candidates are more stable than estimates based on smaller numbers. For this reason, reliability estimates are calculated for tests that are taken by one hundred or more candidates.
- **Self-selection of candidates by test administration date.** Typically, candidates can decide when to take a particular test. OAE tests are administered throughout the year, and candidates can select when to take and retake the tests. This self-selection can affect the composition, ability level, and variability of the group taking a particular test at a given test administration.
- **Variability of the group tested.** In general, the larger the true variance or true spread of the scores of the candidate group (i.e., the greater the individual differences in the true level of knowledge and skills of the candidates in the particular group taking a test on a particular occasion), the greater will be the reliability coefficient. Reliability estimates tend to be higher if candidates in the group have widely varying levels of knowledge, and lower if they tend to have similar levels of knowledge.
- **Test length.** Statistical estimates of reliability are typically higher for tests with greater numbers of questions. A more reliable estimate of a person's knowledge is obtained by asking more questions.
- **Test content.** Reliability estimates are typically higher for tests that cover narrow, homogeneous content than for tests (such as many used for educator licensure) that cover a broad range of content. Tests for educator licensure must typically test a broad base of knowledge and skills that pertain to licenses that will apply in a wide range of educational settings, grade levels, and teaching assignments.

Because the tests included in the OAE program are used to make high-stakes decisions, several indicators of decision consistency (that is the degree to which the same decisions are made from two tests) and measures that indicate score reliability (consistency of scores across administrations) are calculated. Statistics presented not only consider the reliability of the test scores, but also indicate the reliability of the decisions made using the test results.

Several measures are employed to assess the reliability of each test in the OAE program. These measures are described below.

- **Livingston-Lewis estimate of decision consistency.** For a test used to make licensure requirement decisions such as the OAE, the consistency of such

decisions becomes a primary focus (Crocker & Algina, 1986). Decision consistency refers to the degree to which the same decisions are made from two tests. For the OAE program, the Livingston and Lewis (1995) estimate of decision consistency is used. This multi-stage method calculates decision consistency and accuracy using four types of input including: distribution of scores on one form, the minimum and maximum possible scores, the cut points used for classification, and the reliability coefficient (Livingston & Lewis, 1995). Decision consistency is reported in the range of 0 to 1, with estimates close to 1 indicating more consistent or reliable decisions.

- **Kuder-Richardson formula 20 (KR20).** The Kuder-Richardson index of item homogeneity (KR20) is an overall test consistency (reliability) estimate based on a single test administration (Kuder & Richardson, 1937). It is applicable to the multiple-choice section of tests. KR20 is reported in the range 0 to 1, with a higher number indicating a greater level of consistency (reliability). Homogeneity refers to the degree to which the items on the test are consistent with one another. For the OAE, KR20 is computed for tests composed of multiple-choice items only as well as for multiple-choice sections of tests that also include constructed-response items.
- **Stratified coefficient alpha.** Stratified coefficient alpha is an estimate of total test score reliability for a test containing a mixture of item types (e.g., multiple-choice and constructed-response) (Qualls, 1995). Each item type component of the test is treated as a subtest. Internal consistency estimates for the separate subtests are combined to compute stratified coefficient alpha. Stratified coefficient alpha is reported in the range 0 to 1, with a higher number indicating a greater level of consistency (reliability). This statistical estimate was deemed most appropriate for estimating total reliability of tests with both multiple-choice and constructed-response items for the OAE because it takes into account differences in test length and variance of the two item types.
- **Standard error of measurement.** The *Standards for Educational and Psychological Testing* define the standard error of measurement as the estimate of the difference between observed scores and estimated true scores by estimating the variability of measurement errors. This statistic speaks to the reliability of test scores, with smaller standard errors of measurement indicating more reliable test scores. (AERA, APA, & NCME, 2014).
- **Generalizability coefficient (G).** The Generalizability (G) coefficient is a measure of the percent of total score variance that is attributable to persons (i.e., factors within the candidate, such as subject matter knowledge). It reflects the proportion of variability in individuals' scores that is attributable to true score variability rather than to measurement error (Brennan, 2001). It is reported in the range 0 to 1, with a higher number indicating a greater level of generalizability. The G-coefficient is applicable to test sections composed of constructed-response items. It gauges the degree to which the results from one test form of the constructed-response items are generalizable to other forms, or other test administrations.
- **Scorer Agreement.** Scorer agreement is the degree of agreement between constructed-response scores assigned by independent scorers. Independent scorers are in agreement if the scores they award are either exact or adjacent. The scorers are not in agreement if the scores awarded differ by more than one point. The percent of cases in which the first two independent scorers are in

agreement is computed as a measure of scorer agreement (reliability). The following scorer agreement statistics are reported:

- *Percent Agreement.* Overall agreement determined by summing exact and adjacent agreement.
- *Percent Exact.* This is the percentage of scores in which the first two scorers were in exact agreement.
- *Percentage Adjacent.* This is the percentage of scores in which the two scorers assigned adjacent scores.
- *Inter-rater Reliability.* This is the intraclass correlation between the first and second score assigned to each response, corrected using the Spearman-Brown formula.

Validity

The *Standards for Educational and Psychological Testing* state that validity is a fundamental consideration in developing and evaluating tests (AERA, APA, & NCME, 2014). Validity relates to the use and interpretation of test scores rather than describing a test itself. For the OAE program, collection of both content-based and construct-based validity evidence is a continuous process.

Content-Based Validity Evidence

Because the OAE program is composed of licensure tests, gathering content-based validity evidence is essential to confirm that the assessment frameworks (which represent the content domain) are representative and relevant of the knowledge and skills required to be an entry-level educator in the classroom. Content-based validity evidence for the OAE program was gathered throughout the assessment validation process, starting with the frameworks.

OAE Frameworks. In validating the frameworks, content-based validity evidence is initially gathered through an alignment study. Alignment can be used to gather content-based validity evidence by corroborating that the knowledge, skills, and other constructs measured by the tests are consistent with those specified in the frameworks (Koretz & Hamilton, 2006). The OAE assessment frameworks are based on state-approved and nationally recognized professional and academic standards, and contain a competency component with information about the knowledge and/or skills necessary for performing the job of a licensed educator in Ohio public and non-public schools. Therefore, the competencies collectively define the range of content to be measured by the test. Pearson conducted an alignment study of each OAE framework to Ohio Educational Preparation Standards, Ohio Student Standards, and Ohio Educator Standards. These alignment studies were carried out to confirm that the test content, and therefore the tests are aligned with appropriate standards related to the intended purpose of the test. Additional information on the alignment studies can be found in the Ohio Assessments for Educators Technical Report, Development and Validation, 2011-2014.

Another source of content-based validity evidence was gathered by conducting an initial review of the frameworks by Ohio educators and teacher educators. During this review, Ohio educators and teacher educators checked that the content domain represented the knowledge and skills required by an entry-level educator in that test field by considering the criteria of alignment, completeness, clarity of language and terminology, and freedom from bias.

OAE Content Validation Surveys. Content Validation Surveys provide content-based validity evidence through the input of experts and stakeholders in educator licensure regarding the importance of the necessary knowledge and skills specified in each framework for an entry-level educator. Results of these surveys guided the final definition of the content domain. Practicing educators rated the test components of each framework on a 1–5 scale. Results of the surveys were used to determine the final status of all of the assessment framework components. Results of the content validation surveys for the OAE indicated that across test fields, almost all competencies and descriptive statements achieved a rating of at least 4.0, representing “great importance.” Additional information on the content validation surveys is included in the Ohio Assessment for Educators Technical Report, Development and Validation, 2011–2014.

OAE Test Items. As described in the Ohio Assessments for Educators Technical Report, Development and Validation, 2011–2014, the item validation activities by the Bias Review Committee (BRC) and Content Advisory Committees (CACs) provided additional content-based validity evidence. The BRC validated that the items were free from bias in content, language, offensiveness, and stereotypes, and that the items were fair and represented the diversity of the Ohio population. The CACs validated items to indicate that they matched the test competency or content domain to which they were written and were accurate, free from bias, and job-related. Any items that were judged to be problematic were eliminated.

Construct-Based Validity Evidence

The *Standards for Educational and Psychological Testing* refer to construct validity as the degree to which scores from an assessment can be interpreted as indicating the candidate’s standing on the knowledge and skills assessed by the test (AERA, APA, & NCME, 2014). Some threats to construct validity include construct irrelevance and construct underrepresentation. Construct irrelevance is “variance in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation” (AERA, APA, & NCME, 2014, p 217). Construct underrepresentation is defined as “the extent to which a test fails to capture important aspects of the construct domain that the test is intended to measure” (AERA, APA, & NCME, 2014, p. 217).

OAE Content Validation Surveys. For the OAE program, Pearson conducted content validation surveys that were aimed at gathering evidence to show that each assessment frameworks adequately reflects the knowledge and skills necessary for an entry-level educator in Ohio. Survey participants were asked to rate the importance of the knowledge and skills described by each competency and accompanying descriptive statements. The third item on the survey asked participants to rate a set of competencies in terms of how well they represented important aspects of the knowledge and skills required for performing the job of an entry-level educator. Results showed that the set of competencies for each test adequately represent the knowledge and skills the test is intended to measure.

OAE Test Items. The item validation activities by the BRC and CACs provided additional construct-based validity evidence. The BRC validated that the items were free from bias in content, language, offensiveness, and stereotypes and that the items were fair and represented the diversity of the Ohio population. The CACs validated items to indicate that they matched the test competency or content domain to which they were written and were accurate, free from bias, and job-related. Any items that were judged to be problematic were eliminated.

Score Reporting

After administration of the OAE tests, score reports are provided to candidates to inform them of their passing status and performance on the test. Score reports are also provided to the Ohio Department of Education (ODE) and to institutions, as designated by the candidate, to communicate passing status of individual candidates for educator licensure.

Candidate Score Reports

Candidates receive unofficial test results (reported as pass or not pass) on-site following the administration of multiple-choice-only tests. For tests that contain constructed-response items, candidates receive a receipt of completion on-site following the test administration. All candidates may request that a score report be e-mailed to the address provided during the registration process on the score report date published on the program website. Score reports are posted to the candidate's online account for 45 days as PDF documents, which a candidate may view, print, and save for their records. After the 45 days, candidates may request a copy of their scores through their online account.

Score reports include the following information:

- The date the candidate took the test
- The candidate's overall scaled score based on the number of items answered correctly converted to a scale ranging from 100 to 300
- The candidate's passing status based on the state-approved passing standard
- Details of candidate performance on each content domain assessed by the test for both multiple-choice and constructed-response sections

Candidates who do not pass a test are also provided with a Detailed Performance Summary that includes more detailed information about the candidate's performance on each test competency and constructed-response item, if applicable.

A sample candidate score report can be found in Appendix A. Score reports are accompanied by an interpretive guide to help candidates understand the reports. Additional information on how to read the score report can be found on the OAE program website (http://www.oh.nesinc.com/Content/Docs/OAE_ScoreReportBacker.pdf).

Other Score Recipients

Candidate scores are sent to the Ohio Department of Education (ODE) and institutions as designated by the candidate. These scores are delivered electronically through Pearson's secure web-based score-reporting system. Through this system, the ODE and institutions also have access to a web-based tool called *ResultsAnalyzer*®. This tool is interactive and allows ODE and institutions to view, analyze, reorganize, download, and print reports based on test data and generate customized reports of their choice.

ResultsAnalyzer® can support the needs of institutions of higher education through the following features:

- Create relevant and timely reports on candidate results
- Filter reports by candidate demographics, specific tests, or testing year

- Organize data into customized tables and graphs
- Customize data queries to align with institution goals and areas of interest
- Aggregate performance data across testing program years
- Export data to Microsoft® Excel or other report software and print graphics
- Analyze candidate data for numerous purposes, including Title II reporting, institutional research, program evaluation, and curriculum improvement

ResultsAnalyzer® also allows ODE to generate reports based on its needs and areas of focus using the features previously mentioned. In addition, *ResultsAnalyzer*® can create reports and filter results by institution, as well as support the analysis of institutional data. These data can assist ODE in assessing the performance of institutions across the state and help in future statewide policy decisions.

Test Statistics Reports

This document provides statistical reports for each test and reports on candidate performance for OAE tests administered during the 2014-2015 program year. Test statistics reports consist of test form and field statistics reports that outline statistical characteristics of multiple-choice only tests as well as tests composed of both multiple-choice and constructed-response items administered during the program year. Total scaled score distributions also are provided. Specific details of each report are outlined below.

Test Form Statistics Reports

The Test Form Statistics Report provided in Appendix B provides selected statistics for test forms administered to at least 100 candidates during the 2014-2015 program year. This report includes the following information:

- Test Field Code and Name
- Form Designation
- *Number of Tests Taken*. Scores are included in the analyses for this report for candidates who attempted at least one multiple-choice item and provided scorable responses to the constructed-response items, if applicable. In the cases in which the same test forms were administered at more than one test administration during the reporting period and candidates took the same test form at more than one test administration, both scores for those candidates' several attempts are included in the analyses.
- *Mean*. The mean of the total test scaled scores achieved by the candidates taking the test form.
- *Standard Error of Measurement (SEM)*. The standard error of measurement of the total test scaled scores is reported.
- *Decision Consistency*. The Livingston Lewis estimate of decision consistency that is reported value is based on all administrations of a test form during the program year. Test forms are considered to be identical if they contain identical sets of scorable multiple-choice items, regardless of the order of the items.
- *Stratified Alpha*. This statistic provides an estimate of total test score reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items) (Qualls, 1995).
- *Test Length*. The number of scorable items of the multiple-choice section.

- Kuder-Richardson formula 20 (KR20)

A second Test Form Statistics Report is provided in Appendix C. This report contains the total test scaled score mean and the number of scorable multiple-choice items for test forms that were administered to less than 100 candidates during the 2014-2015 program year. Statistical indices such as the SEM or reliability coefficients cannot be confidently interpreted when computed using data from small groups of candidates; therefore, such analyses were omitted for test forms administered to less than 100 candidates in the reporting period.

Test Field Statistics Report

The Test Statistics Report by Test Field shown in Appendix D provides selected statistics for test fields comprised of constructed-response items administered to at least 100 candidates during the program year. This report includes the following information:

- **Scorer Agreement.** Measures of scorer agreement reported are:
 - *Percent Agreement.* Overall agreement determined by summing exact and adjacent agreement.
 - *Percent Exact.* This is the percentage of scores in which the first two scorers were in exact agreement.
 - *Percentage Adjacent.* This is the percentage of scores in which the two scorers assigned adjacent scores.
 - *Inter-rater Reliability.* This is the intraclass correlation between the first and second score assigned to each response, corrected using the Spearman-Brown formula.

Total Scaled Score Distribution by Test Field

The Total Scaled Score Distribution by Test field shown in Appendix E provides information about the scaled score distributions associated with all test fields taken by 10 or more candidates. For the OAE, results are reported on a scale ranging from 100 to 300. A scaled score of 220 represents the passing standard for each test.

The scale score distributions include the following information:

- Total Scaled Score: the inclusive list of observed total test scaled scores, in intervals of five scale-score points.
- N: the number of scores observed within each interval.
- N at or above: the number of scores observed at or above each scale score interval.
- Percent: the percent of scores observed within each interval.
- Percent at or above: the percent of scores observed at or above each scaled score interval.

Title II Reporting

The federal government requires that states submit annual reports on quality of their teacher preparation programs for accountability purposes. The reporting requirements are spelled out in sections 205 through 208 of Title II of the Higher Education Act (HEA). The act, which was amended in 2008, was put in place to encourage transparency and accountability among states. The data are intended to inform stakeholders on the quality of teacher preparation.

Pearson helps Ohio meet Title II requirements by working with both ODE and institutions of higher education to produce the reports. The Title II reporting system allows the submission, review, and transmission of data that supports collaboration among institutions and ODE. Pearson supports the reporting needs and requirements through the following actions:

- producing Test Pass Rates Tables and Summary Pass Rates Tables for Title II Reporting.
- providing web-based training to institutions of higher education on the purpose and process for data collection.
- maintaining a secure, password-accessible website for data collection.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.
- Angoff, W. H. (1984). *Scales, Norms and Equivalent Scores*. Princeton, NJ: Educational Testing Service.
- Berk, R. A. (1980). A consumer's guide to criterion-referenced test reliability. *Journal of Educational Measurement* 14(4):323–349.
- Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: Springer-Verlag.
- Crocker, L., and J. Algina. (1986). *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadsworth Group.
- Kolen, M. J. and Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer Science and Business Media, LLC.
- Koretz, D. M. and Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 531–578). Westport, CT: American Council on Education and Praeger Publishers.
- Kuder, G. F. and Richardson, M. W. (1937) The theory of the estimation of test reliability. *Psychometrika* 2:151–160.
- Livingston, S. A., and Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement* 32(2):179–197.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed.) (pp. 13–103). Washington, DC: American Council on Education.
- Qualls, L. A. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education* 8(2):111–120.

Appendices

Appendix A: Sample OAE Candidate Score Report

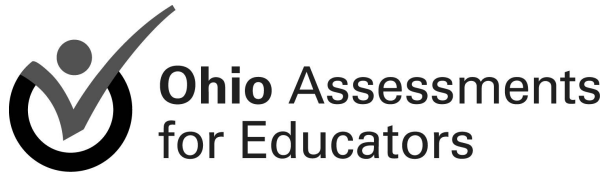
Appendix B: Test Form Statistics Report for Test Fields with More than 100 Tests Taken

Appendix C: Test Form Statistics Report for Test Fields with Less than 100 Tests Taken

Appendix D: Test Field Statistics Report

Appendix E: Total Scaled Score Distribution by Test Field

**Appendix A:
Sample OAE Candidate Score Report**



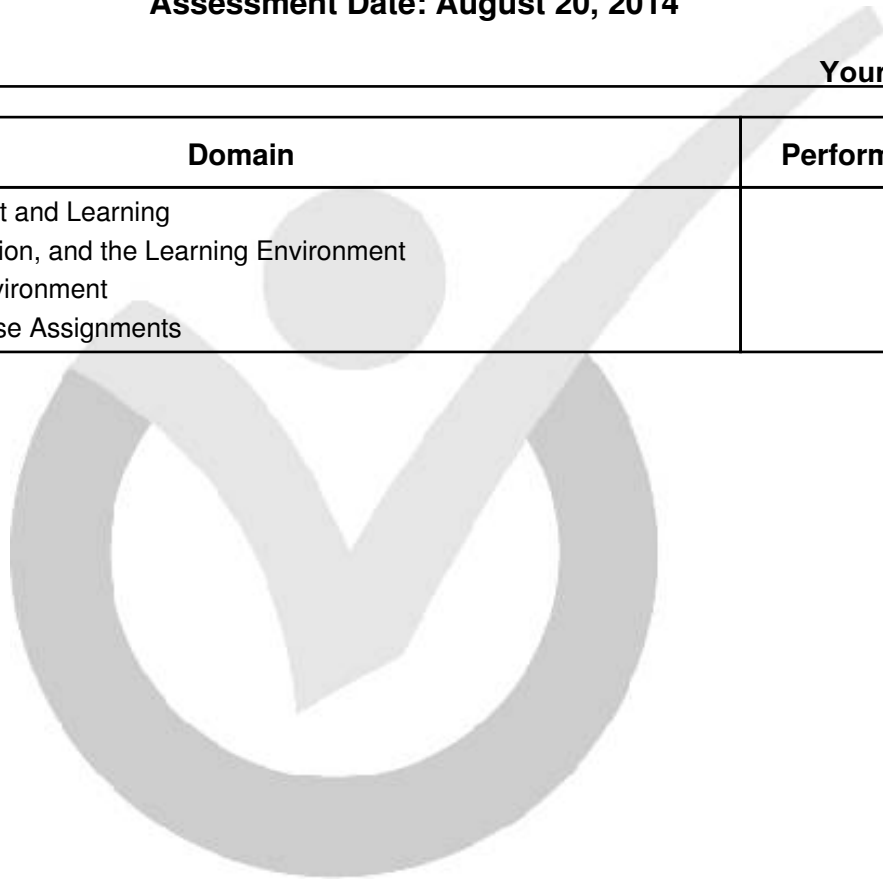
FirstName LastName
300 Venture Way
Sample City, OH XXXXX

Assessment of Professional Knowledge: Early Childhood (PK-3) 001
Minimum Passing Score: 220
Assessment Date: August 20, 2014

Status: Pass

Your Score: 231

Domain	Performance Index
Student Development and Learning	++
Assessment, Instruction, and the Learning Environment	+++
The Professional Environment	+++
Constructed Response Assignments	+++

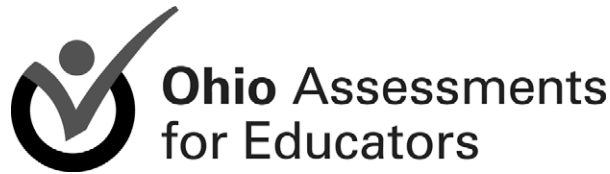


Examinee Name: FirstName LastName

Social Security Number: XXX-XX-XXXX



This barcode contains unique candidate information.



HOW TO READ YOUR SCORE REPORT

Overview. This score report provides your assessment results for the Ohio Assessments for Educators (OAE) that you took on the assessment date indicated on the report. The purpose of the OAE is to assess the knowledge and skills of a prospective Ohio educator. The OAE program helps the Ohio Department of Education and the State Board of Education of Ohio meet their goal of ensuring that candidates have the knowledge and skills needed to perform the job of an educator in Ohio schools.

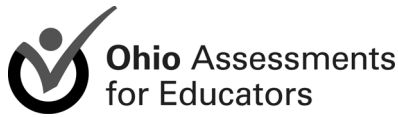
Assessment Score. Your total assessment score is a scaled score. A scaled score is a combination of the number of scorable questions you answered correctly on the multiple-choice section of the assessment and the scores you received on any constructed-response assignments (if applicable) converted to a scale from 100 to 300, with a score of 220 representing the passing score for the assessment. The number of questions per competency may vary by test form. The total number of questions needed to pass the assessment may also vary across test forms as a result of variations in question difficulty. The conversion of raw scores to scaled scores takes such differences into account. The passing score for each assessment is established by the State Board of Education of Ohio and is based on the professional judgments and recommendations of Ohio educators. "Pass" or "Did Not Pass" status is based on your total score for each assessment.

Performance Indices. Performance indices are provided for each domain of the assessment. These indices can help you understand your areas of strength and weakness. This information should be interpreted with caution, since different domains contain different numbers of assessment questions. For each domain, your performance is reported according to the following performance indices:

Performance Indices: Multiple-Choice Questions	
++++	Performance on the competency is well above the level represented by the minimum passing score.
+++	Performance on the competency is just at or above the level represented by the minimum passing score.
++	Performance on the competency is just below the level represented by the minimum passing score.
+	Performance on the competency is well below the level represented by the minimum passing score.

Performance Indices: Constructed-Response Assignments			
Constructed-Response Assignments (Assessments with a 4-Point Scoring Scale*)		Constructed-Response Assignments (Assessments with a 3-Point Scoring Scale*)	
++++	Response reflects a thorough understanding	+++	Response reflects a thorough understanding
+++	Response reflects a general understanding	++	Response reflects a general understanding
++	Response reflects a limited understanding	+	Response reflects limited or no understanding
+	Response reflects little or no understanding		

*See study guides at www.oh.nesinc.com



Each assessment framework describes the content knowledge assessed by the OAE. You may view, print, or download the framework for any OAE assessment by selecting "Assessment Frameworks" in the "Prepare" section of the OAE website at www.oh.nesinc.com.

If a response to a constructed-response assignment is designated "Blank" or "Unscorable," you will see one of the following codes:

Codes for Blank/Unscorable Responses	
B	Response was blank
U-1	Response was unrelated to assigned topic
U-2	Response was unintelligible
U-3	Response was not primarily in English
U-4	Response lacked sufficient amount of original work

Summary Report for Assessments with Two Subtests. Some OAE assessments are comprised of two subtests. If you took a two-subtest assessment, your score report will include an assessment summary. The assessment summary lists passing status information for each subtest within the assessment and includes the date on which you passed each subtest. Please note that for assessments composed of more than one subtest, you must pass both subtests for that assessment to meet Ohio certification requirements.

Reporting of Scores. Your scores are reported directly to the Ohio Department of Education and the Ohio educator preparation institution(s) you indicated during the registration process. This score report is for your information only. Keep a copy for your permanent records.

ADDITIONAL INFORMATION

Retaking the Assessment. You may retake an OAE assessment by following the same registration procedures you completed for previous assessment administrations. For information about retake policies, visit www.oh.nesinc.com.

Assessment Preparation Materials. Assessment preparation materials are available through the website you used to register for this assessment.

