Score Reporting in Teacher Certification Testing: A Review, Design, and Interview/Focus Group Study

Heather Susan Klesch

University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/open_access_dissertations

Part of the Education Commons

Recommended Citation


https://scholarworks.umass.edu/open_access_dissertations/251

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
SCORE REPORTING IN TEACHER CERTIFICATION TESTING:
A REVIEW, DESIGN, AND INTERVIEW/FOCUS GROUP STUDY

A Dissertation Presented
by
HEATHER S. KLESCH

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 2010

Education
SCORE REPORTING IN TEACHER CERTIFICATION TESTING:
A REVIEW, DESIGN, AND INTERVIEW/FOCUS GROUP STUDY

A Dissertation Presented
by
HEATHER S. KLESCH

Approved as to style and content by:

____________________________________
Ronald K. Hambleton, Chair

____________________________________
Lisa A. Keller, Member

____________________________________
Robert D. Marx, Member

____________________________________
Christine B. McCormick, Dean
School of Education
ACKNOWLEDGMENTS

This dissertation would not have been possible without the support and encouragement of my family, friends, and colleagues:

To my husband, for his unwavering support, understanding, confidence in me, and constant sacrifice—thank you. Thank you to my son, who may not understand what I do, but understands how important it is to me. In addition, thank you to my mom and dad, for always believing in me.

To my friends, thank you for continuing to encourage me, regardless of the hour.

To Ronald Hambleton, thank you for your scholarly support all these years. You have been an advocate for me and have never failed to help, no matter how small the matter.

Thank you to Lisa Keller and Bob Marx for seeing me through this milestone and guiding me along the way.

Thank you to the educators, testing professionals and doctoral students who assisted me by volunteering their time for this study. I found your interviews and group sessions open, insightful and thought-provoking.

To my colleagues at the Evaluation Systems group of Pearson: I must extend a special thanks and acknowledgment to William Gorth, Paula Nassif, Matthew DiBartolomeo, and Heather Perry. Your continued encouragement and support of my studies have been a welcome motivation. I thank you for your insight, flexibility, and interest.
ABSTRACT

SCORE REPORTING IN TEACHER CERTIFICATION TESTING:
A REVIEW, DESIGN, AND INTERVIEW/FOCUS GROUP STUDY
MAY 2010
HEATHER S. KLESCH, B.A., FLORIDA STATE UNIVERSITY
Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST
Directed by: Professor Ronald K. Hambleton

The reporting of scores on educational tests is at times misunderstood, misinterpreted, and potentially confusing to examinees and other stakeholders who may need to interpret test scores. In reporting test results to examinees, there is a need for clarity in the message communicated. As pressure rises for students to demonstrate performance at a certain level, the communication of scores to the public needs to be examined. Although public school student testing often is placed in the spotlight, this study examines score reporting in teacher certification, which may not have the same complexities of student test score reporting, but does have the equally critical need to effectively communicate scoring information.

The purpose of this study was to create multiple teacher certification examinee score reports based on findings in the literature on educational test score reporting, as well as marketing and design principles, and to conduct interviews and focus groups to gather feedback on the comprehension and preferences in interpreting the designed score
reports and results. Different approaches for reporting test scores were used to design the score reporting materials for a hypothetical teacher certification testing examinee who had not passed. Educators and educational testing professionals were convened and interviewed to review the score reports and offer feedback, suggestions and discussion. The findings are covered in great detail. Using the findings, a final model score report was designed, which was then reviewed with doctoral students in educational measurement.