**Appendix B:
Test Form Statistics Report for Test Fields with
More than 100 Tests Taken**

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
001 Assessment of Professional Knowledge: Early Childhood (PK-3)	A	743	237	13.1	0.87	0.72	80	0.73	0.22
	B	822	238	12.9	0.89	0.76	80	0.78	0.17
	C	452	238	12.8	0.87	0.81	80	0.83	0.22
	D	448	239	12.8	0.90	0.80	80	0.82	0.17
002 Assessment of Professional Knowledge: Middle Childhood (4-9)	A	114	231	15.1	0.83	0.75	80	0.76	0.42
	B	519	236	15.0	0.83	0.79	80	0.79	0.45
	C	505	235	14.8	0.85	0.77	80	0.77	0.41

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
003 Assessment of Professional Knowledge: Adolescence to Young Adult (7-12)	A	140	257	12.4	0.98	0.76	80	0.71	0.52
	B	633	252	12.1	0.96	0.80	80	0.78	0.44
	C	694	252	12.1	0.96	0.79	80	0.77	0.45
004 Assessment of Professional Knowledge: Multi-Age (PK-12)	A	207	240	14.2	0.89	0.78	80	0.82	0.25
	B	1,006	241	13.6	0.90	0.83	80	0.84	0.39
	C	964	242	13.6	0.90	0.82	80	0.83	0.38

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
006 Art	A	151	231	10.5	0.83	--	120	0.83	--
	B	41	231	--	--	--	120	--	--
	C	51	231	--	--	--	120	--	--
007 Biology	A	116	224	10.3	0.85	--	120	0.90	--
	B	41	209	--	--	--	120	--	--
	C	56	220	--	--	--	120	--	--
012 Early Childhood Education	A	701	251	7.5	0.97	--	120	0.80	--
	B	683	249	7.5	0.96	--	120	0.83	--
	C	387	249	7.3	0.95	--	120	0.89	--
	D	389	250	7.3	0.93	--	120	0.87	--

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
013 Early Childhood Special Education	A	124	222	13.2	0.72	--	80	0.63	--
	B	205	223	13.2	0.72	--	80	0.71	--
	C	198	218	14.0	0.73	--	80	0.70	--
015 Educational Leadership	A	120	232	10.4	0.85	0.65	60	0.66	0.50
	B	538	225	10.2	0.81	0.73	60	0.69	0.60
	C	508	224	10.3	0.82	0.69	60	0.69	0.55
018 Elementary Education Subtest I	A	102	232	12.6	0.82	--	60	0.68	--
	B	377	240	12.2	0.84	--	60	0.77	--
	C	374	238	12.3	0.82	--	60	0.72	--

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
019 Elementary Education Subtest II	A	128	220	12.8	0.85	--	60	0.85	--
	B	452	231	13.1	0.78	--	60	0.80	--
	C	403	233	13.0	0.81	--	60	0.79	--
020 English Language Arts	A	69	222	--	--	--	120	--	--
	B	372	225	11.0	0.84	--	120	0.87	--
	C	371	222	11.1	0.82	--	120	0.87	--
021 English to Speakers of Other Languages (ESOL)	A	198	247	8.1	0.92	--	120	0.87	--
023 Health	A	151	242	10.9	0.83	--	80	0.79	--
024 Integrated Science	A	191	246	6.7	0.93	--	120	0.89	--

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
025 Integrated Social Studies	A	63	232	--	--	--	120	--	--
	B	304	230	7.6	0.84	--	120	0.89	--
	C	285	230	7.6	0.87	--	120	0.90	--
027 Mathematics	A	71	227	--	--	--	120	--	--
	B	275	223	8.2	0.85	--	120	0.93	--
	C	254	222	8.1	0.90	--	120	0.94	--
028 Middle Grades English Language Arts	A	63	228	--	--	--	120	--	--
	B	330	226	11.1	0.86	--	120	0.85	--
	C	348	228	11.0	0.85	--	120	0.85	--

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
029 Middle Grades Science	A	72	223	--	--	--	120	--	--
	B	354	224	8.3	0.82	--	120	0.87	--
	C	353	224	8.3	0.84	--	120	0.86	--
030 Middle Grades Mathematics	A	80	217	--	--	--	120	--	--
	B	413	223	7.8	0.85	--	120	0.89	--
	C	433	223	7.8	0.86	--	120	0.91	--
031 Middle Grades Social Studies	A	96	219	--	--	--	120	--	--
	B	383	217	8.2	0.81	--	120	0.85	--
	C	380	218	8.2	0.84	--	120	0.87	--

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
032 Music	A	81	236	--	--	--	120	--	--
	B	163	240	10.9	0.85	--	120	0.80	--
	C	140	236	11.0	0.85	--	120	0.82	--
034 Physical Education	A	149	227	9.1	0.84	--	120	0.86	--
	B	45	228	--	--	--	120	--	--
	C	48	220	--	--	--	120	--	--
038 Reading Subtest I	A	225	236	14.5	0.85	0.78	54	0.78	0.31
	B	648	244	14.8	0.88	0.78	54	0.78	0.25
	C	634	244	14.7	0.88	0.78	54	0.79	0.29

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

Test Form Statistics Report For Test Fields With More Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Indices				Multiple-Choice Section		Open Response Section
			Mean ^a	SEM ^b	Decision Consistency ^c	Stratified Alpha ^d	Length ^e	KR20 ^f	G Coefficient ^g
Test Field [*]	Form ^{**}								
039 Reading Subtest II	A	208	242	14.9	0.87	0.75	54	0.76	0.32
	B	624	246	14.7	0.90	0.77	54	0.77	0.40
	C	594	248	14.8	0.90	0.77	54	0.77	0.39
040 School Counselor	A	27	227	--	--	--	116	--	--
	B	176	235	10.8	0.84	--	120	0.77	--
	C	149	234	10.8	0.81	--	120	0.76	--
043 Special Education	A	519	231	9.8	0.85	--	120	0.87	--
	B	503	232	10.1	0.83	--	120	0.85	--
	C	424	231	9.7	0.84	--	120	0.89	--
	D	436	232	9.8	0.85	--	120	0.88	--

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively

a - the mean total test scaled scores achieved by the candidates taking the test form

b - the standard error of measurement of the total test scaled scores

c - a reliability statistic that describes the consistency of the pass/fail decision on the total test scaled score

d - an estimate of total test reliability for a test containing multiple item types (e.g., multiple-choice, constructed-response items)

e - the number of scorable items in the multiple-choice section

f - an overall test consistency (reliability) estimate based on a single test administration (Kuder-Richardson formula 20)

g - a measure of the proportion of total score variance that is attributable to true score variability rather than to measurement error (Generalizability Coefficient)

**Appendix C:
Test Form Statistics Report for Test Fields with
Less than 100 Tests Taken**

Test Form Statistics Report For Test Fields With Fewer Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Mean ^a	Multiple-Choice Section Length ^b
Test Field [*]	Form ^{**}			
005 Agriscience	A	15	248	80
008 Business Education	A	74	228	120
009 Chemistry	A	92	233	120
010 Computer Information Science	A	19	199	64
014 Earth and Space Science	A	35	232	120
016 Computer/Technology Subtest I	A	58	259	56
017 Computer/Technology Subtest II	A	62	252	32
022 Family and Consumer Sciences	A	34	219	120
026 Marketing	A	3	--	80
035 Physics	A	47	242	120
036 Prekindergarten Subtest I	A	40	230	54
037 Prekindergarten Subtest II	A	38	229	54
041 School Library Media Specialist	A	46	227	120
042 School Psychologist	A	2	--	64
044 Special Education Specialist: Deaf/Hard of Hearing	A	16	237	80
045 Special Education Specialist: Visually Impaired	A	12	219	64

(Continued)

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively
a - the mean total test scaled scores achieved by the candidates taking the test form
b - the number of scorable items in the multiple-choice section

Test Form Statistics Report For Test Fields With Fewer Than 100 Tests Taken Per Test Form

September 1, 2014 - August 30, 2015		Number of Tests Taken ^{***}	Total Test Scaled Score Mean ^a	Multiple-Choice Section Length ^b
Test Field [*]	Form ^{**}			
046 Technology Education Subtest I	A	8	--	36
047 Technology Education Subtest II	A	6	--	60
048 Theater	A	21	236	85

NOTES: *, **, *** - test field, test form administered, and the number of times the test form was taken, respectively
a - the mean total test scaled scores achieved by the candidates taking the test form
b - the number of scorable items in the multiple-choice section

**Appendix D:
Test Field Statistics Report**

Test Field Statistics Report For Test Fields With Open-Response Items

Sep 1, 2014 - Aug 31, 2015	Number of Tests Taken ^a	Open-Response Prompts							
		Item 1 of 4				Item 2 of 4			
		Scorer Agreement			Inter-rater Reliability ^e	Scorer Agreement			Inter-rater Reliability ^e
		Percent Agreement ^b	Percent Exact ^c	Percent Adjacent ^d		Percent Agreement ^b	Percent Exact ^c	Percent Adjacent ^d	
001 Assessment of Professional Knowledge: Early Childhood (PK-3)	2,465	98.0	58.5	39.5	0.66	97.5	56.4	41.0	0.79
002 Assessment of Professional Knowledge: Middle Childhood (4-9)	1,138	97.1	56.6	40.5	0.72	96.2	54.0	42.2	0.76
003 Assessment of Professional Knowledge: Adolescence to Young Adult (7-12)	1,467	98.0	58.9	39.0	0.65	97.7	58.6	39.0	0.67
004 Assessment of Professional Knowledge: Multi-Age (PK-12)	2,177	96.1	51.1	45.0	0.67	95.3	56.1	39.2	0.79
015 Educational Leadership	1,166	95.2	51.4	43.8	0.60	95.5	48.5	47.0	0.59
038 Reading Subtest I	1,507	98.2	64.1	34.1	0.77	98.1	59.9	38.3	0.69
039 Reading Subtest II	1,426	98.5	64.9	33.6	0.74	97.9	62.3	35.6	0.71

NOTES: *a* - the number of times the test was taken

b - overall agreement determined by summing exact and adjacent agreement

c - the percentage of scores in which the first two scorers were in exact agreement

d - the percentage of scores in which the two scorers assigned adjacent scores

e - the intraclass correlation between the first and second score assigned to each response, corrected using the Spearman-Brown formula

Test Field Statistics Report For Test Fields With Open-Response Items

Sep 1, 2014 - Aug 31, 2015	Open-Response Prompts							
	Item 3 of 4				Item 4 of 4			
	Scorer Agreement			Inter-rater Reliability ^e	Scorer Agreement			Inter-rater Reliability ^e
	Percent Agreement ^b	Percent Exact ^c	Percent Adjacent ^d		Percent Agreement ^b	Percent Exact ^c	Percent Adjacent ^d	
001 Assessment of Professional Knowledge: Early Childhood (PK-3)	--	--	--	--	--	--	--	--
002 Assessment of Professional Knowledge: Middle Childhood (4-9)	--	--	--	--	--	--	--	--
003 Assessment of Professional Knowledge: Adolescence to Young Adult (7-12)	--	--	--	--	--	--	--	--
004 Assessment of Professional Knowledge: Multi-Age (PK-12)	--	--	--	--	--	--	--	--
015 Educational Leadership	95.8	56.0	39.8	0.66	97.0	53.6	43.4	0.60
038 Reading Subtest I	--	--	--	--	--	--	--	--
039 Reading Subtest II	--	--	--	--	--	--	--	--

NOTES: *a - the number of times the test was taken*

b - overall agreement determined by summing exact and adjacent agreement

c - the percentage of scores in which the first two scorers were in exact agreement

d - the percentage of scores in which the two scorers assigned adjacent scores

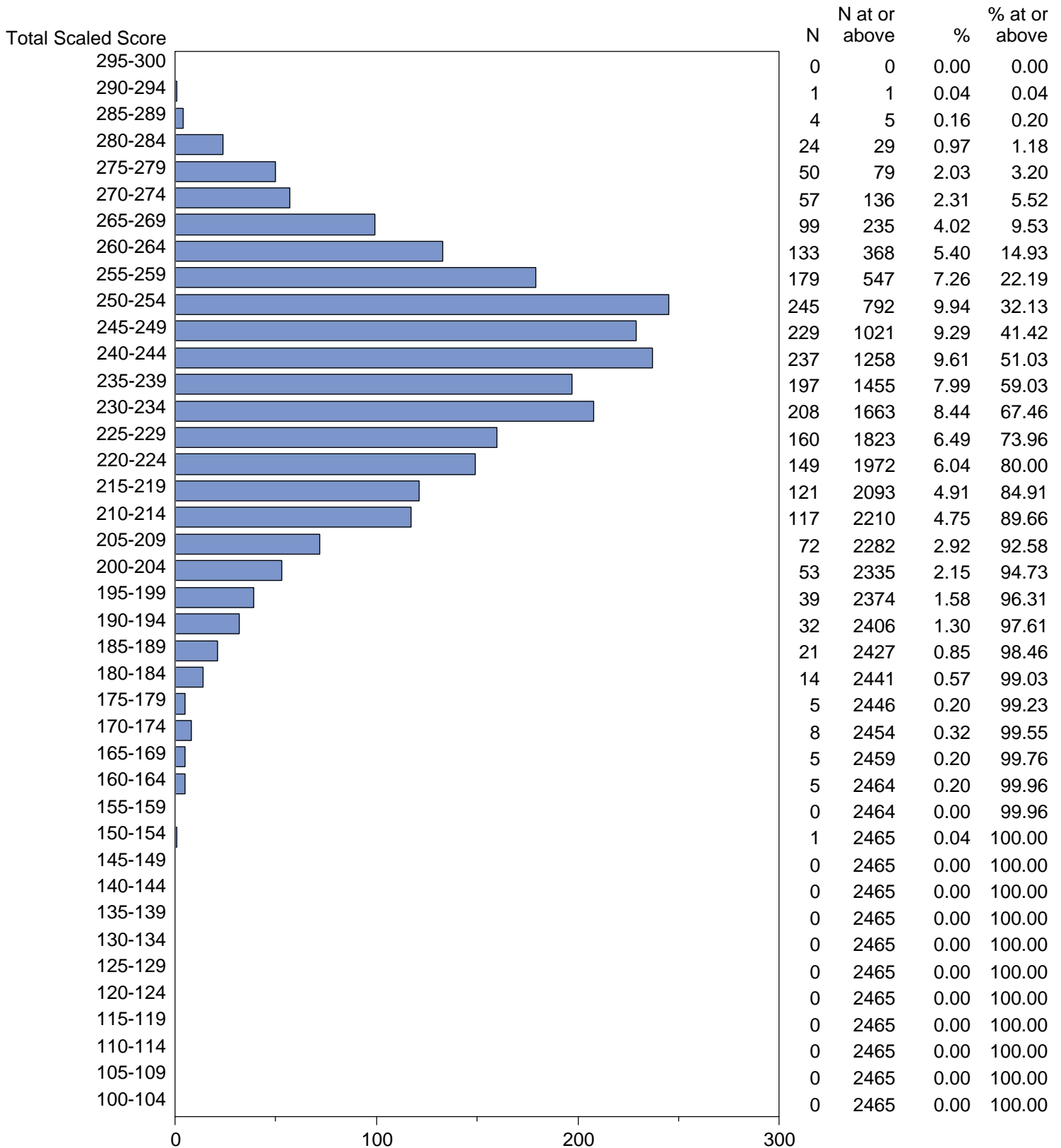
e - the intraclass correlation between the first and second score assigned to each response, corrected using the Spearman-Brown formula

**Appendix E:
Total Scaled Score Distribution by Test Field**

Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

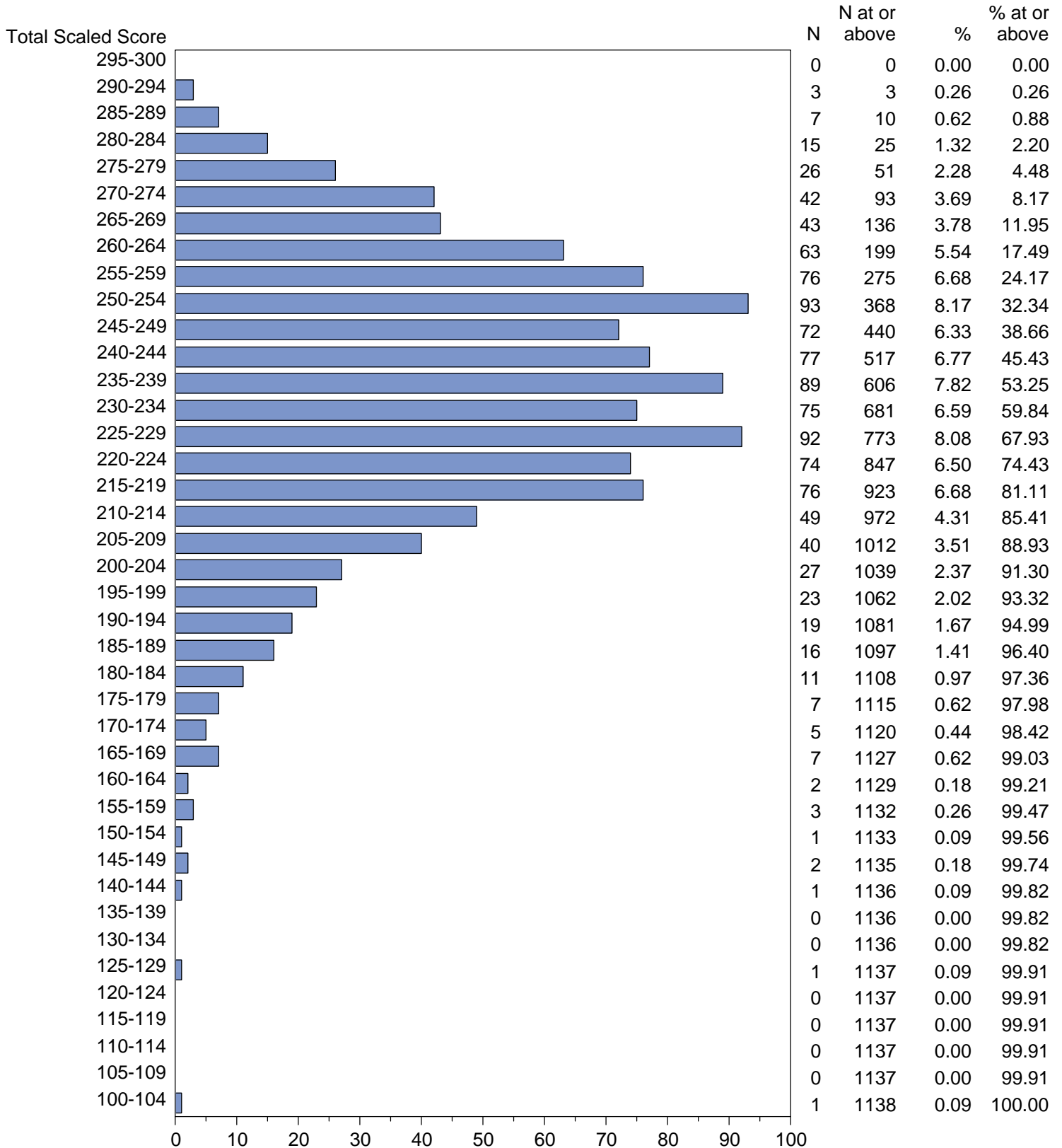
Test Field=001 Assessment of Professional Knowledge: Early Childhood (PK-3)



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

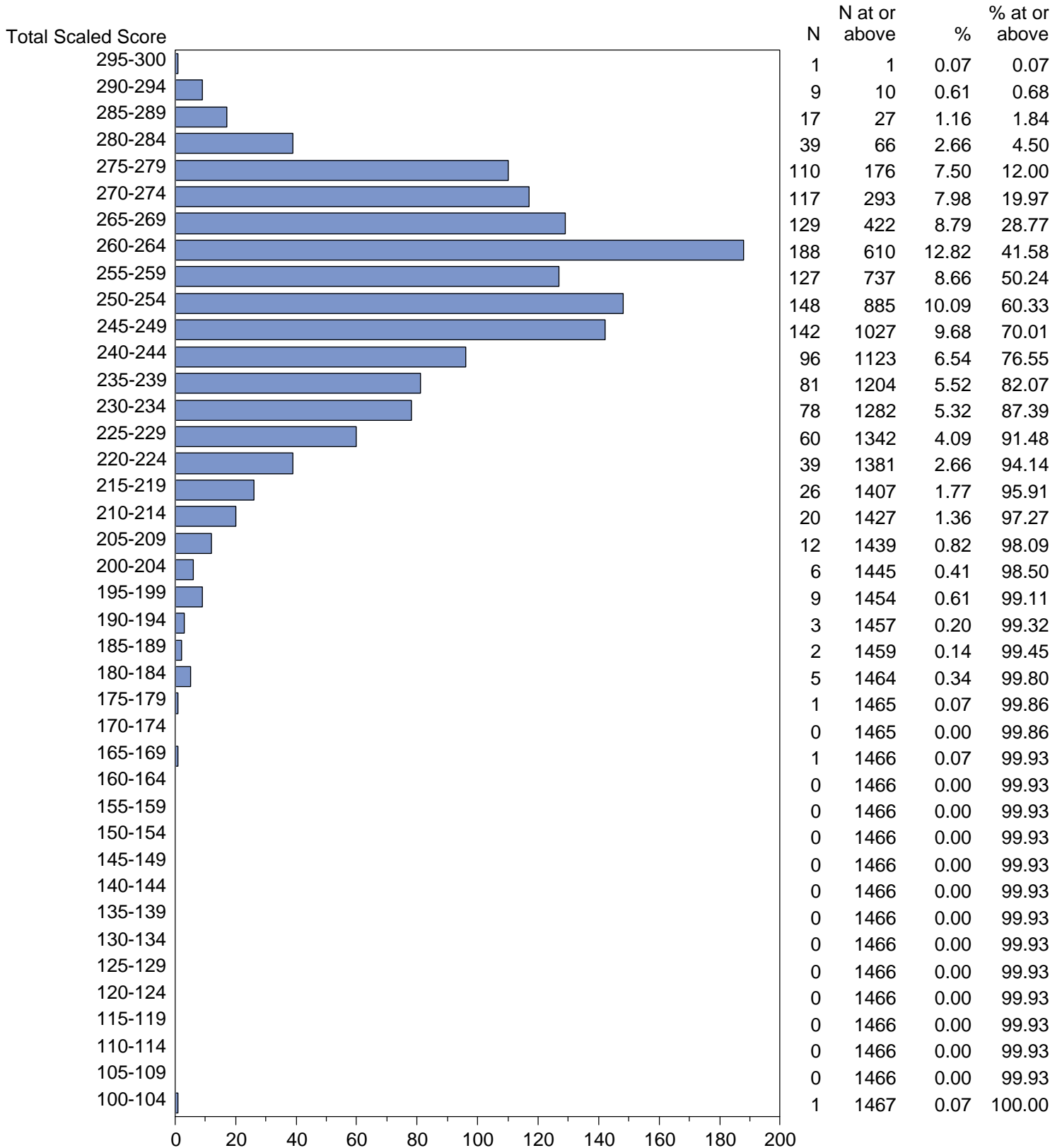
Test Field=002 Assessment of Professional Knowledge: Middle Childhood (4-9)



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

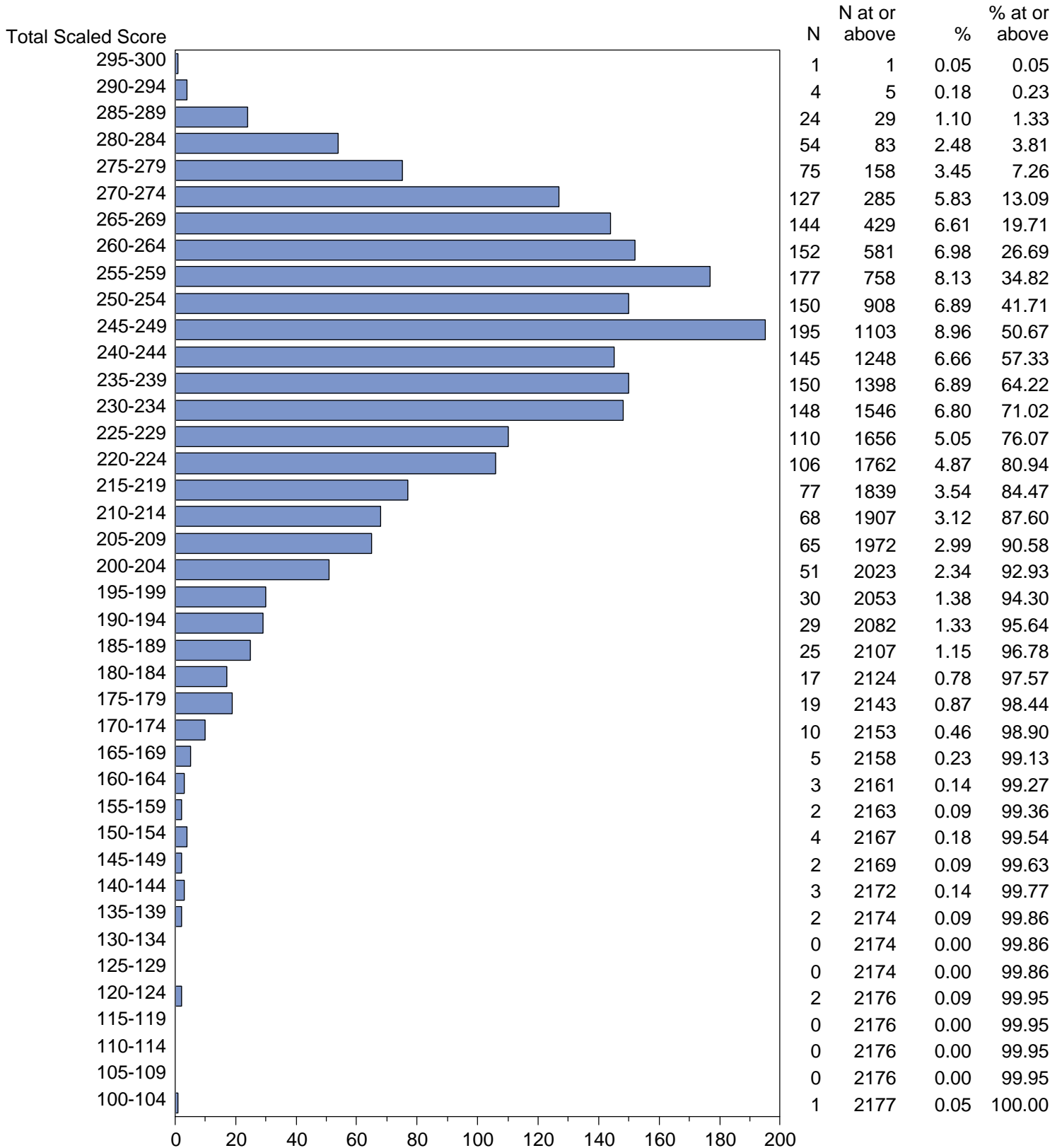
Test Field=003 Assessment of Professional Knowledge: Adolescence to Young Adult (7-12)



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

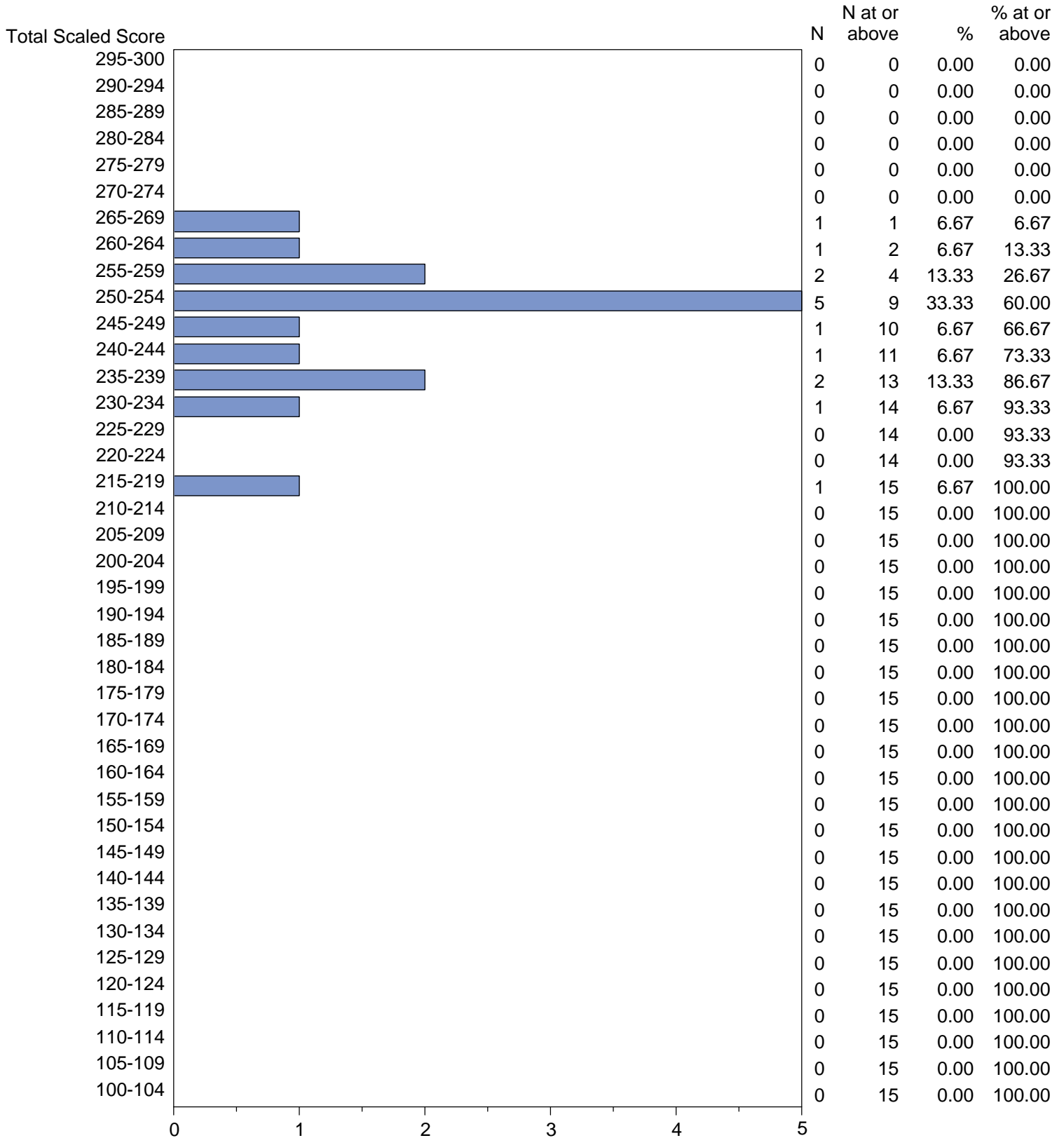
Test Field=004 Assessment of Professional Knowledge: Multi-Age (PK-12)



Candidate Performance Statistical Report
 September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

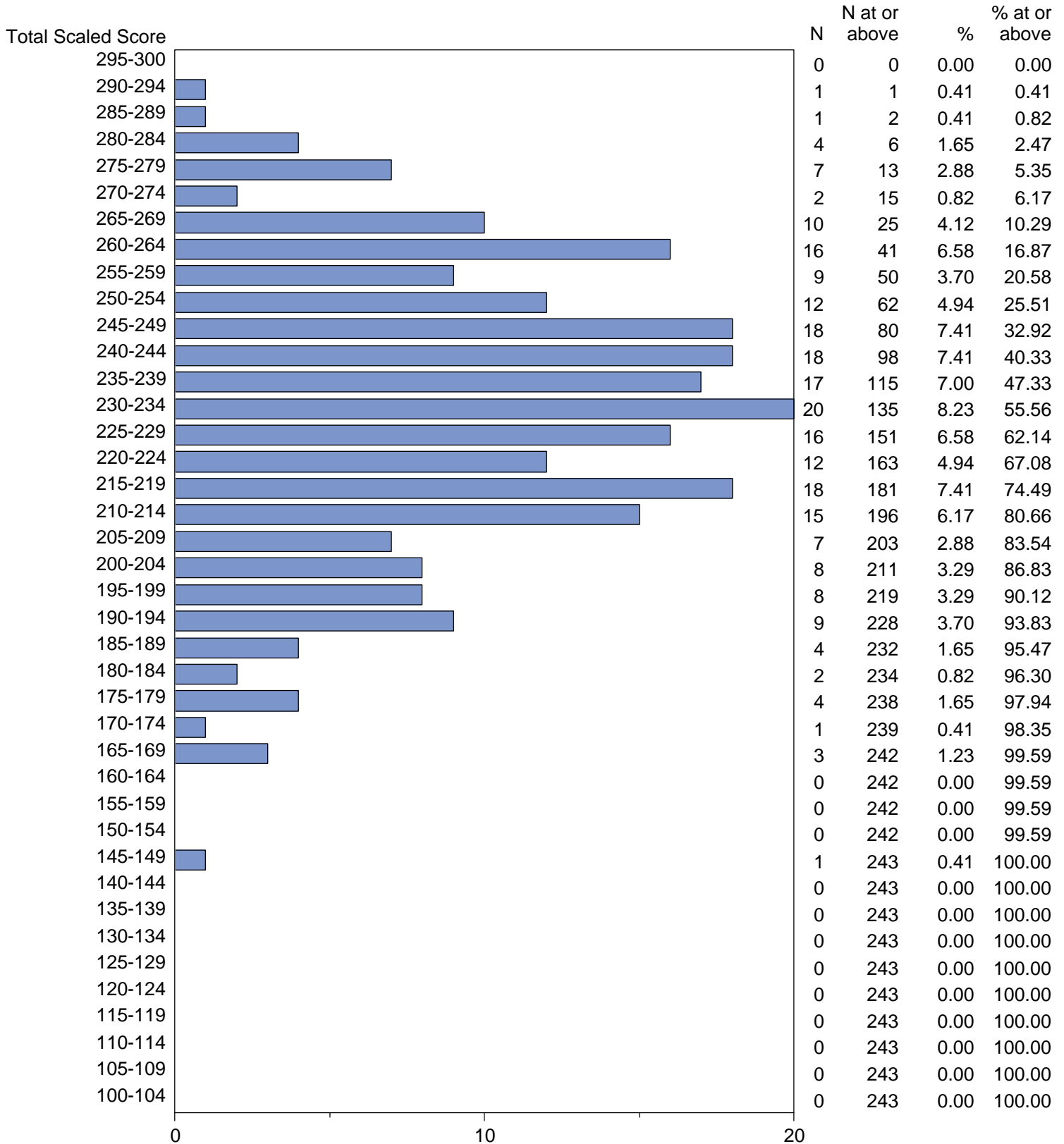
Test Field=005 Agriscience



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

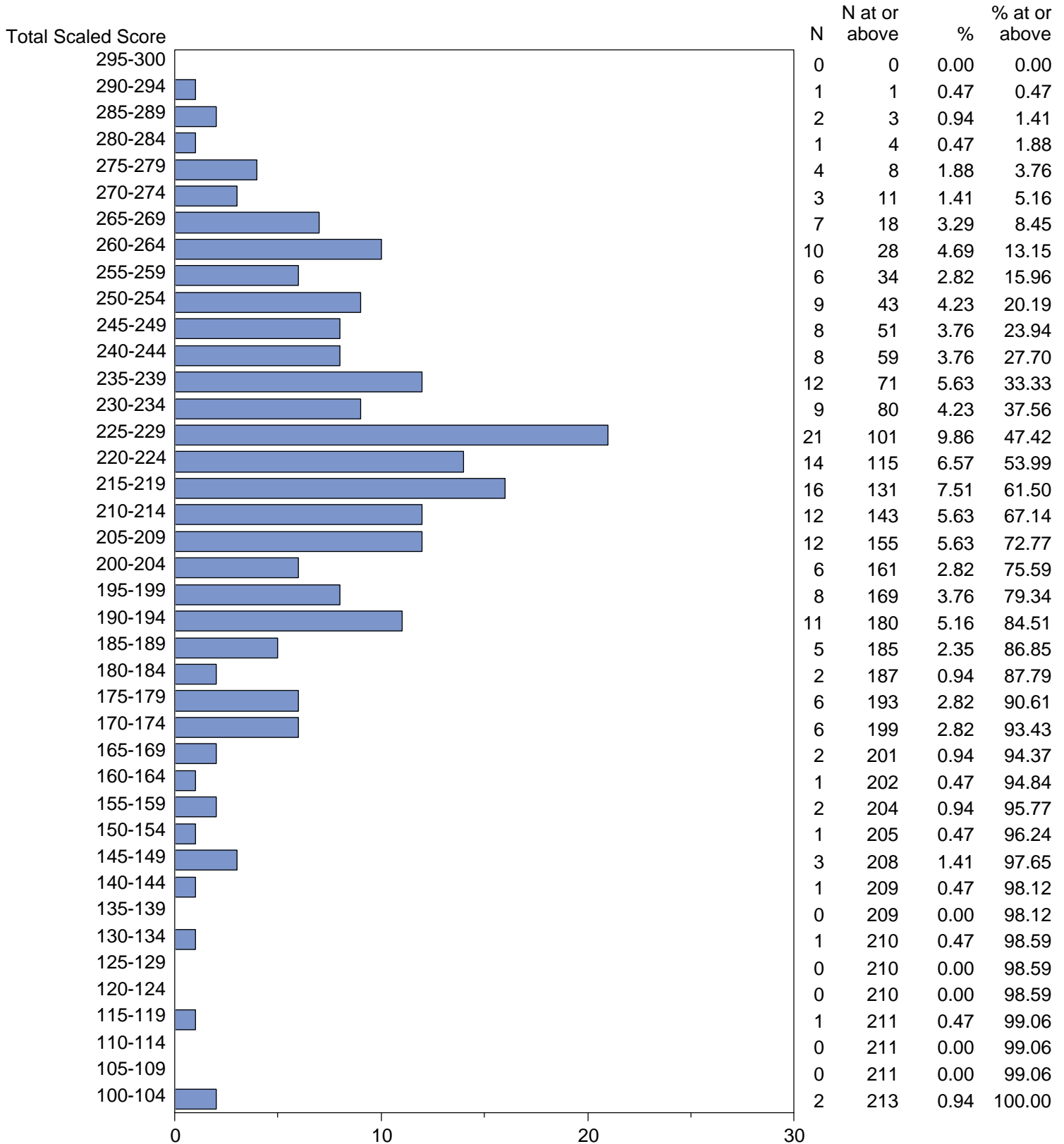
Test Field=006 Art



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

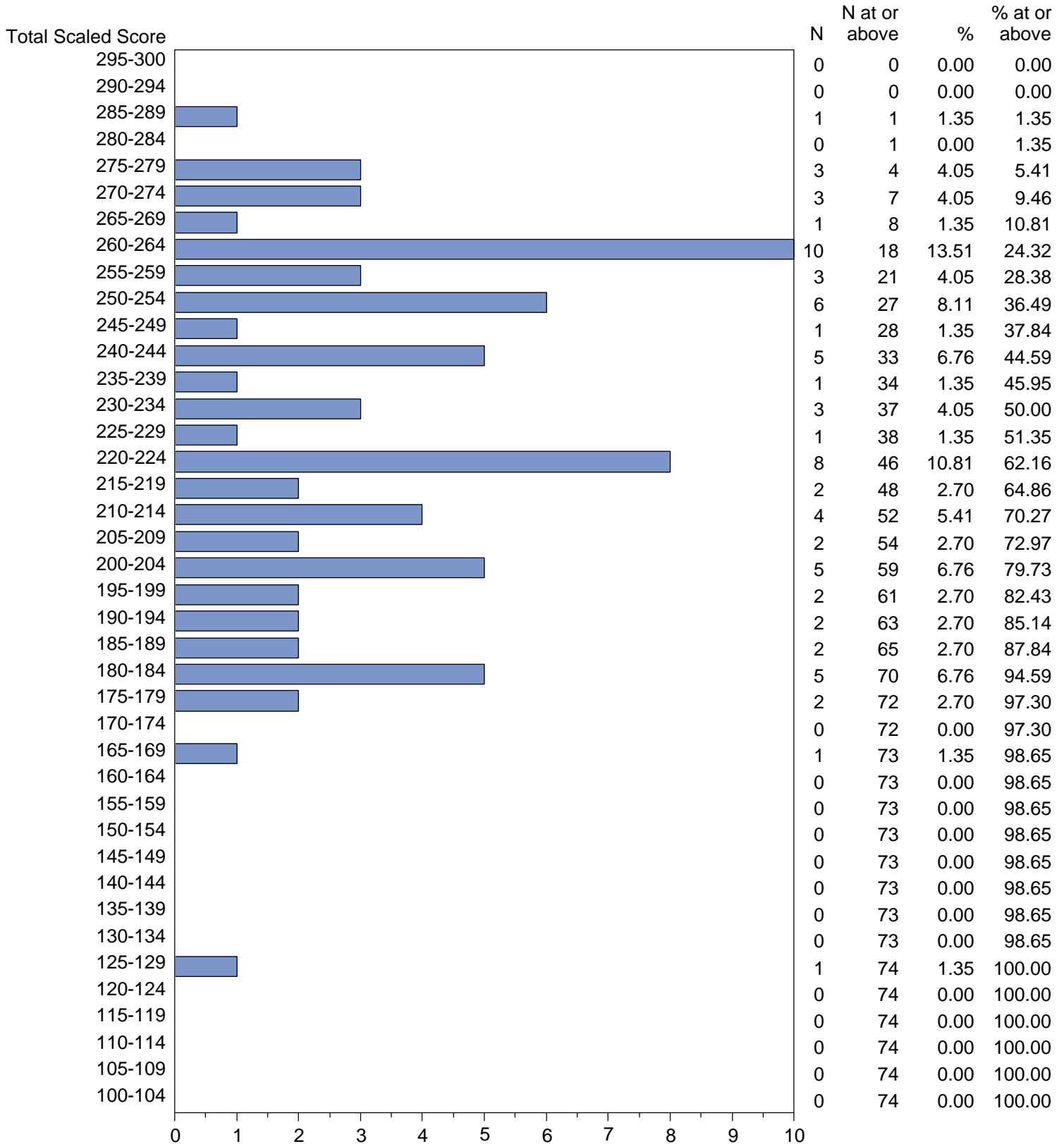
Test Field=007 Biology



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

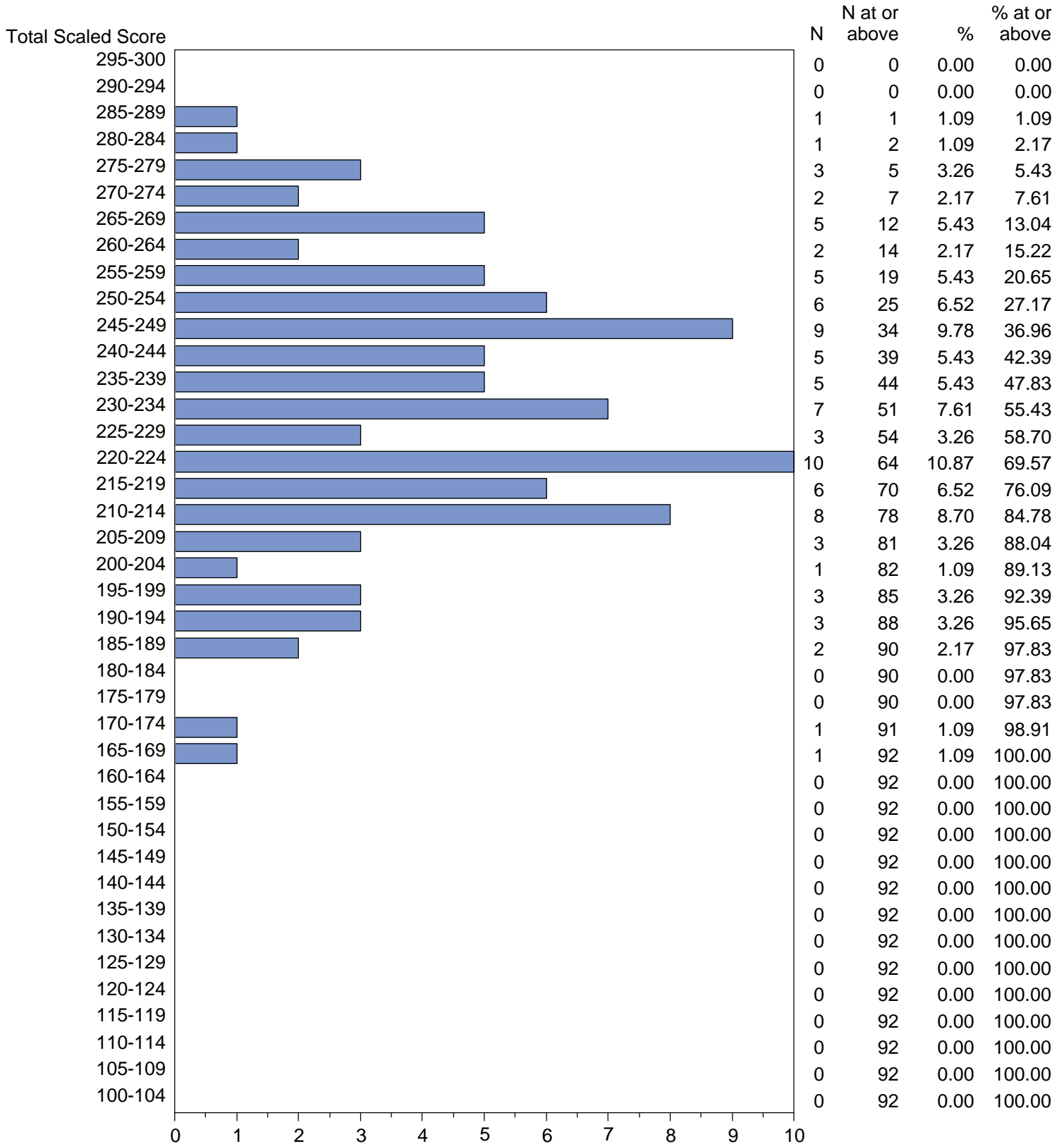
Test Field=008 Business Education



Candidate Performance Statistical Report
 September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

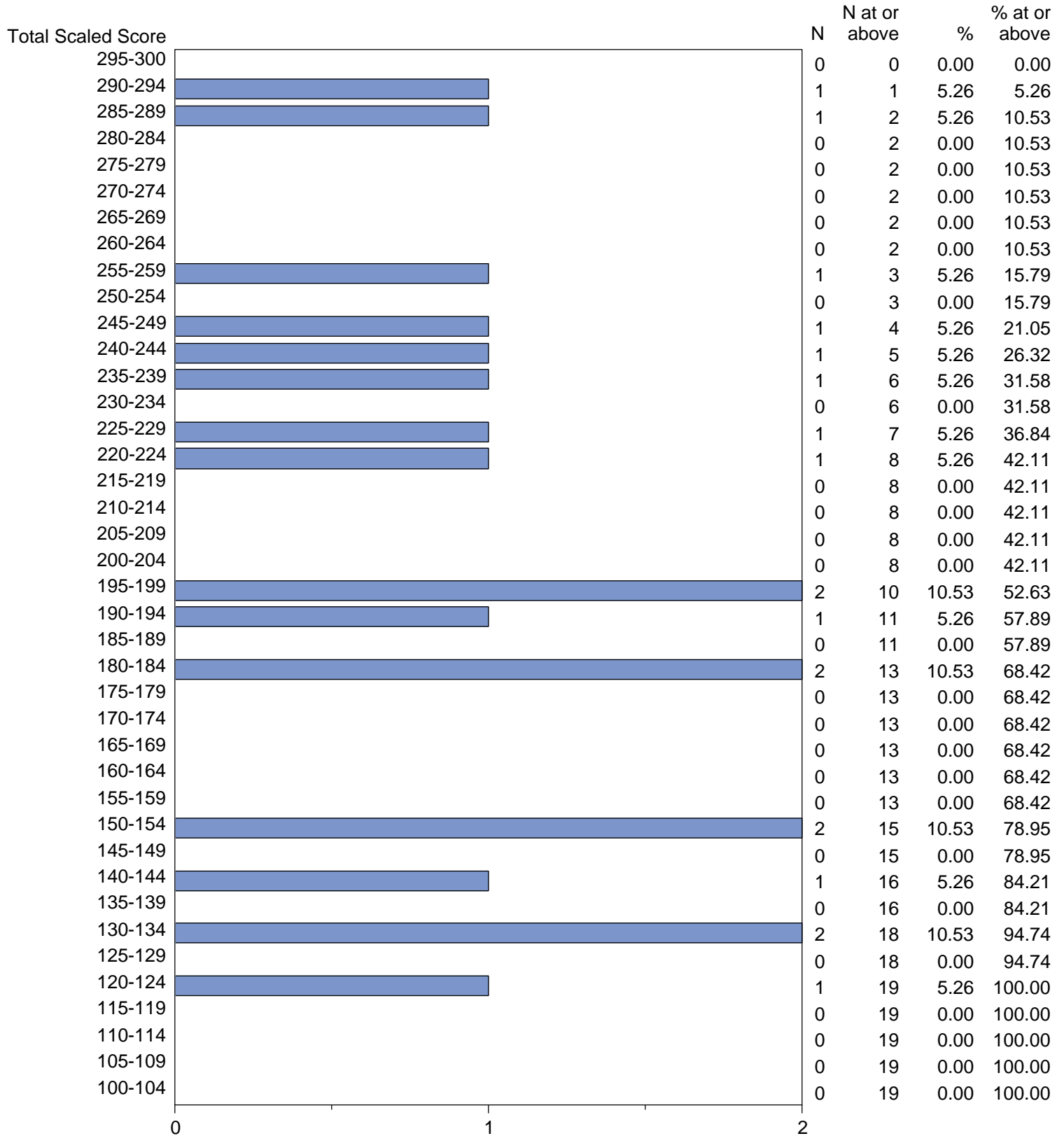
Test Field=009 Chemistry



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

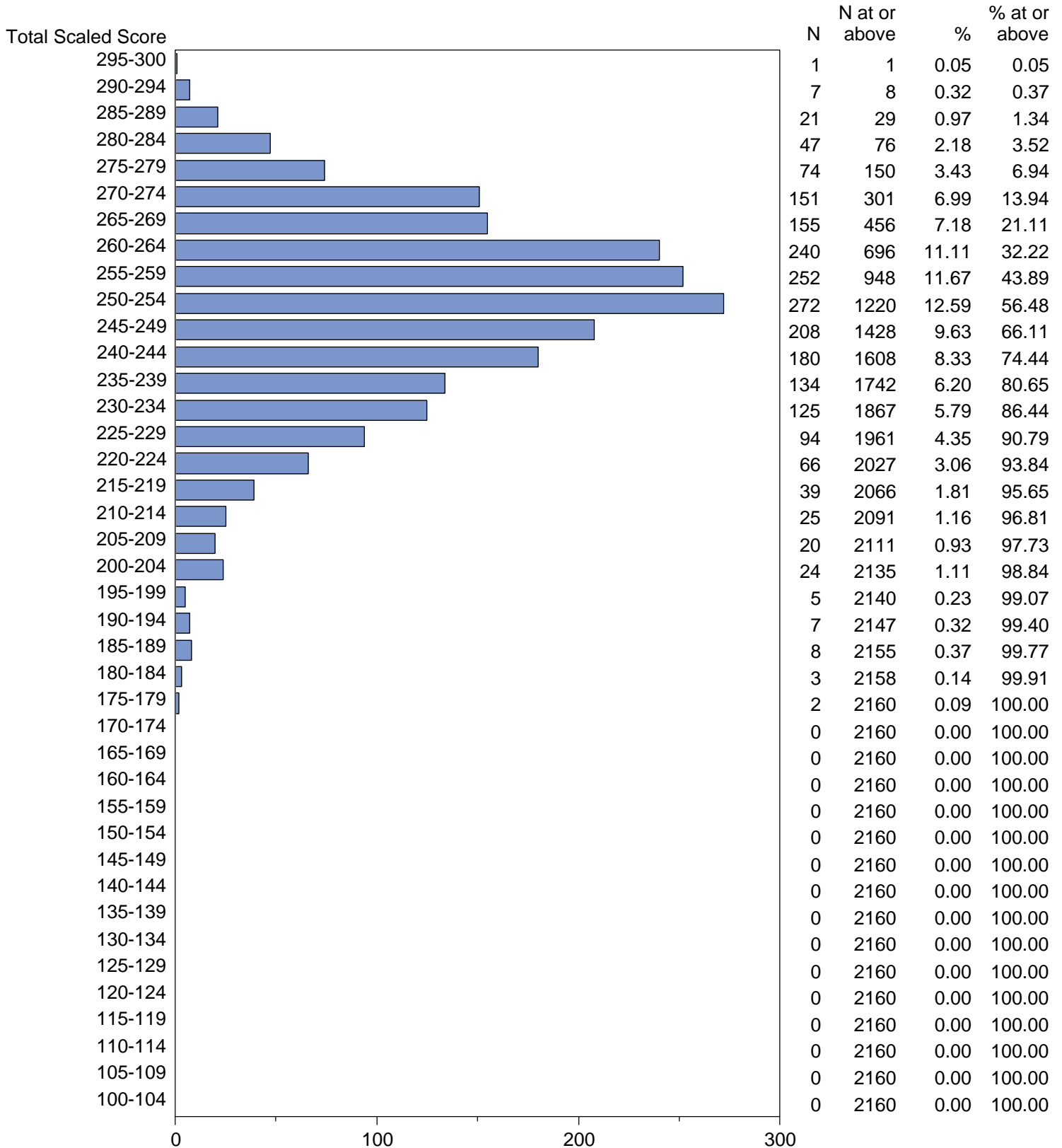
Test Field=010 Computer Information Science



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

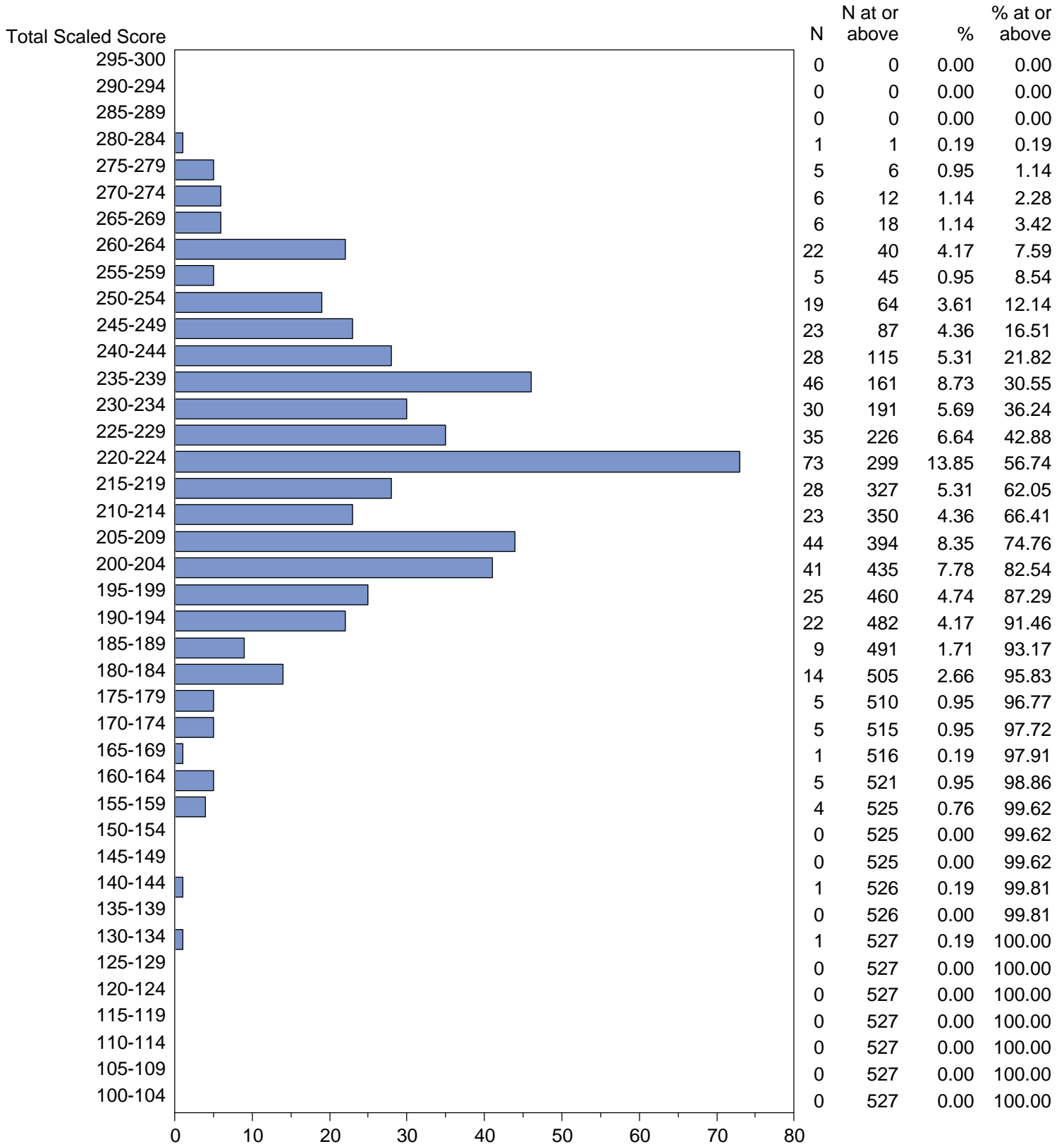
Test Field=012 Early Childhood Education



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

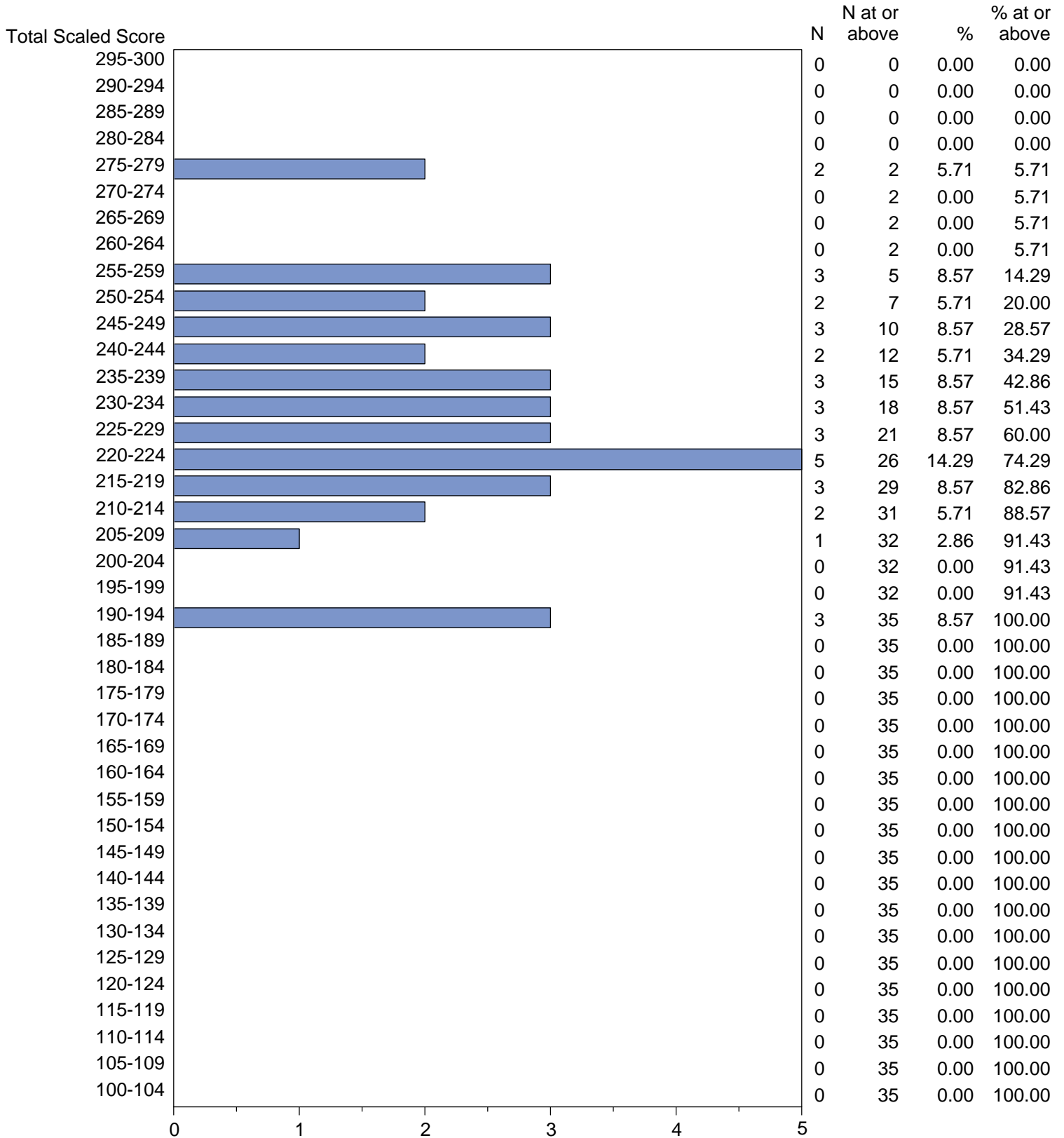
Test Field=013 Early Childhood Special Education



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

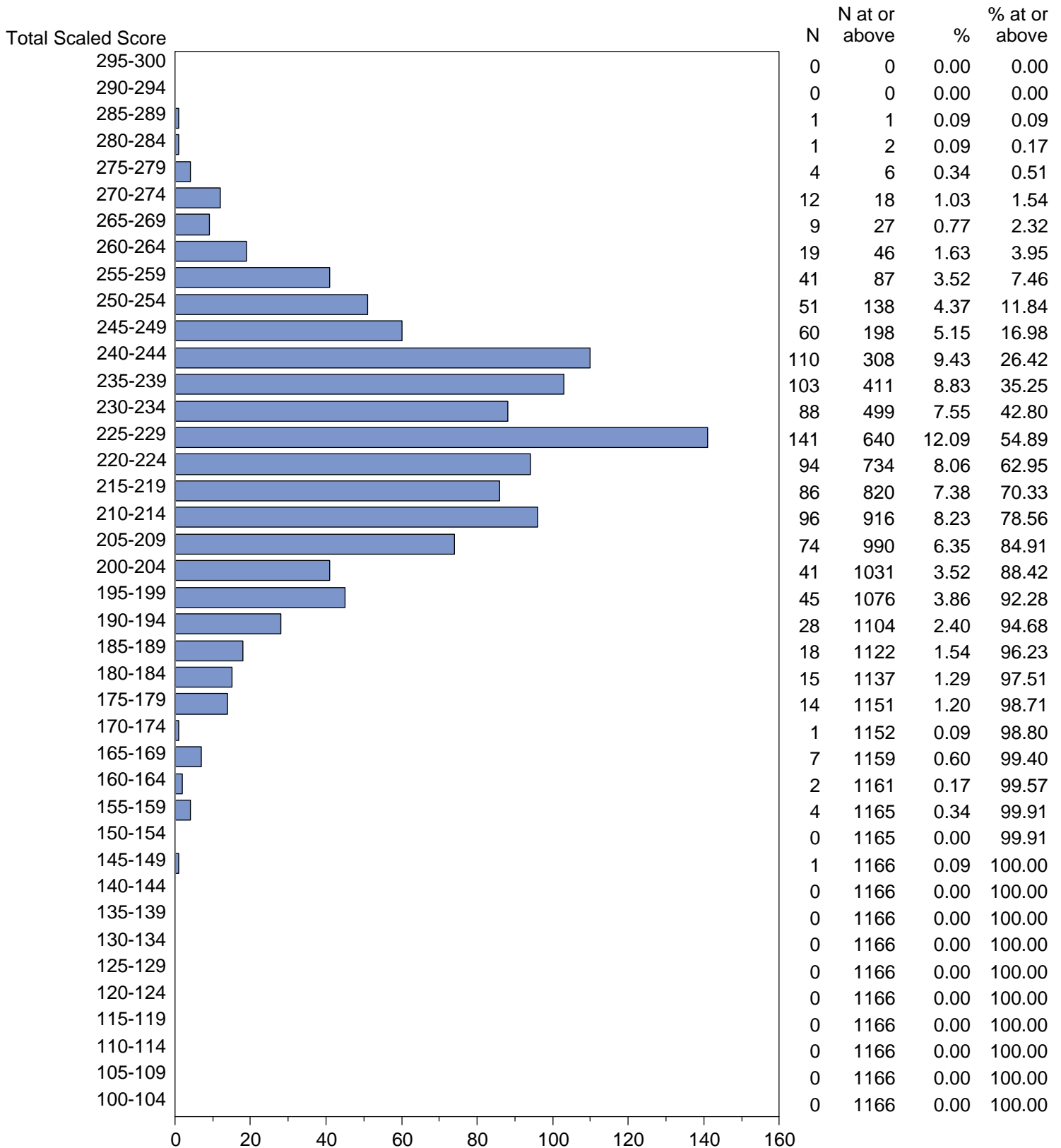
Test Field=014 Earth and Space Science



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

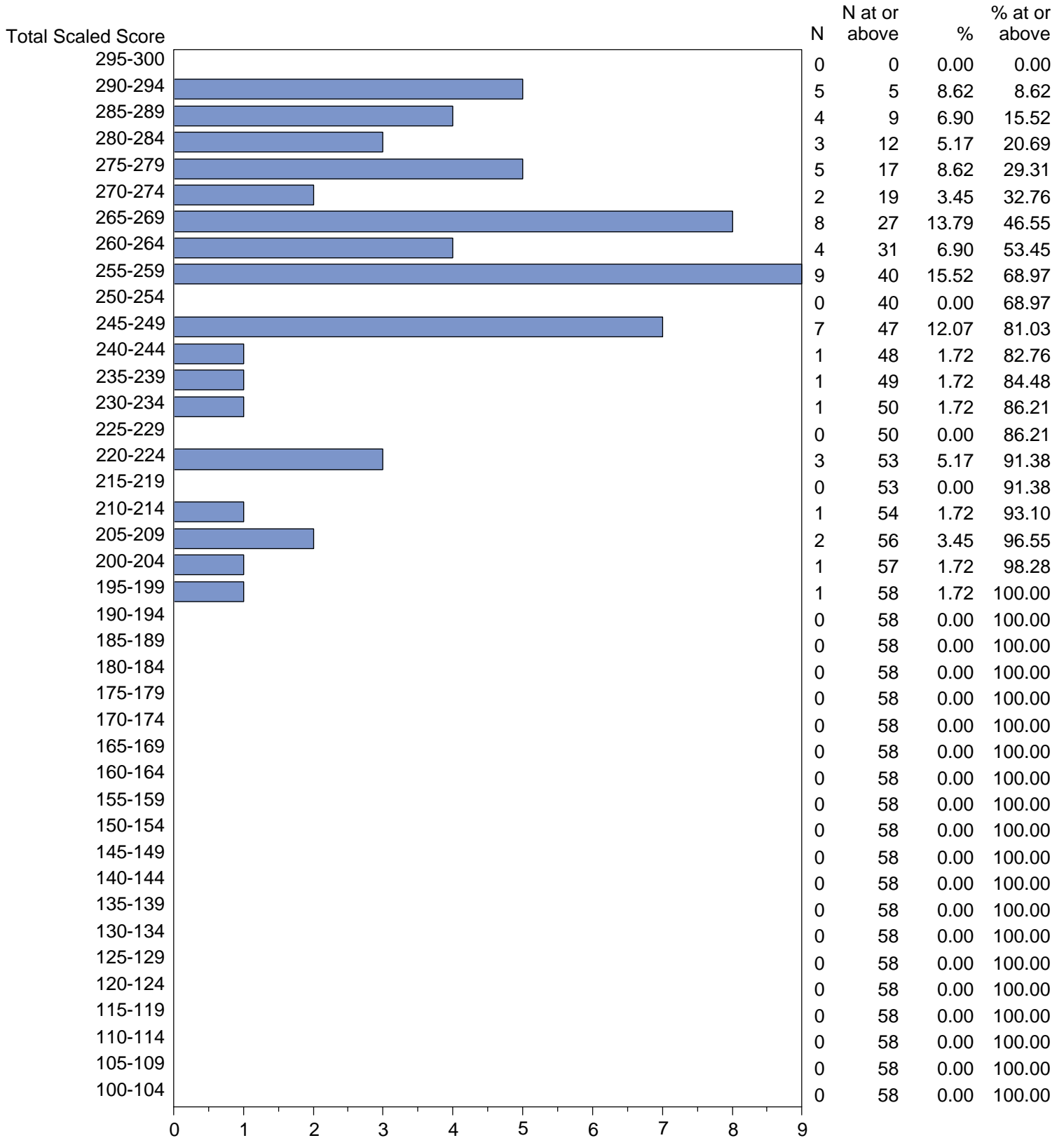
Test Field=015 Educational Leadership



Candidate Performance Statistical Report
 September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

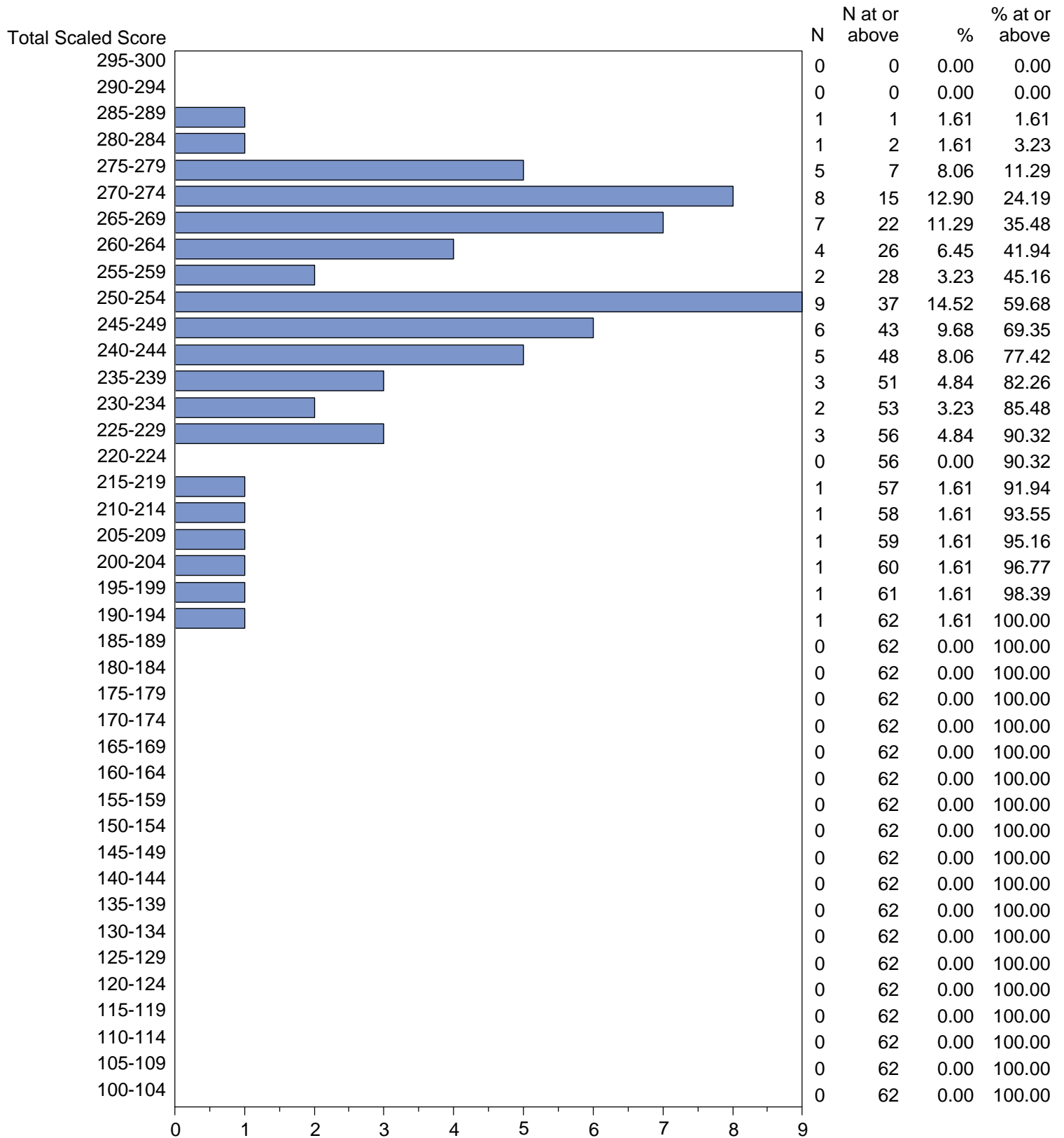
Test Field=016 Computer/Technology Subtest I



Candidate Performance Statistical Report
 September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

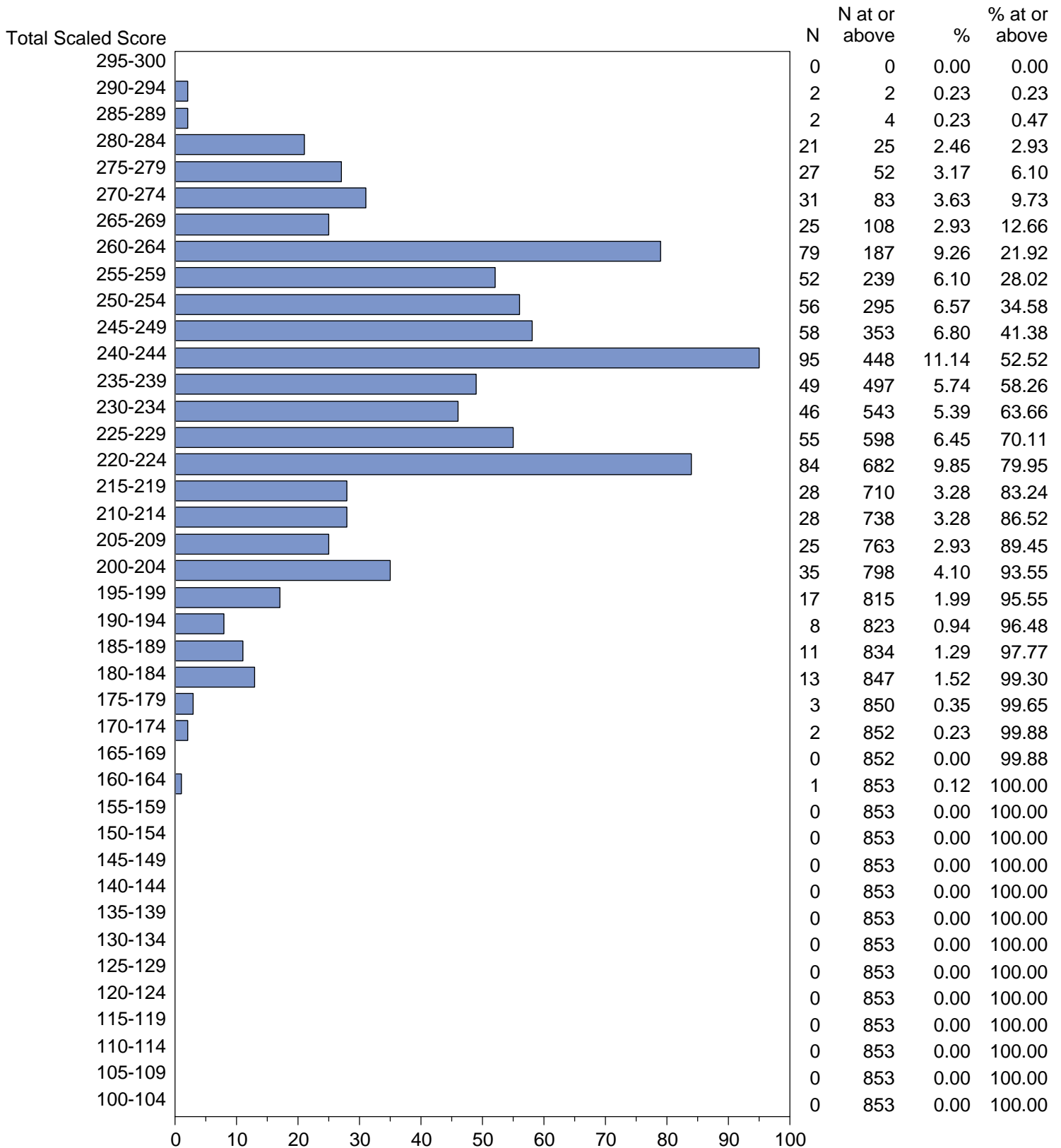
Test Field=017 Computer/Technology Subtest II



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

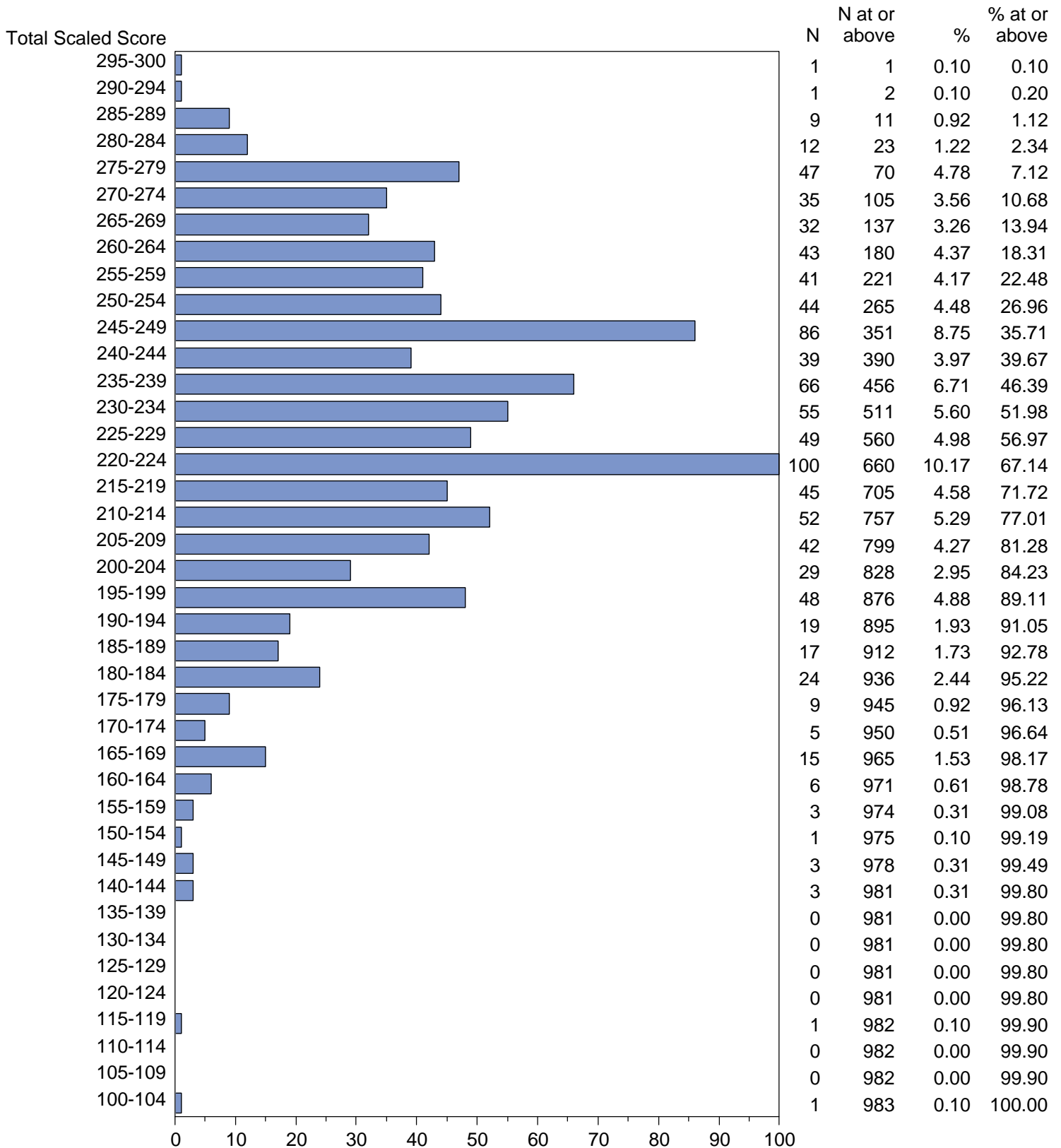
Test Field=018 Elementary Education Subtest I



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

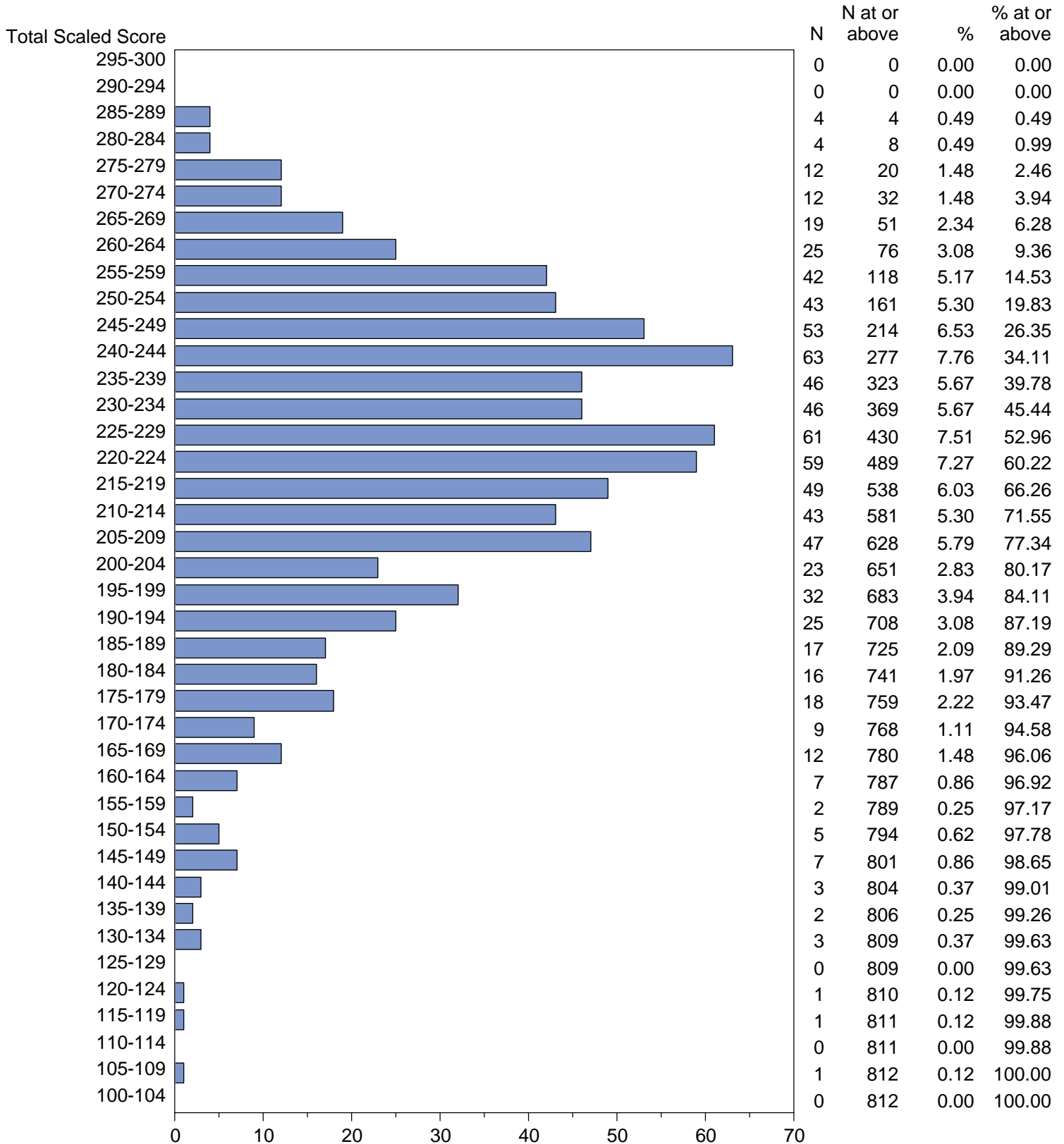
Test Field=019 Elementary Education Subtest II



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

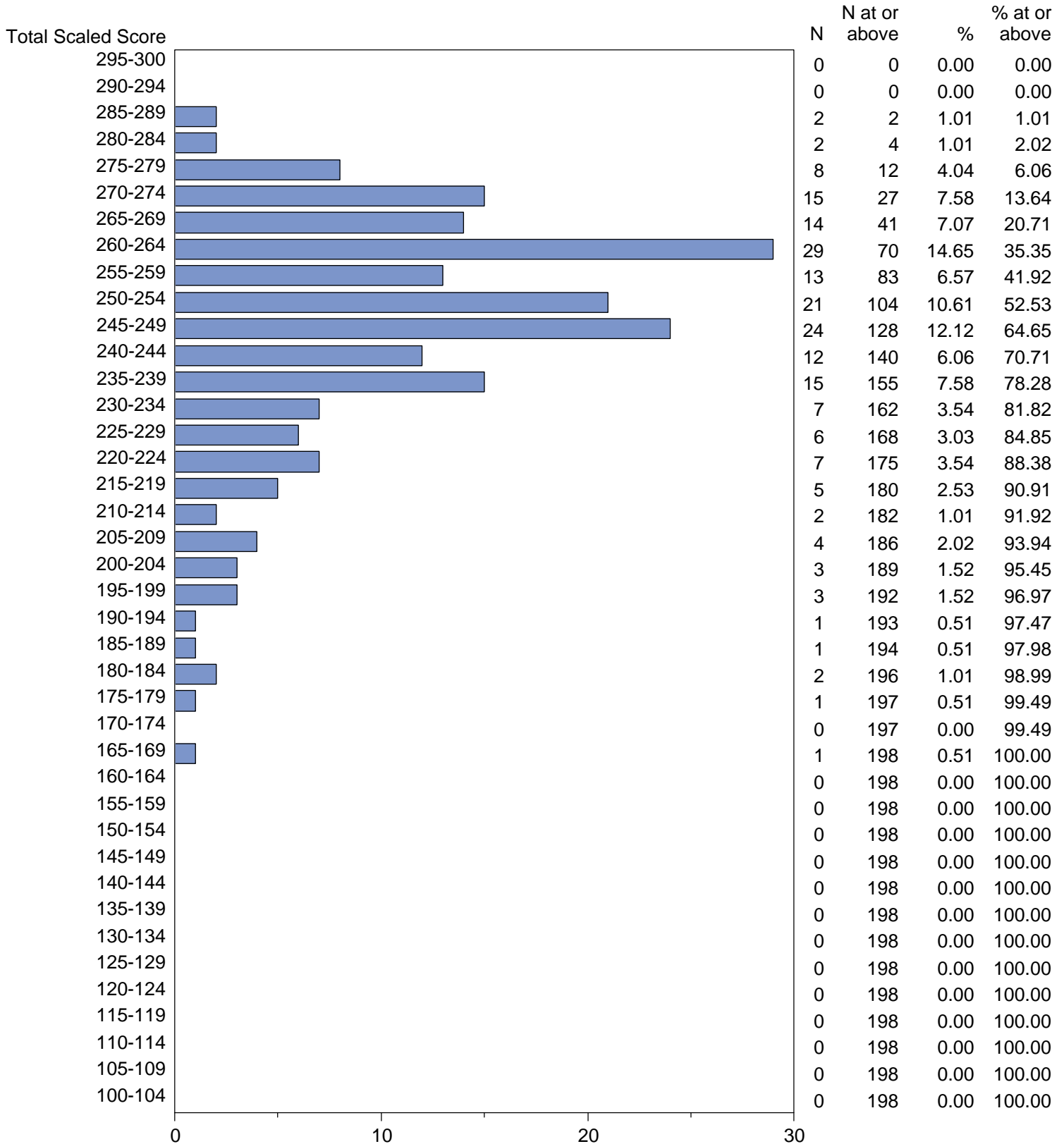
Test Field=020 English Language Arts



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

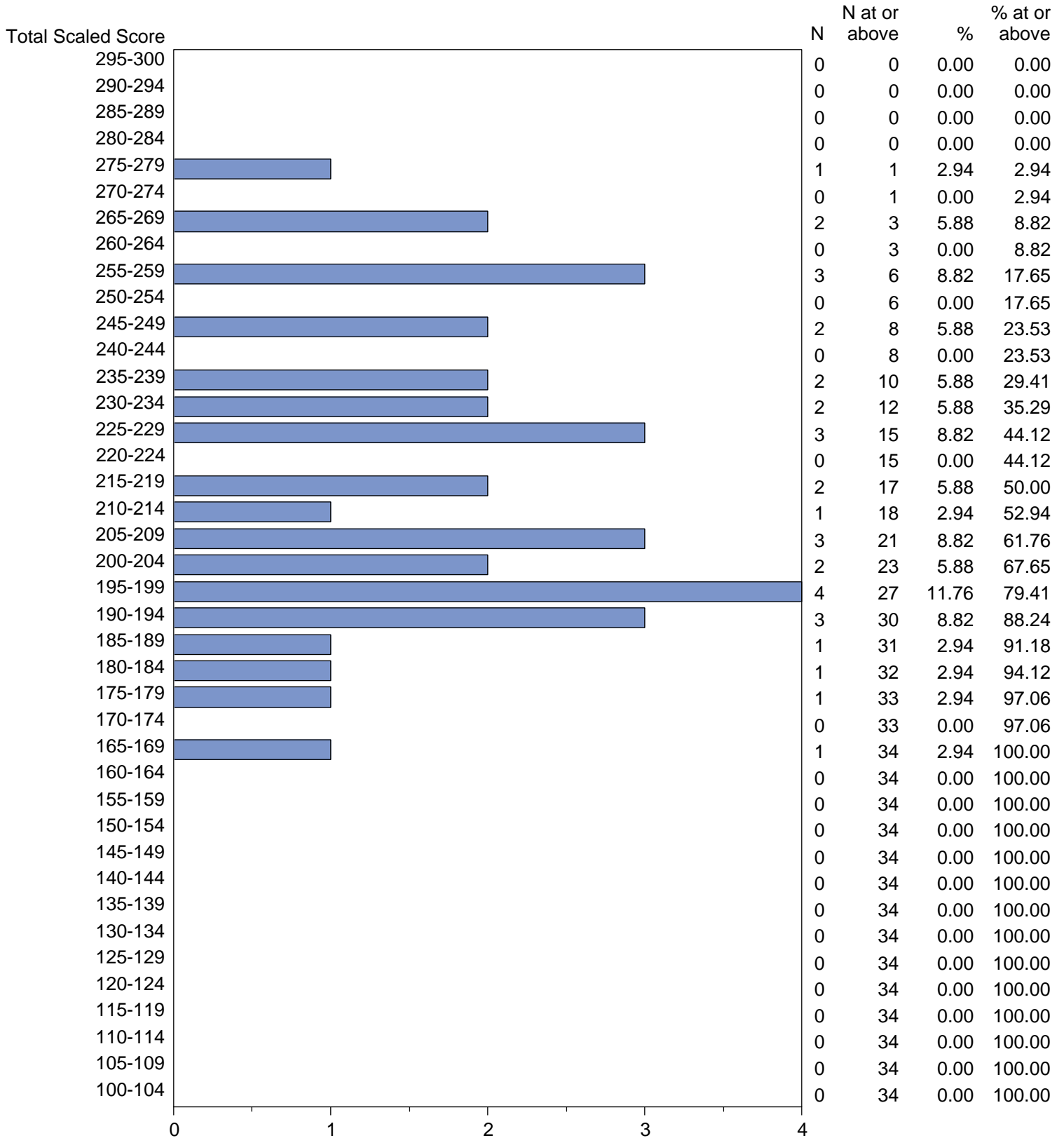
Test Field=021 English to Speakers of Other Languages (ESOL)



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

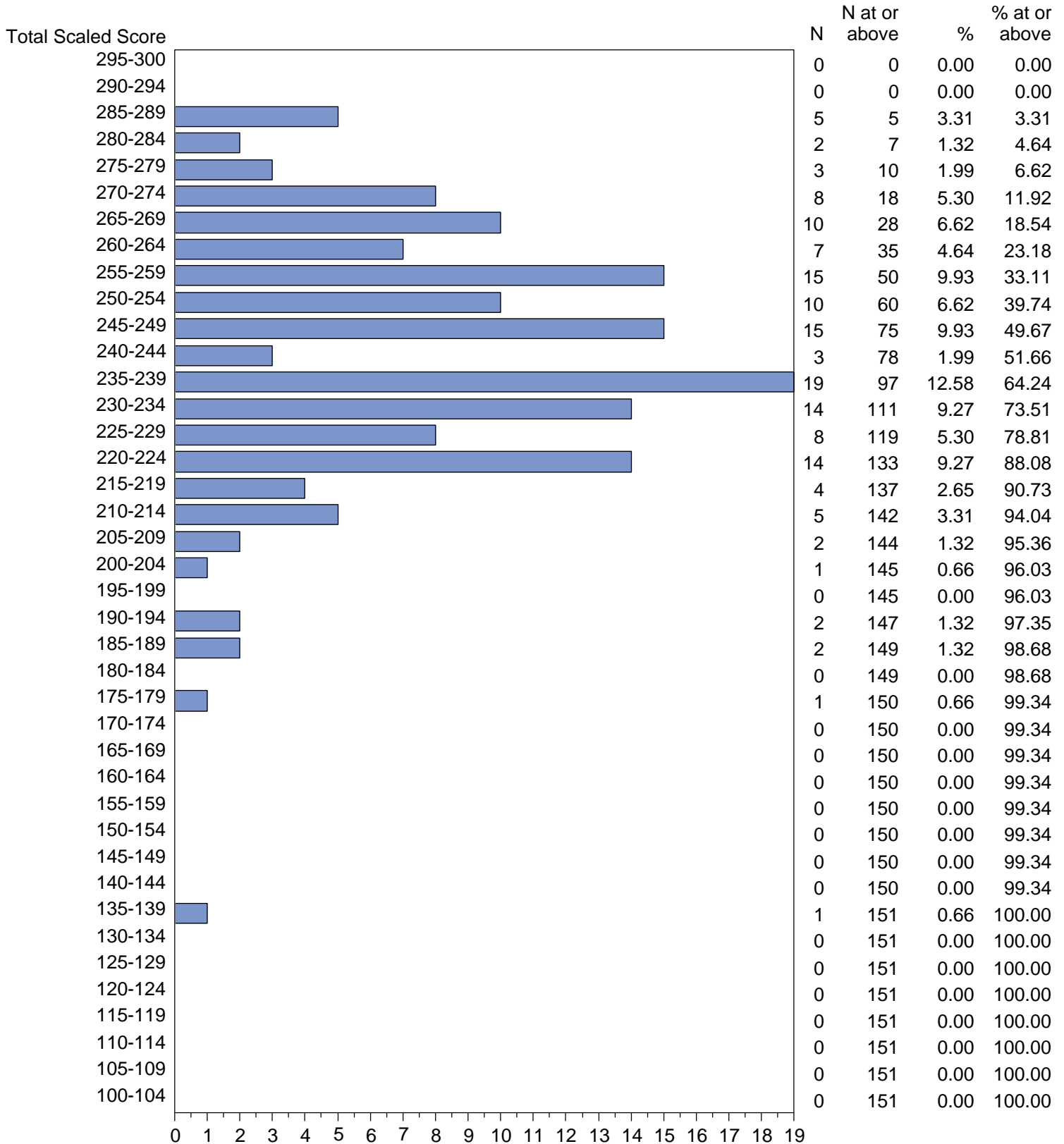
Test Field=022 Family and Consumer Sciences



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

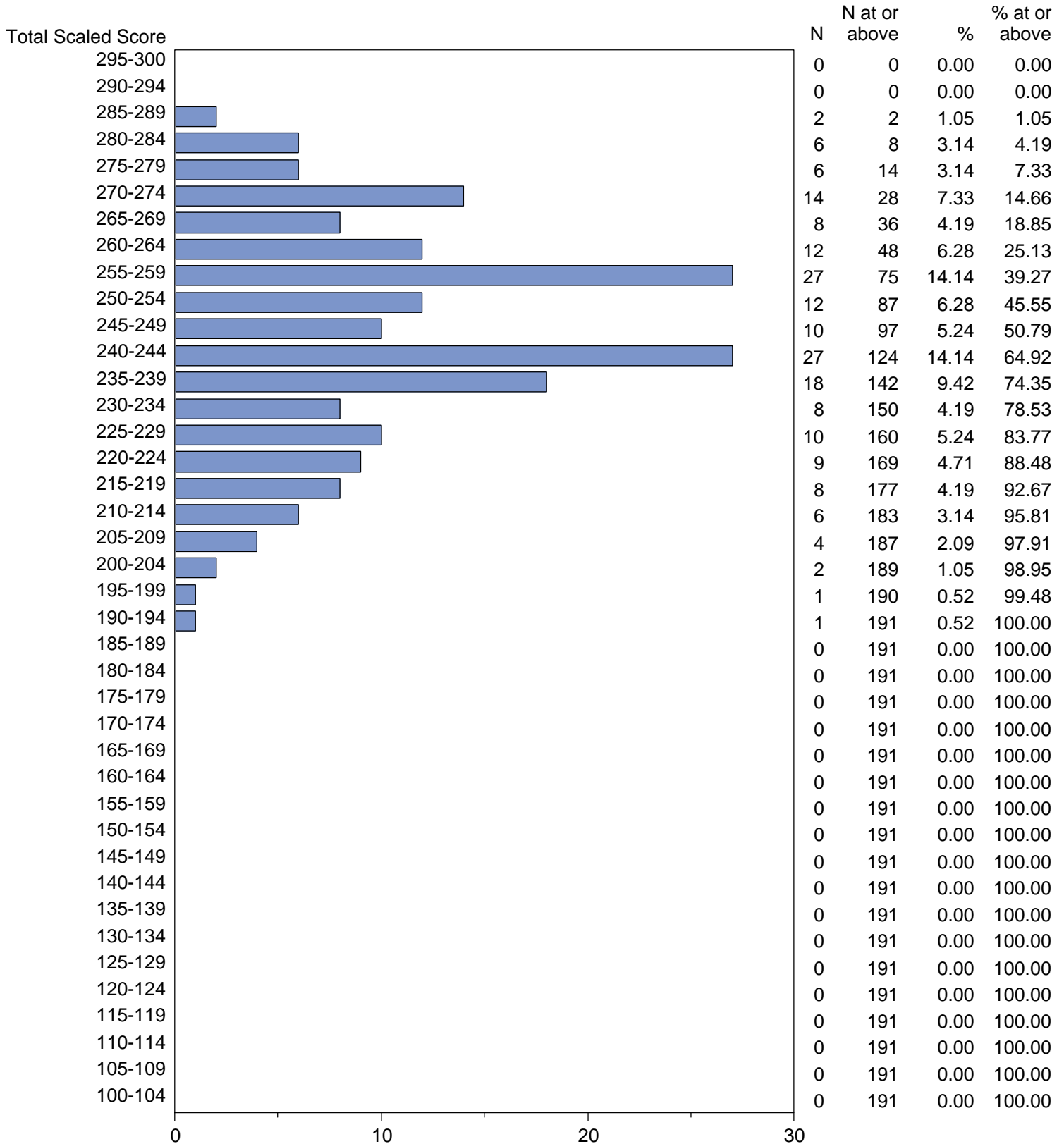
Test Field=023 Health



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

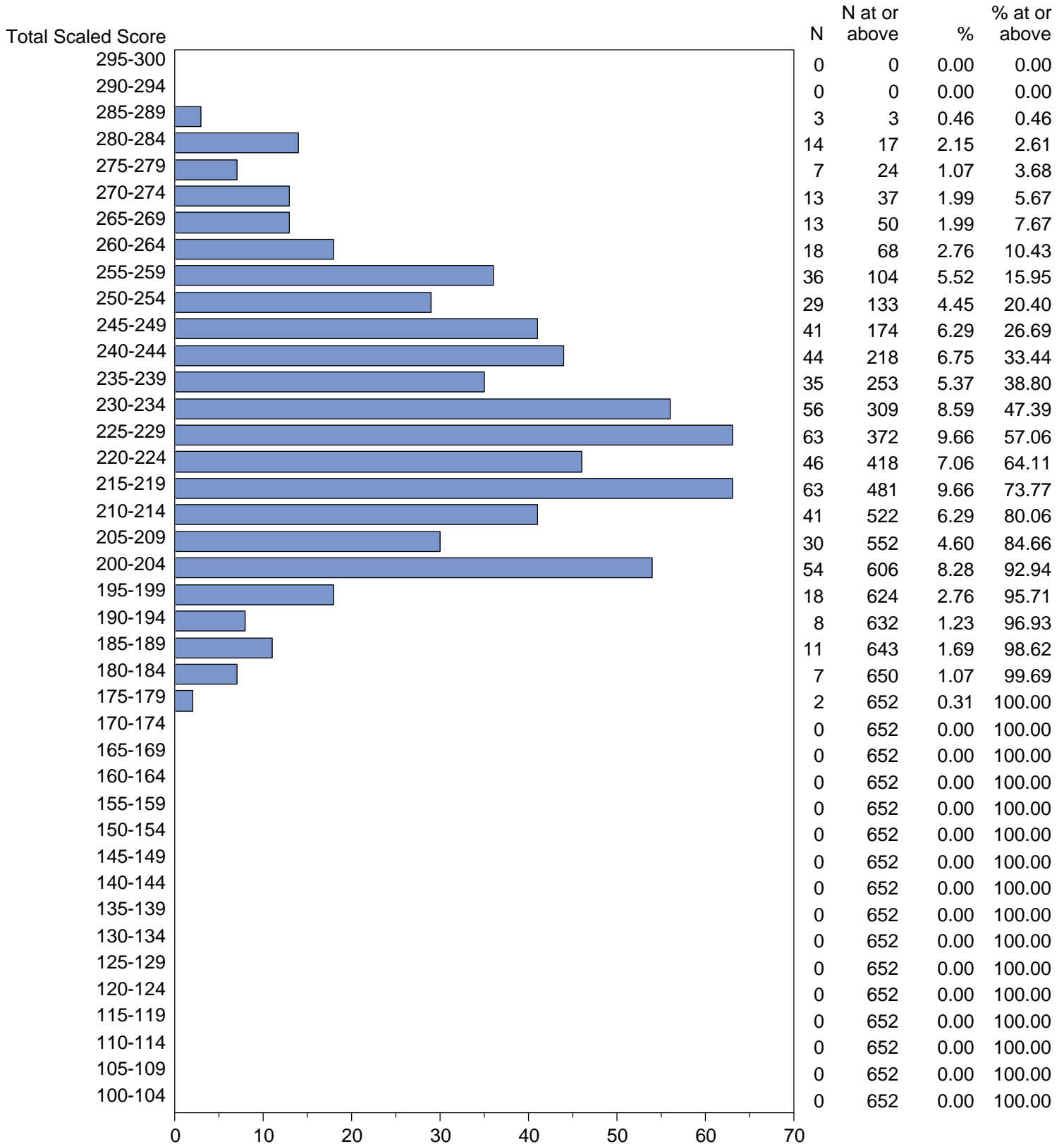
Test Field=024 Integrated Science



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

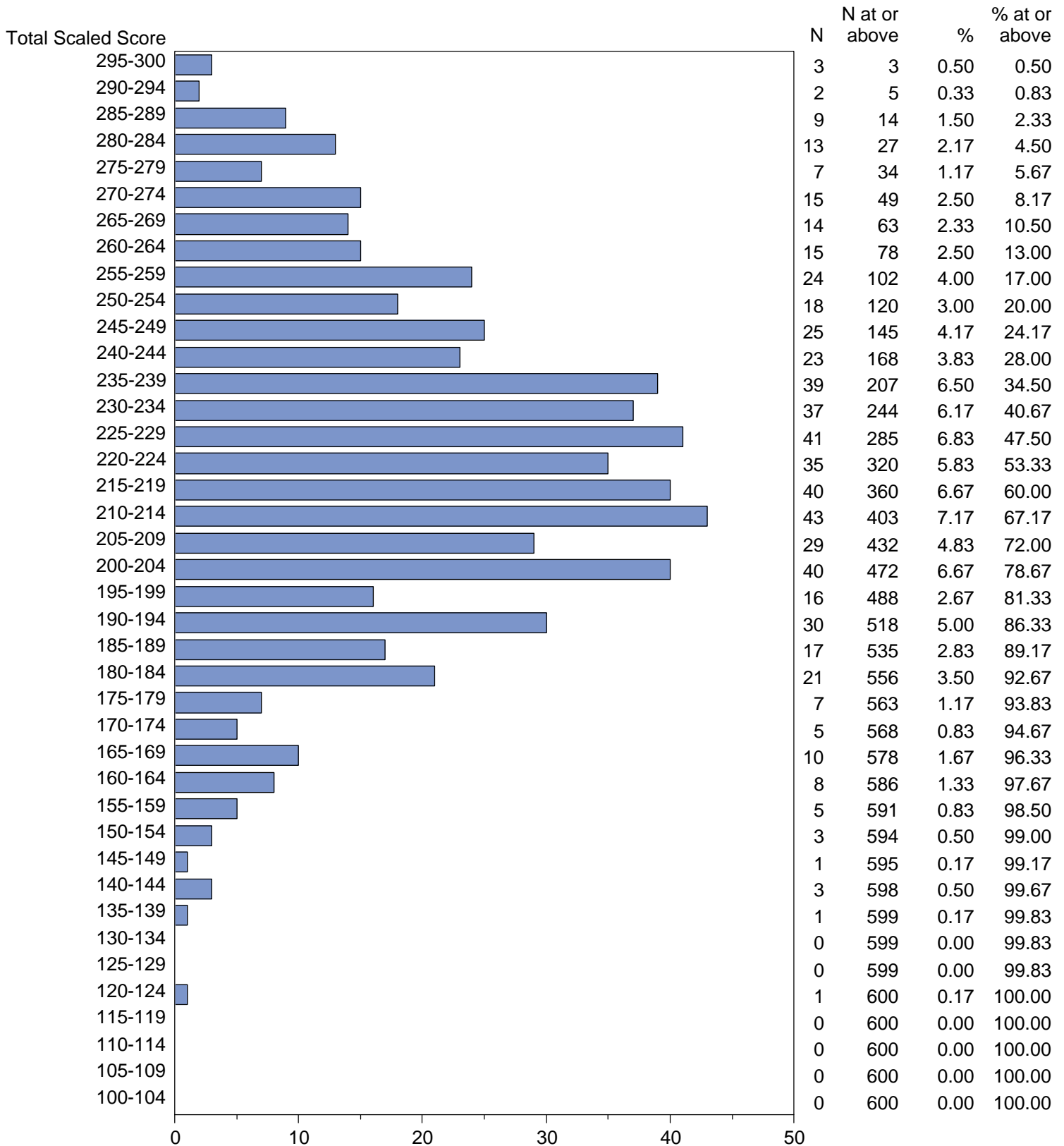
Test Field=025 Integrated Social Studies



Candidate Performance Statistical Report
 September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

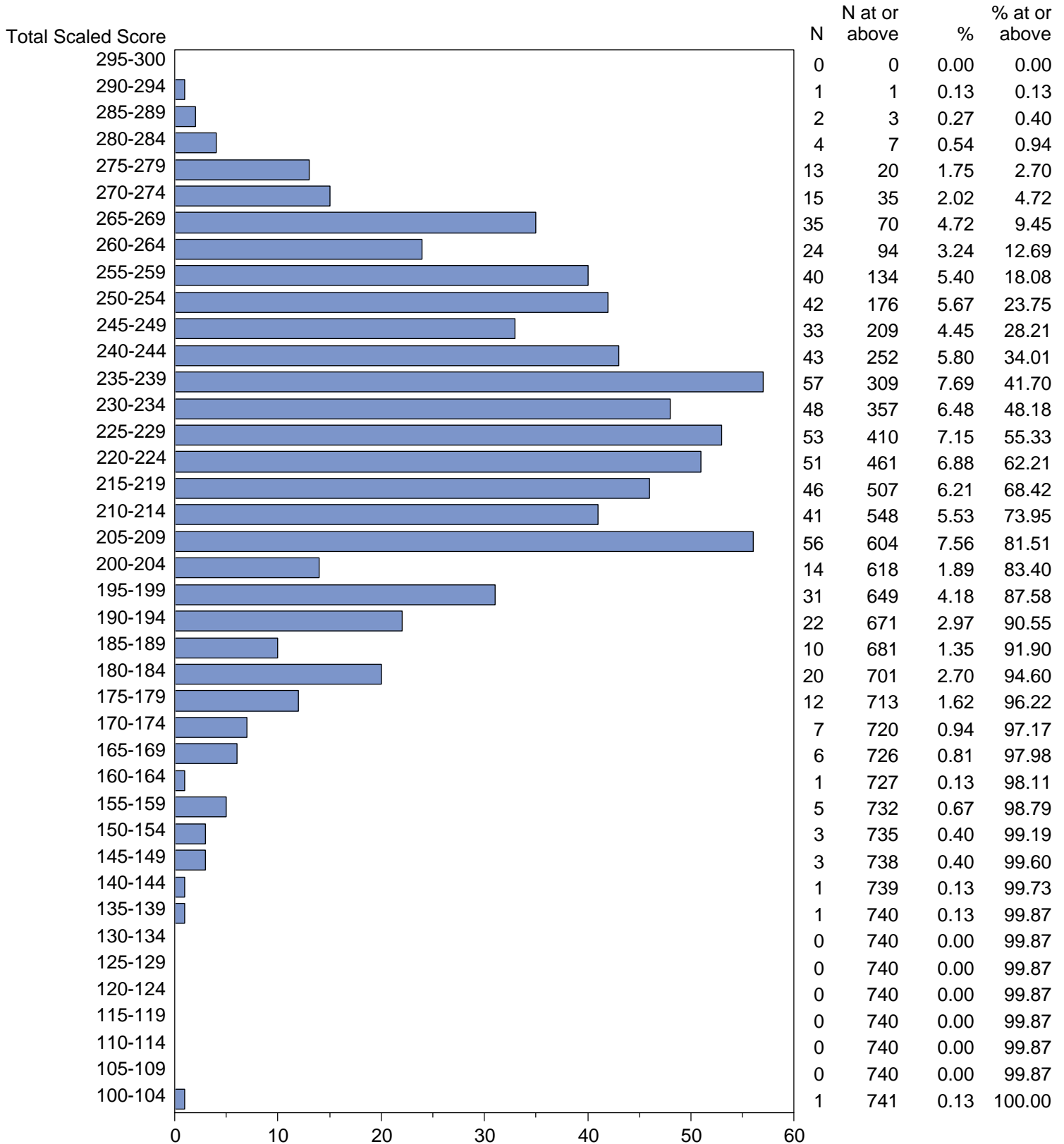
Test Field=027 Mathematics



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

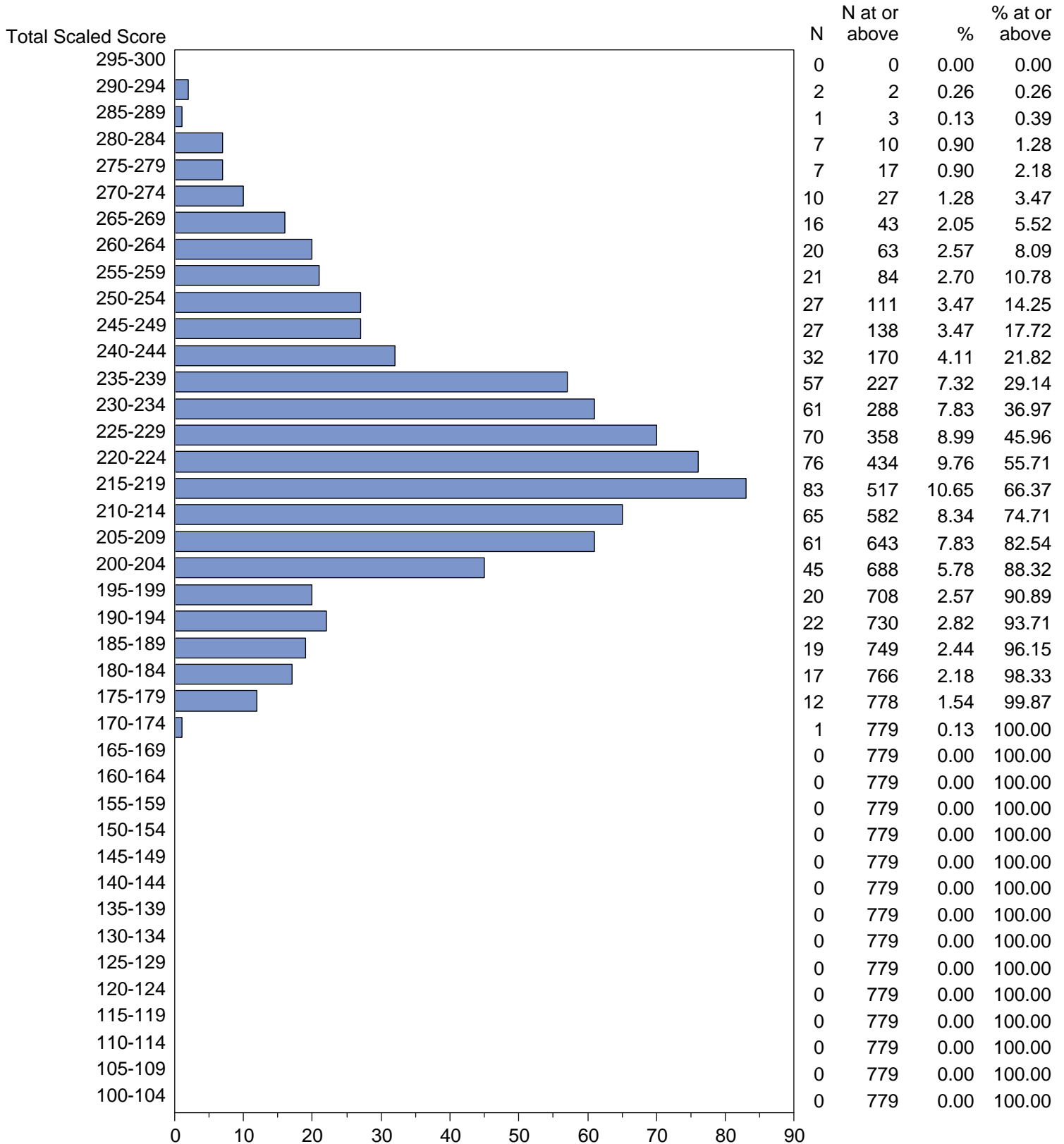
Test Field=028 Middle Grades English Language Arts



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

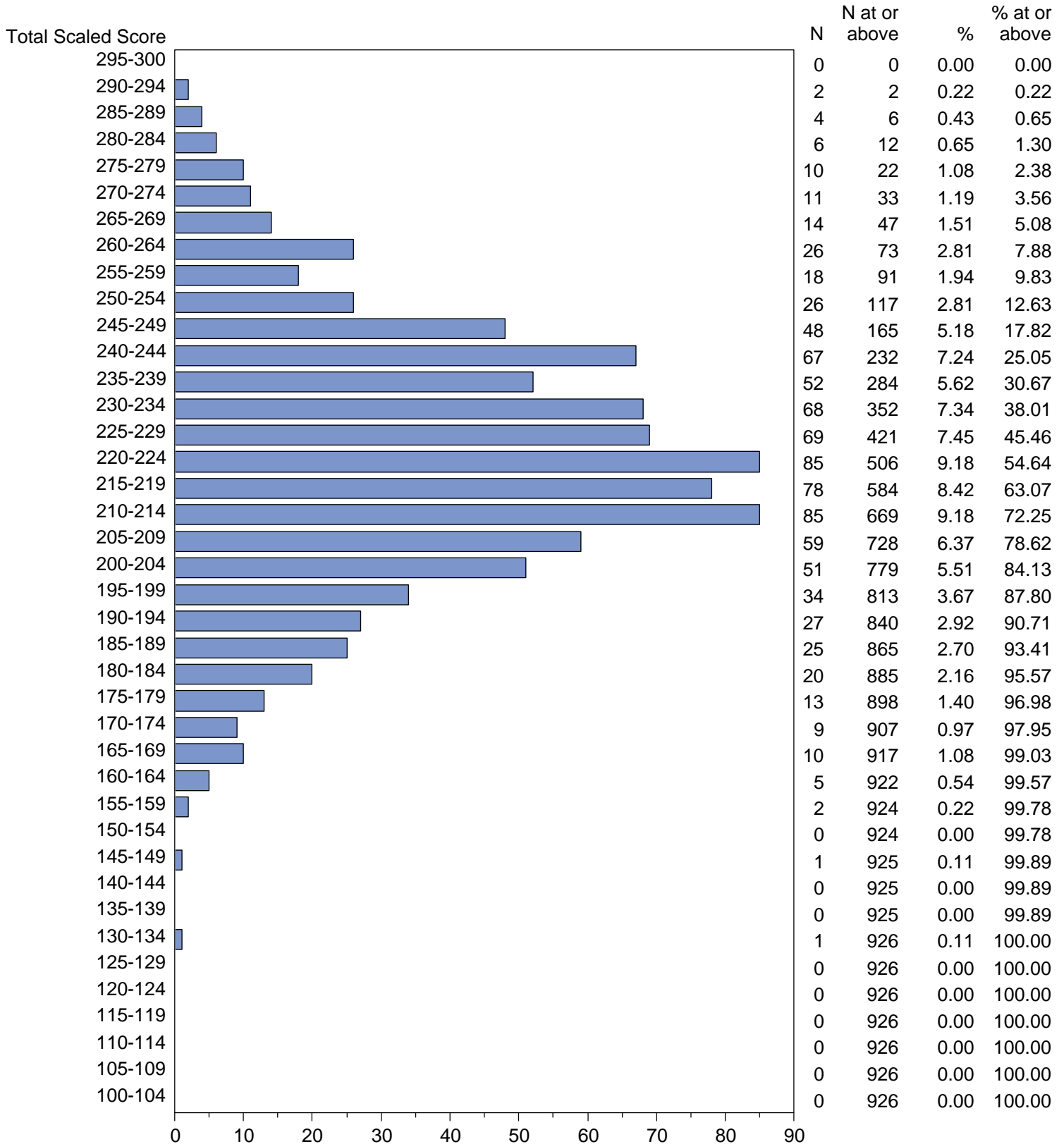
Test Field=029 Middle Grades Science



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

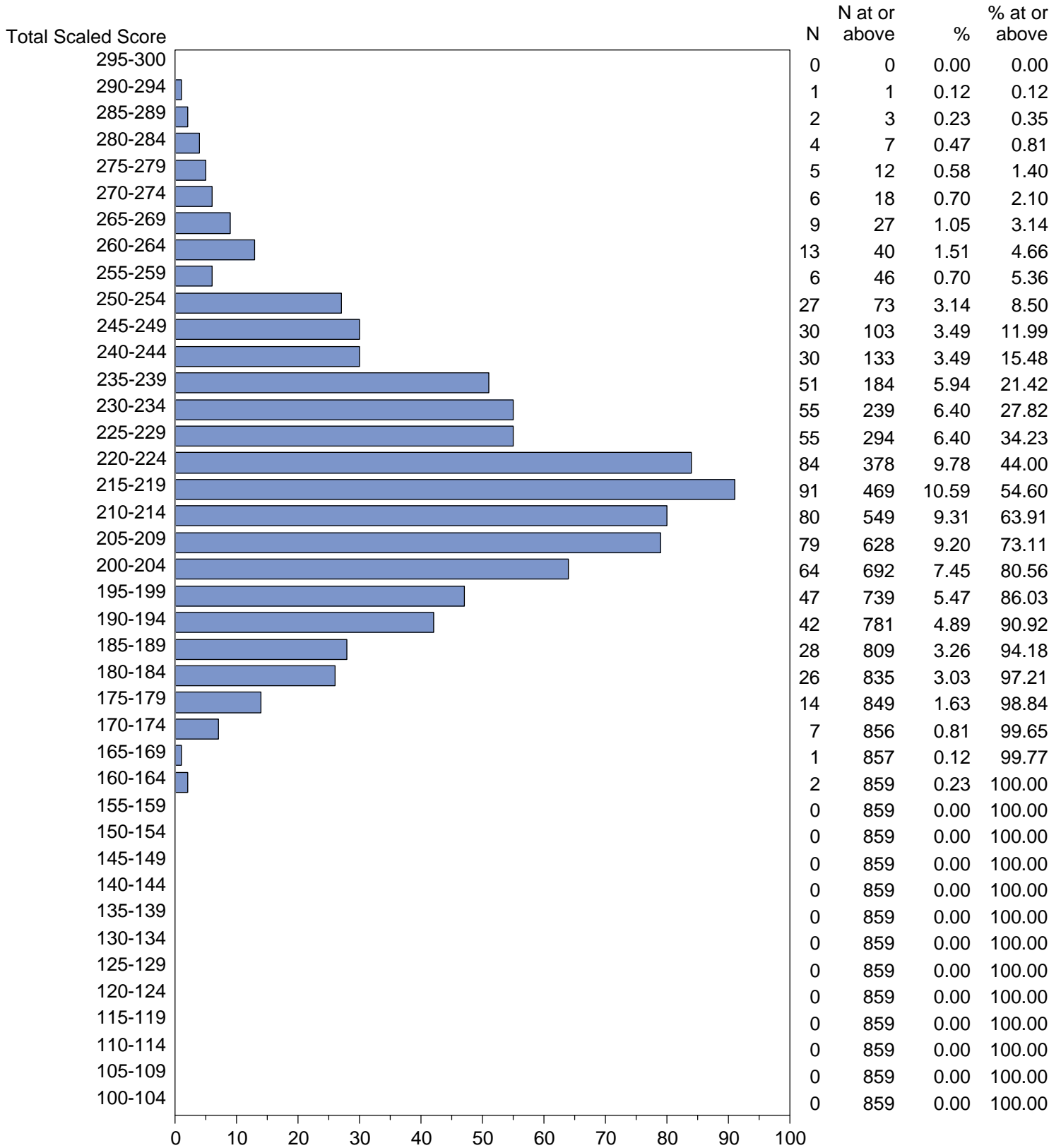
Test Field=030 Middle Grades Mathematics



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

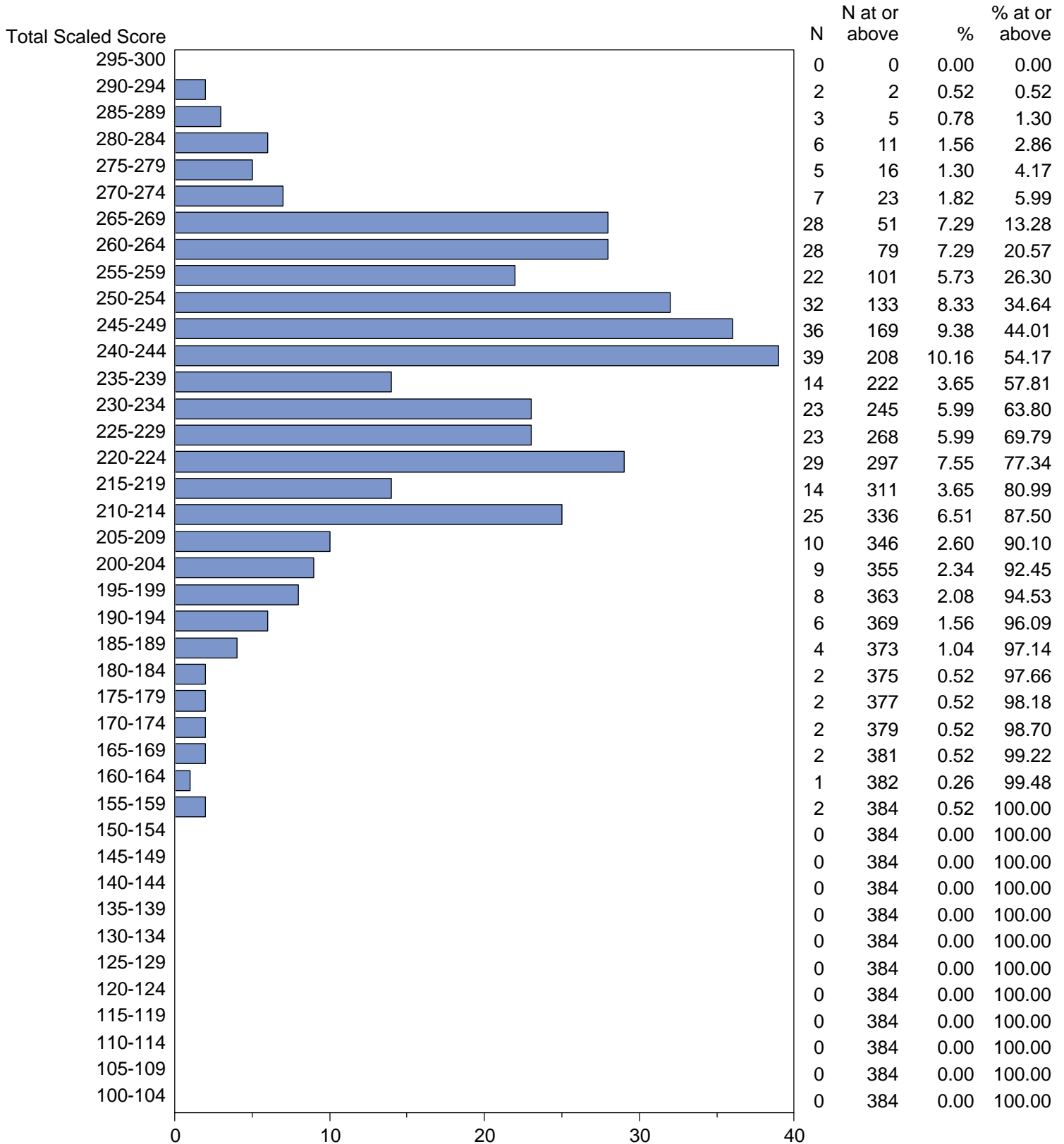
Test Field=031 Middle Grades Social Studies



Candidate Performance Statistical Report
 September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

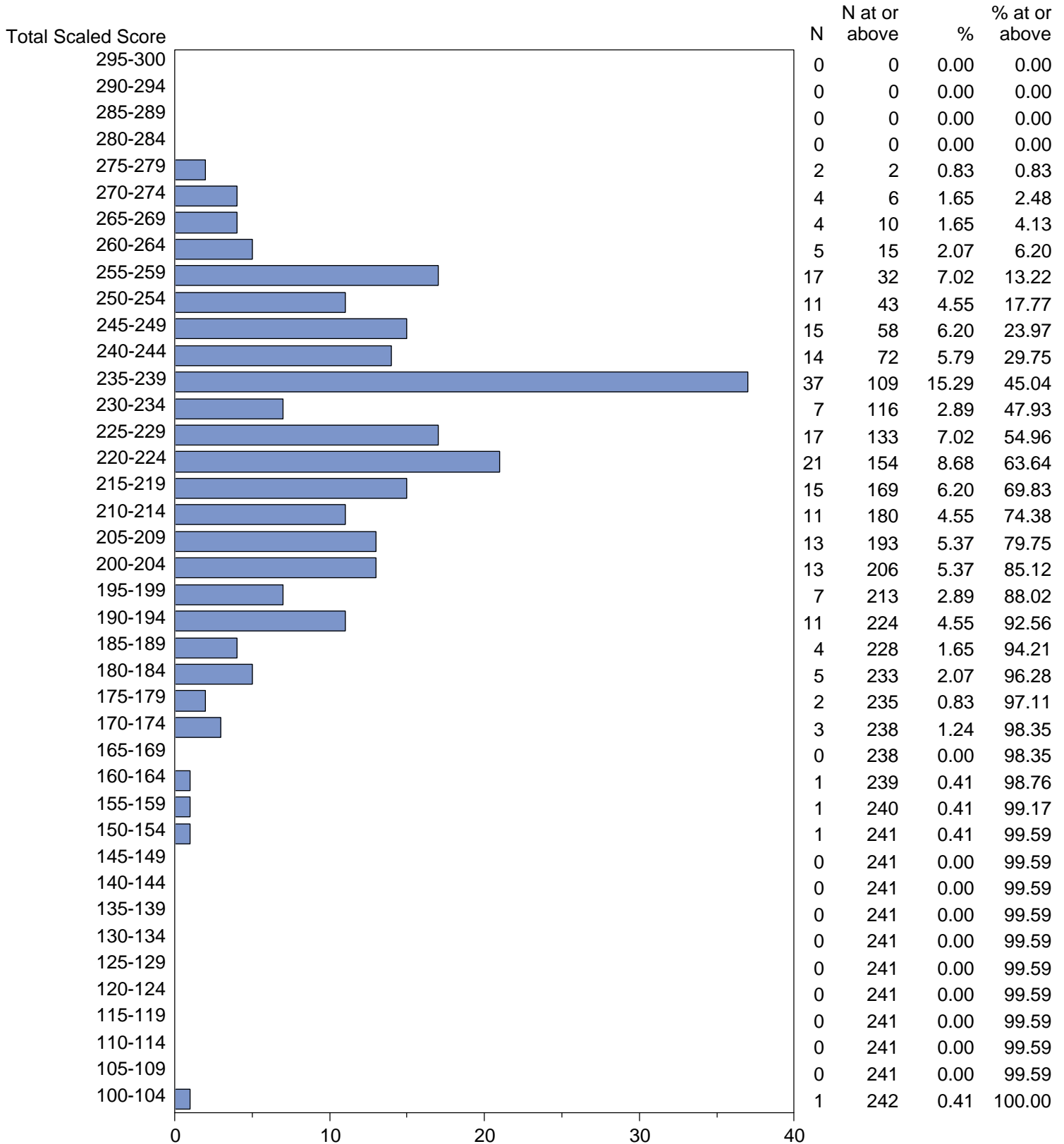
Test Field=032 Music



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

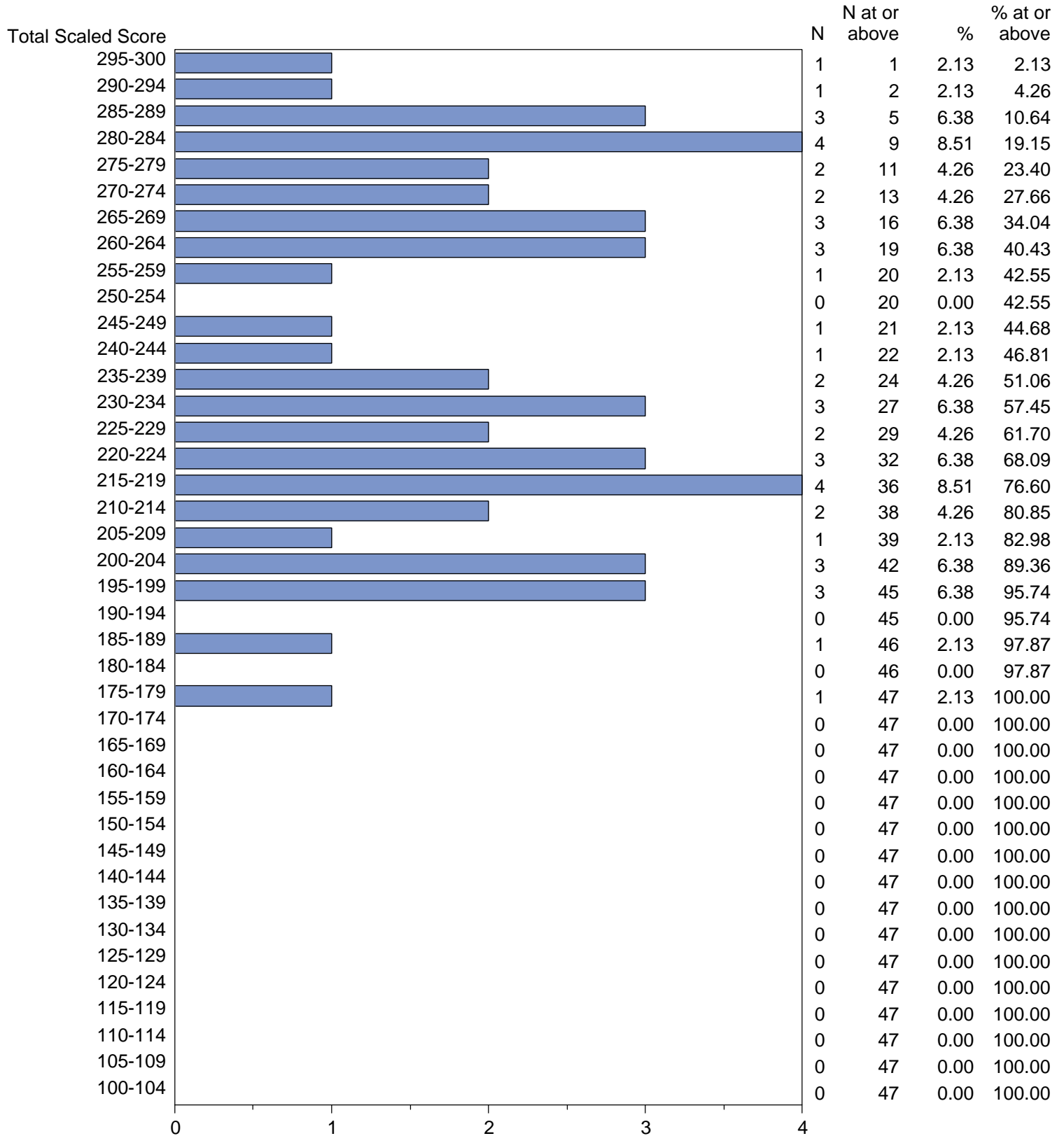
Test Field=034 Physical Education



Candidate Performance Statistical Report
 September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

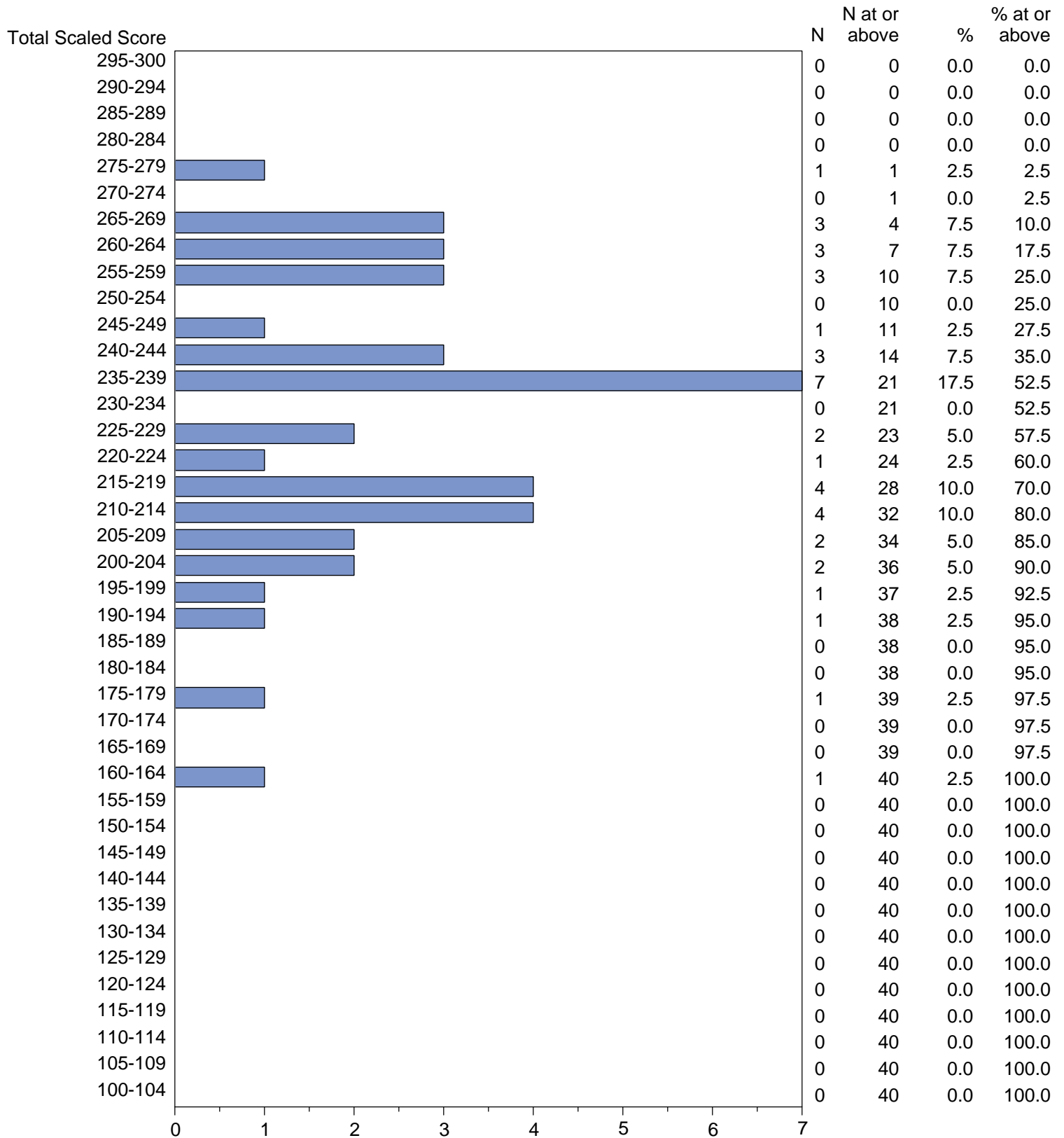
Test Field=035 Physics



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

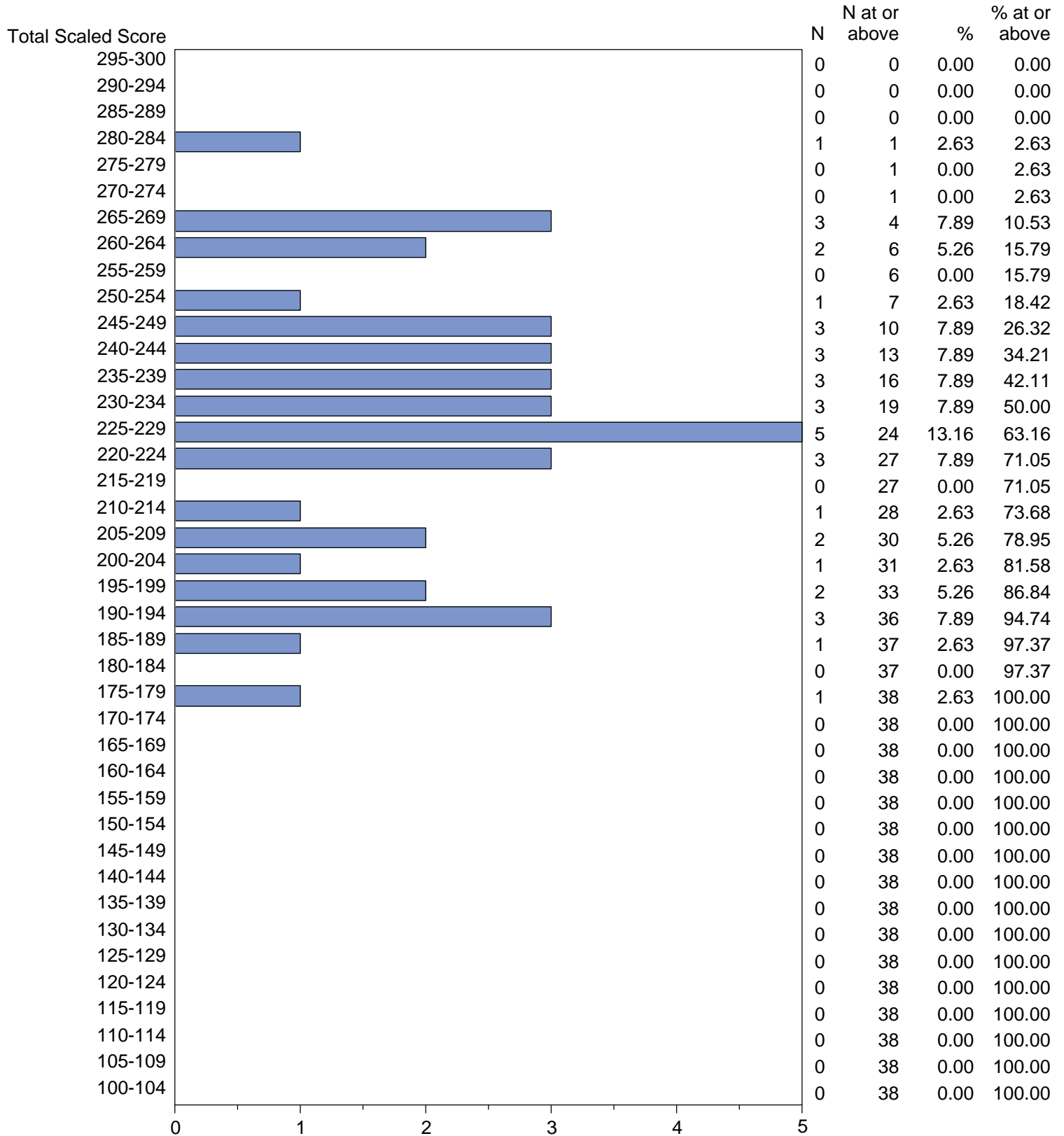
Test Field=036 Prekindergarten Subtest I



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

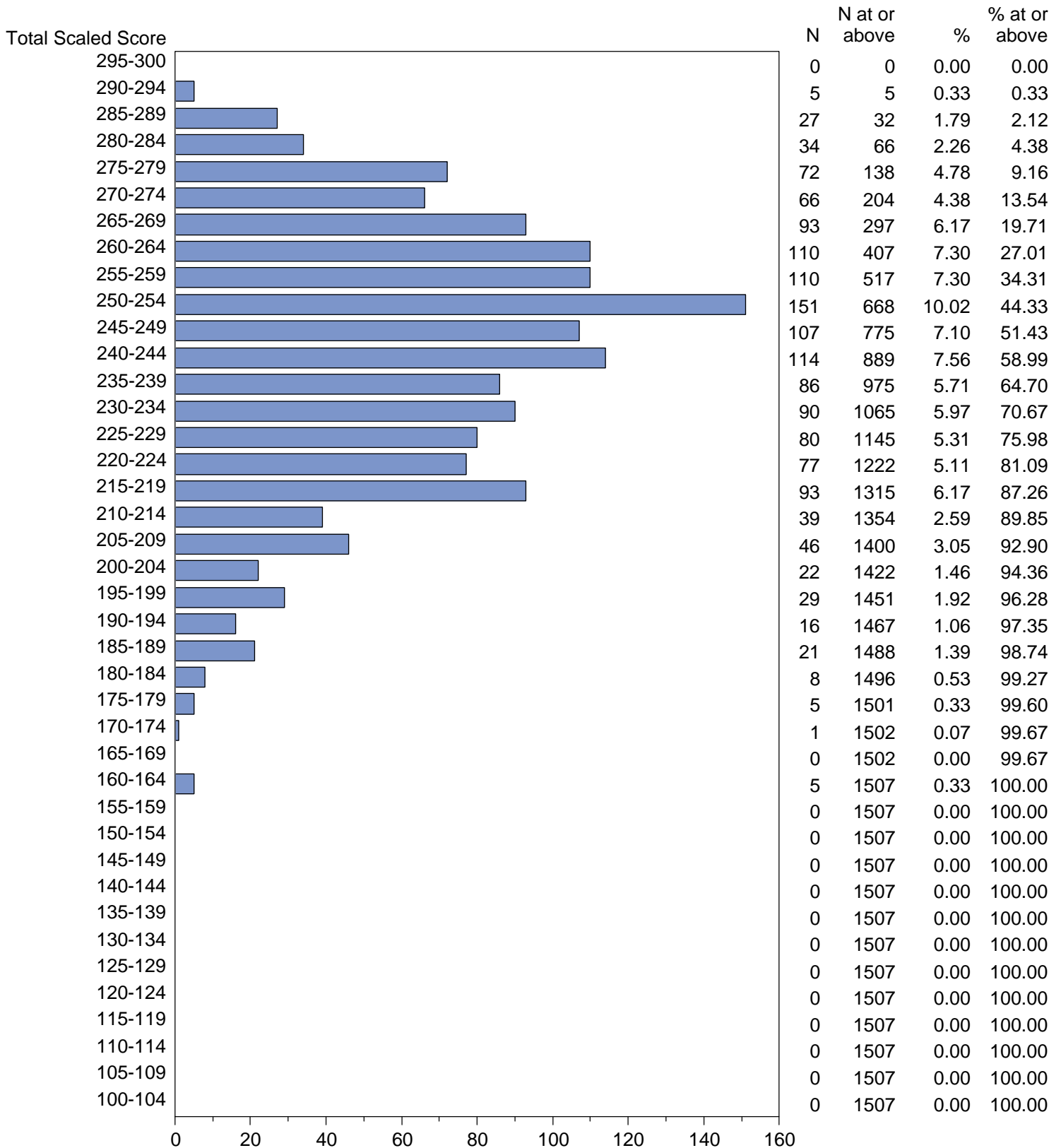
Test Field=037 Prekindergarten Subtest II



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

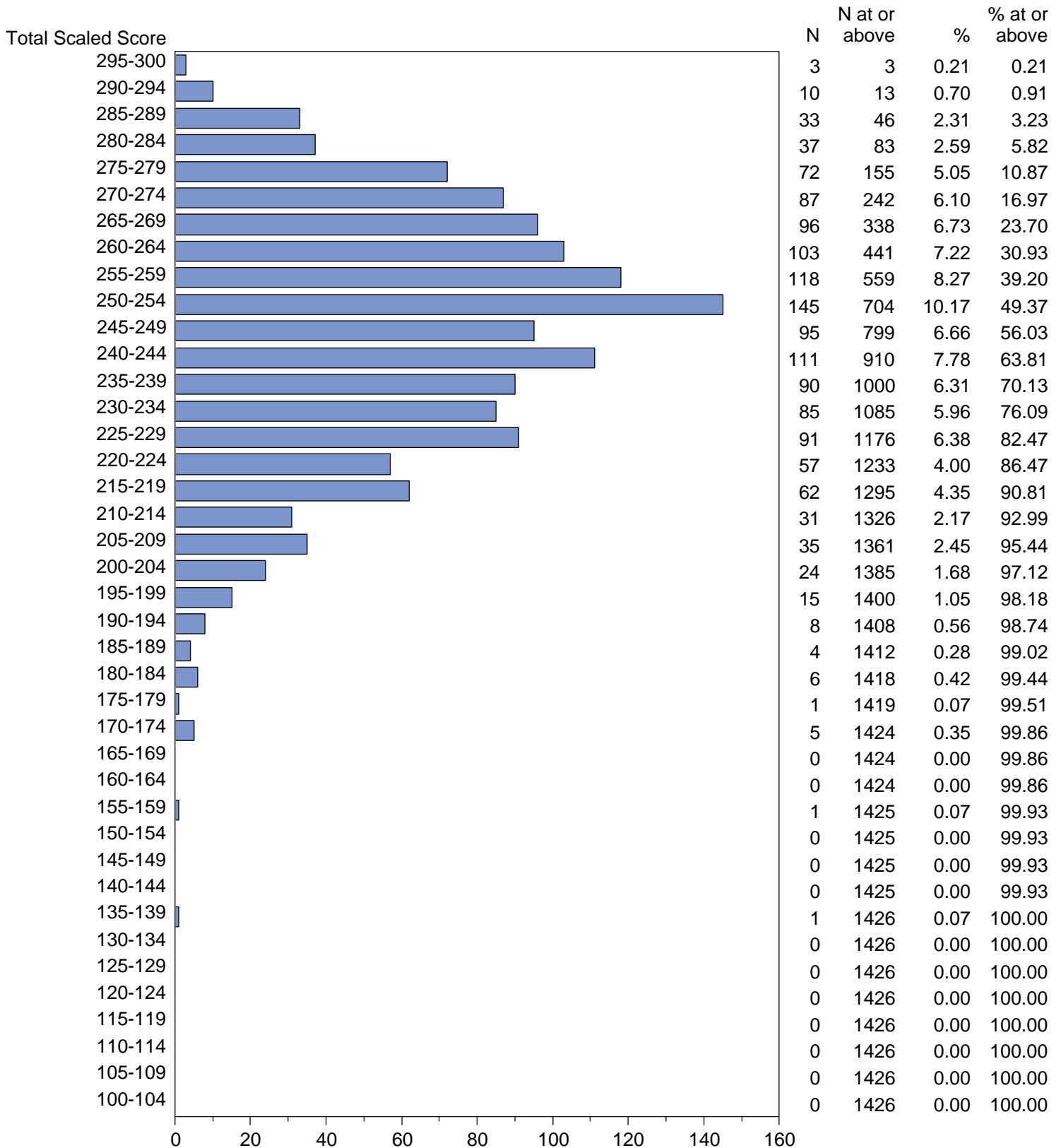
Test Field=038 Reading Subtest I



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

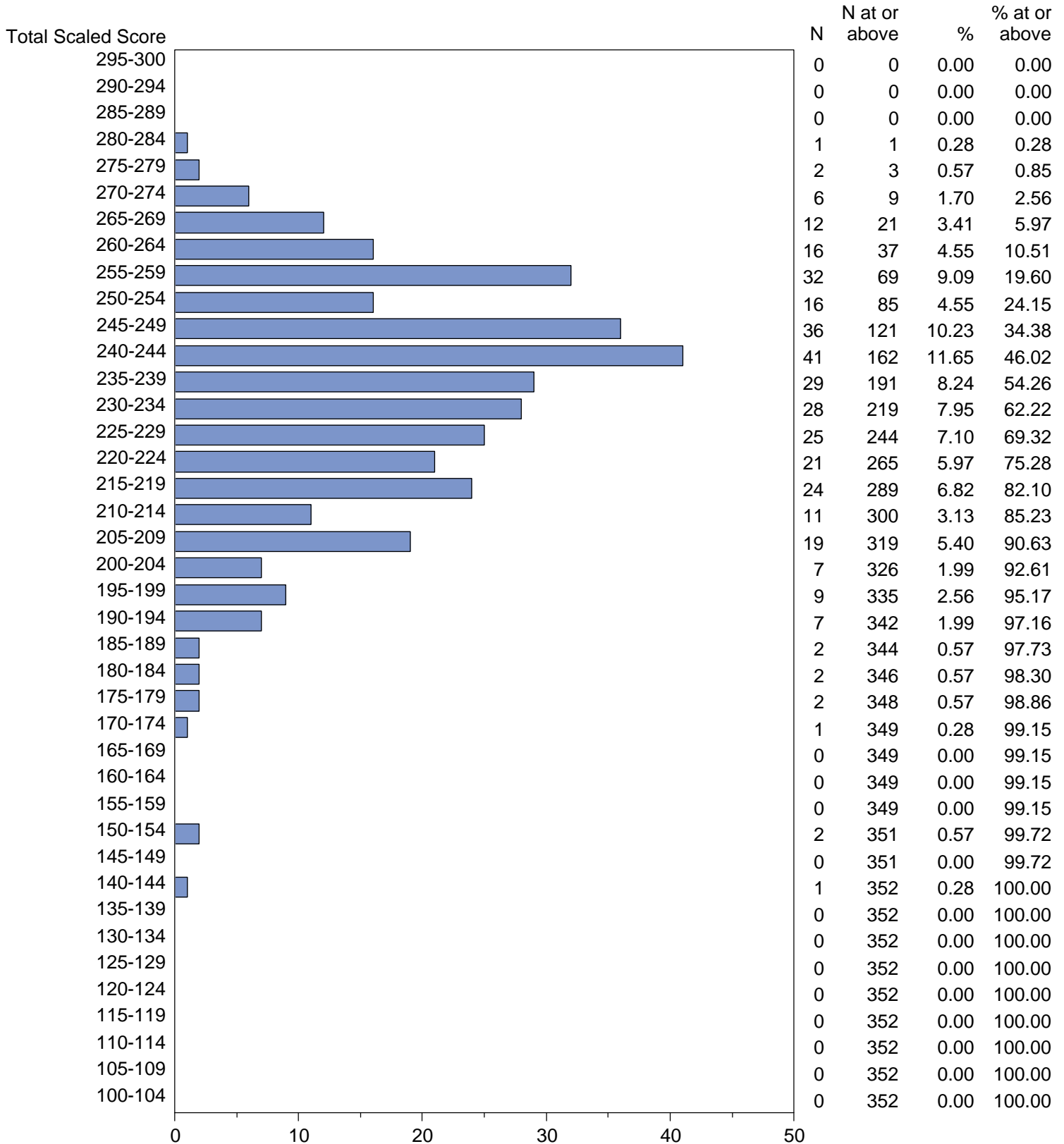
Test Field=039 Reading Subtest II



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

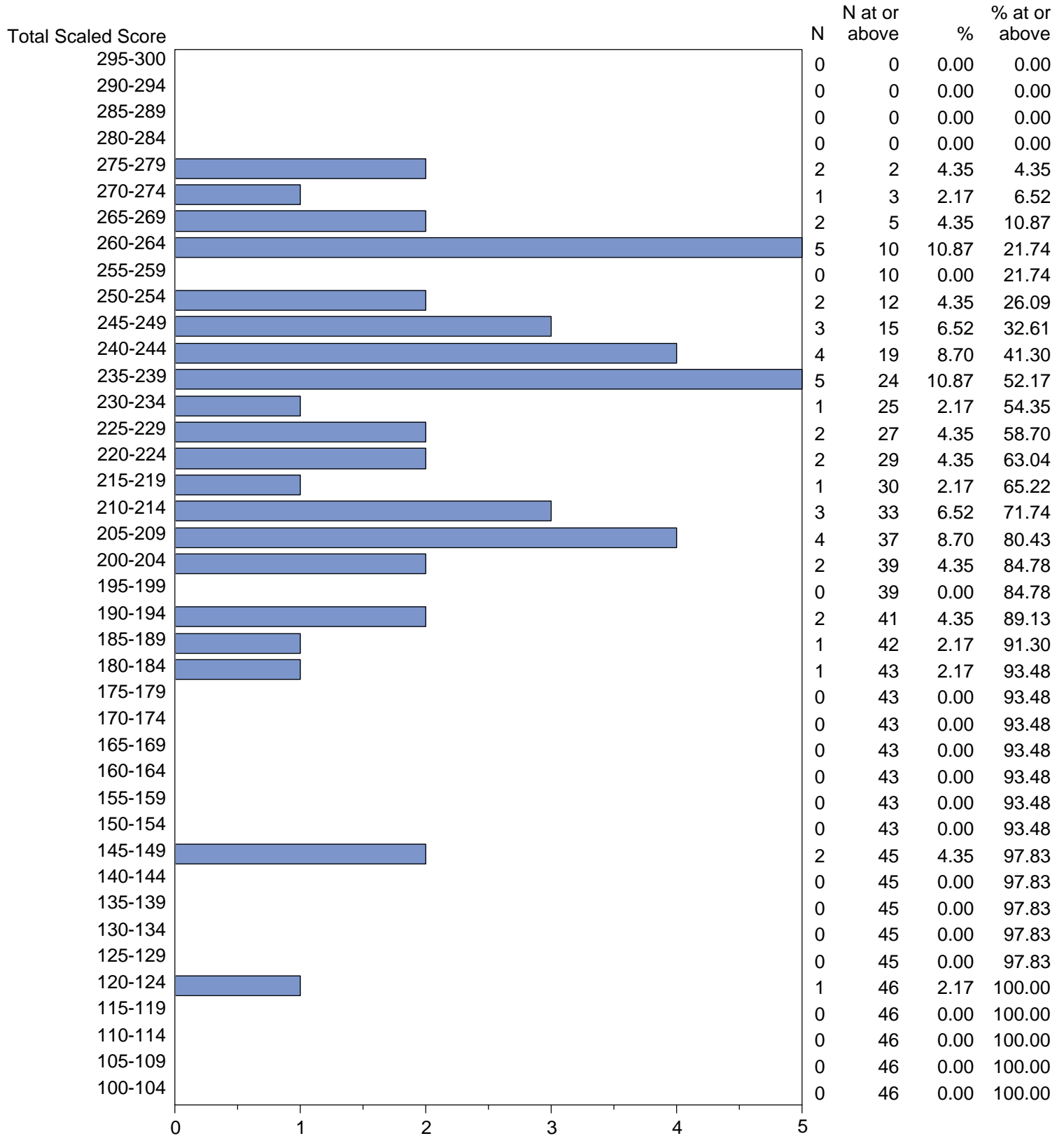
Test Field=040 School Counselor



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

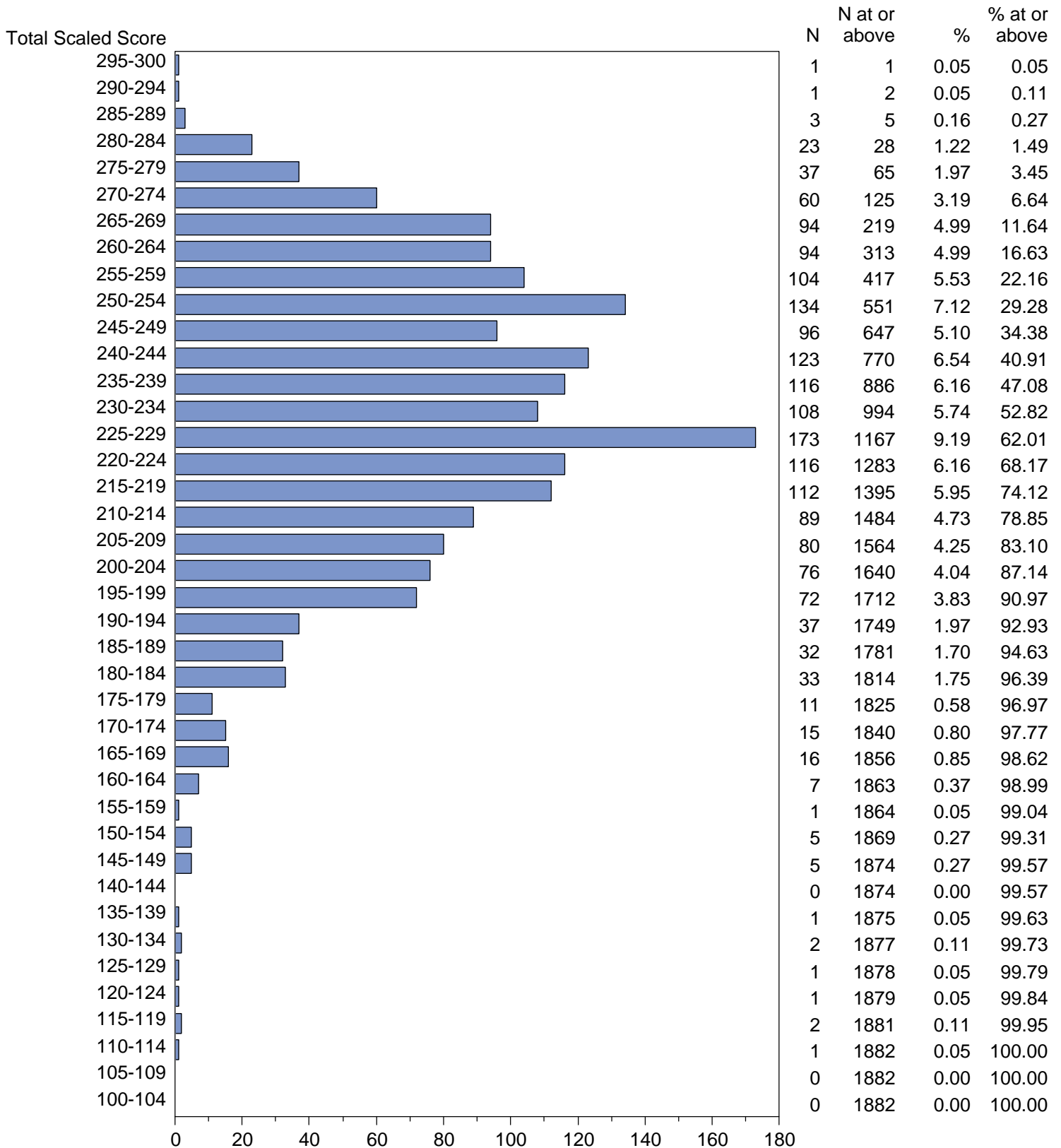
Test Field=041 School Library Media Specialist



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

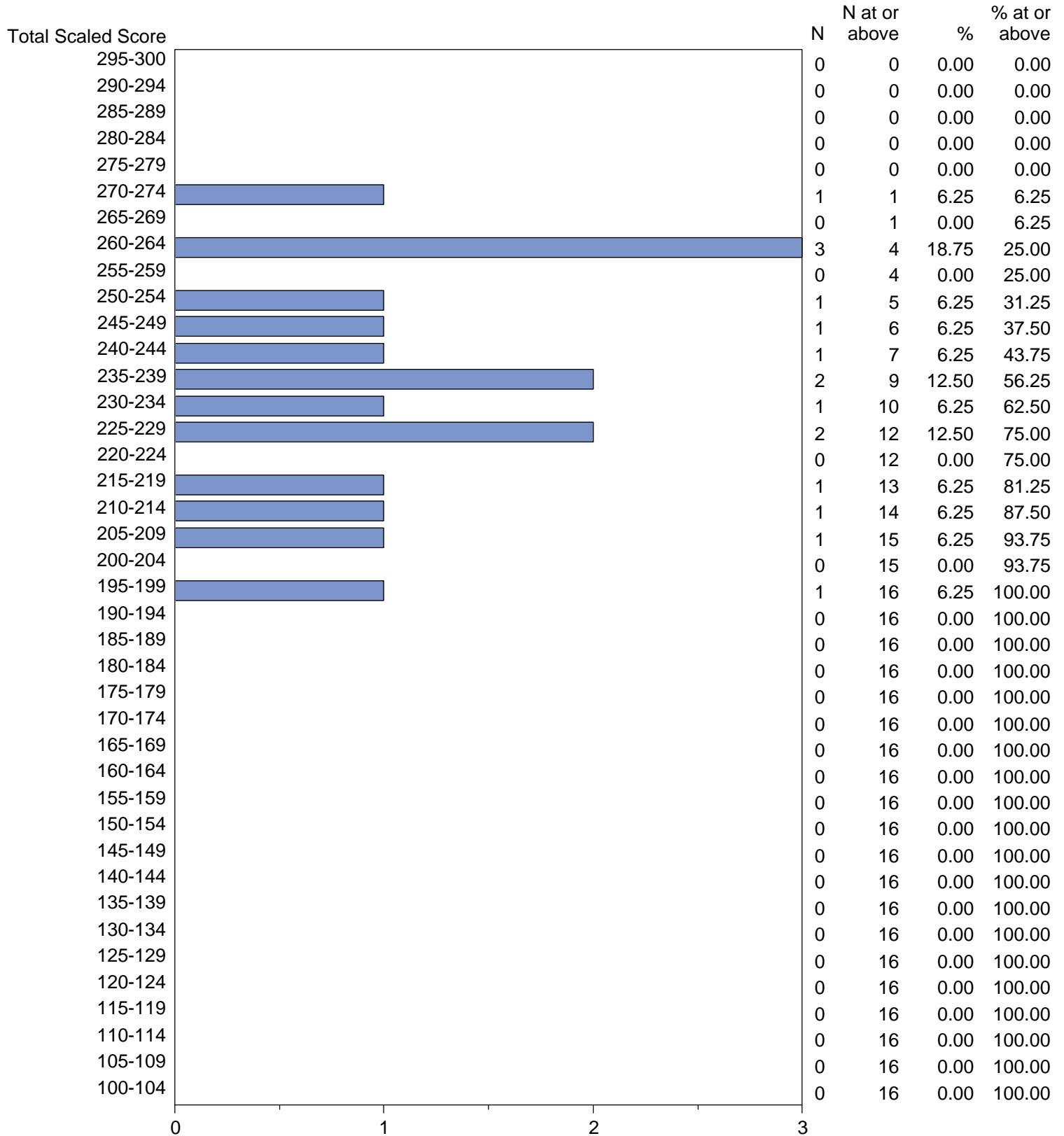
Test Field=043 Special Education



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

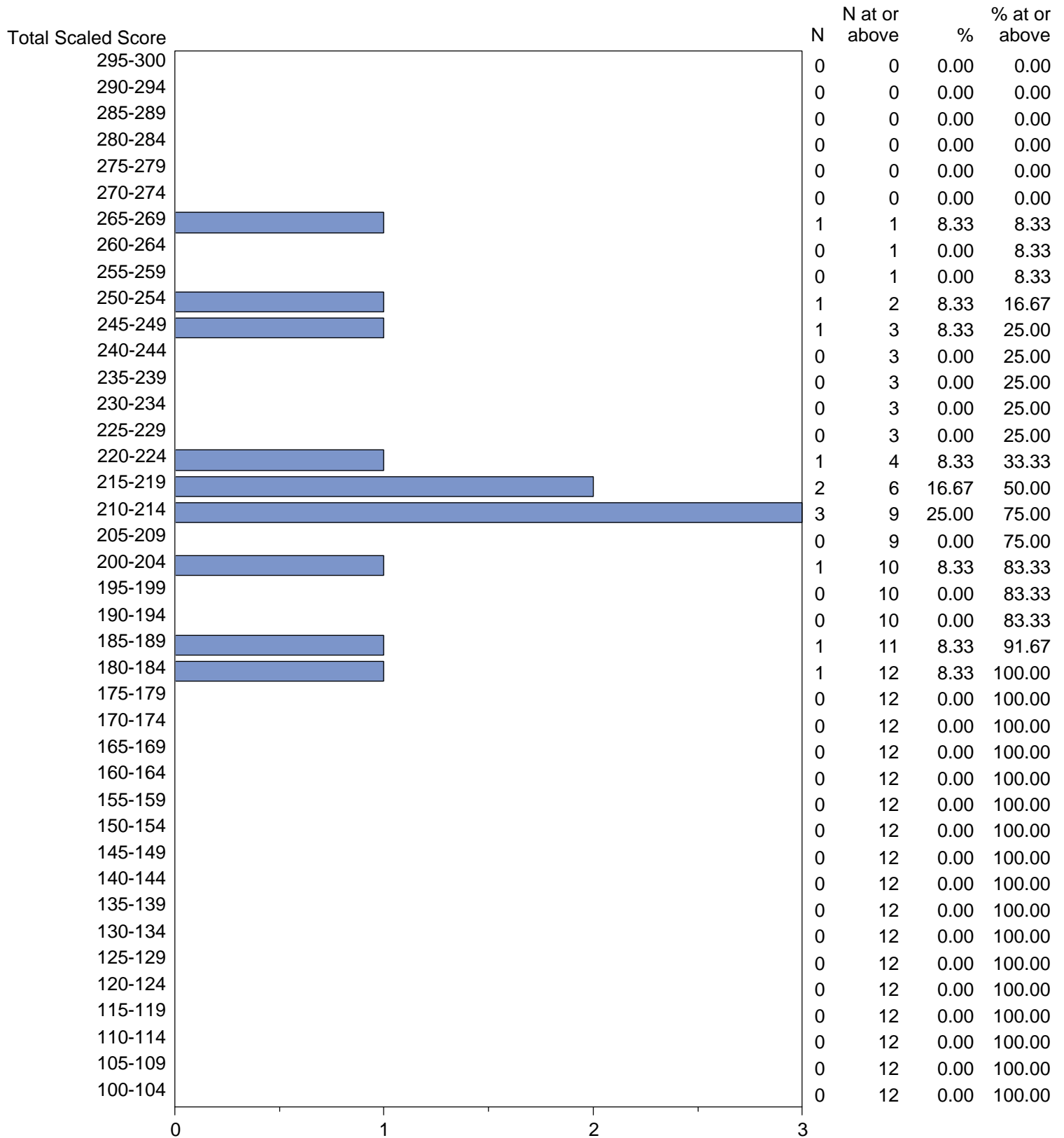
Test Field=044 Special Education Specialist: Deaf/Hard of Hearing



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

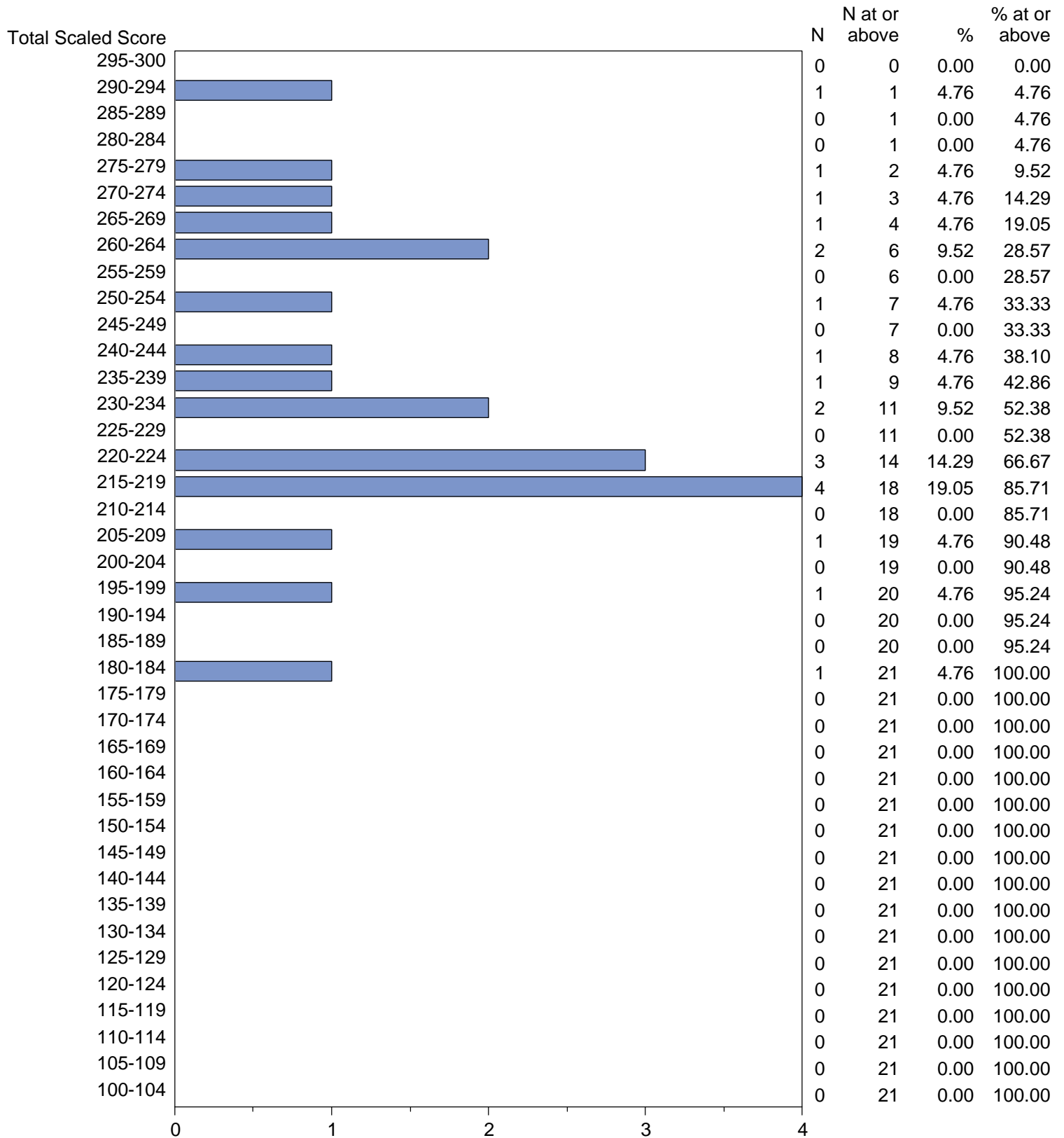
Test Field=045 Special Education Specialist: Visually Impaired



Candidate Performance Statistical Report
September 1, 2014 - August 30, 2015

Total Scaled Score Distribution by Test Field (All Forms)

Test Field=048 Theater



Developed and Produced by



**Evaluation Systems
P.O. Box 226
Amherst, Massachusetts 01004**