Through this process, some clear patterns and differences arose. Overall, there was a desire on the educator and doctoral student end to provide as much information as possible, where supported by sound measurement principles. The reporting of raw performance information, as well as accommodating comprehension styles by providing performance information in contextual, statistical and visual ways were requested. Upon addressing these requests, two areas that may not have full clarity and direction remained: The process of converting raw score performance to a scaled score (participants wanted more information on this process), and information provided that could address candidate weak areas, directing examinees to materials that could improve their studies, understanding, and examination performance.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Statement of the Problem</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Purpose of the Study</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Teacher Certification Testing</td>
<td>5</td>
</tr>
<tr>
<td>1.2.1 Basis and Purpose of Teacher Certification</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Score Reporting</td>
<td>8</td>
</tr>
<tr>
<td>1.3.1 Importance of Score Reporting</td>
<td>9</td>
</tr>
<tr>
<td>1.3.2 Psychometric Considerations in Score Reporting</td>
<td>10</td>
</tr>
<tr>
<td>1.3.3 Common Questions About Score Reports</td>
<td>11</td>
</tr>
<tr>
<td>1.3.4 Functions of Score Reporting</td>
<td>12</td>
</tr>
<tr>
<td>1.3.5 Issues in Score Reporting</td>
<td>13</td>
</tr>
<tr>
<td>2. REVIEW OF LITERATURE</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Common Issues In Score Reporting</td>
<td>18</td>
</tr>
<tr>
<td>2.2.1 Language and Terminology</td>
<td>18</td>
</tr>
<tr>
<td>2.2.2 Context</td>
<td>19</td>
</tr>
<tr>
<td>2.2.3 Design and Presentation of Data</td>
<td>21</td>
</tr>
<tr>
<td>2.2.4 Subscores and Scales</td>
<td>22</td>
</tr>
<tr>
<td>2.2.5 Overall Score Report Weaknesses and Strengths</td>
<td>24</td>
</tr>
<tr>
<td>2.2.6 Presentation of Data and Statistics in Other Forums</td>
<td>25</td>
</tr>
<tr>
<td>2.3 Teacher Certification Testing</td>
<td>26</td>
</tr>
<tr>
<td>2.3.1 The Reporting and Use of Scores</td>
<td>26</td>
</tr>
<tr>
<td>2.3.2 Score Reporting in Current Teacher Certification Testing Programs</td>
<td>29</td>
</tr>
<tr>
<td>2.4 Literature Review and Connection to Score Report Design</td>
<td>35</td>
</tr>
<tr>
<td>3. METHODOLOGY</td>
<td>36</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>36</td>
</tr>
<tr>
<td>3.2 Research Questions</td>
<td>36</td>
</tr>
</tbody>
</table>
3.3 Sample Score Report Design ........................................................................ 37
3.4 Interviews and Focus Groups - Research Design ....................................... 38
  3.4.1 Participants.......................................................................................... 39
  3.4.2 Interview/Focus Group Feedback ................................................... 40
  3.4.3 Data Collection................................................................................. 41
  3.4.4 Data Analysis................................................................................. 42
  3.4.5 Application of Results and Findings.............................................. 43

4. RESULTS ................................................................................................................... 44
  4.1 Introduction to the Results ............................................................................ 44
  4.2 Designing the Score Reports and Explanatory Information ......................... 44
    4.2.1 Score Report A and Explanatory Information ..................................... 48
    4.2.2 Score Report B and Explanatory Information ..................................... 53
    4.2.3 Score Report C and Explanatory Information ..................................... 57
  4.3 Designing The Individual Questionnaires .................................................... 62
  4.4 Educator Interviews ...................................................................................... 66
    4.4.1 Overview........................................................................................ 66
    4.4.2 Participant Demographics.............................................................. 69
    4.4.3 Score Report A Feedback .............................................................. 70
    4.4.4 Score Report B Feedback............................................................. 77
    4.4.5 Score Report C Questionnaire Results ........................................... 85
    4.4.6 Other Feedback From Educator Interviews ................................... 93
  4.5 Focus Group With Educational Testing Professionals ................................. 96
    4.5.1 Overview........................................................................................ 96
    4.5.2 Participant Demographics.............................................................. 96
    4.5.3 Score Report A Feedback .............................................................. 97
    4.5.4 Score Report B Feedback............................................................. 102
    4.5.5 Score Report C Feedback............................................................. 107
    4.5.6 Other Score Report Feedback ...................................................... 111
  4.6 Processing Interview and Group Feedback and Designing Score
    Report D................................................................................................ 112
  4.7 Conducting the Focus Group with University Doctoral Students .............. 117
    4.7.1 Overview...................................................................................... 118
    4.7.2 Demographics .............................................................................. 118
    4.7.3 Score Report A Feedback .............................................................. 120
    4.7.4 Score Report B Feedback............................................................. 124
    4.7.5 Score Report C Feedback............................................................. 127
    4.7.6 Other Feedback ............................................................................ 132
    4.7.7 Score Report D............................................................................. 132

5. DISCUSSION ........................................................................................................... 136
  5.1 Summary of Findings.................................................................................. 136
  5.2 Conclusions ................................................................................................. 138
  5.3 Limitations of the Study............................................................................. 144
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1. Demographics of Educator Participants (Interviews)</td>
<td>69</td>
</tr>
<tr>
<td>4.2. Distribution of Educator Responses to Examinee Expected Performance</td>
<td>89</td>
</tr>
<tr>
<td>4.3. Demographics of Educational Testing Professionals (Focus Group)</td>
<td>97</td>
</tr>
<tr>
<td>4.4. Distribution of Educational Testing Professionals Responses to Examinee Performance</td>
<td>108</td>
</tr>
<tr>
<td>4.5. Demographics of University Education Students/Representatives (Focus Group)</td>
<td>119</td>
</tr>
<tr>
<td>4.6. Distribution of Student Responses to Examinee Performance</td>
<td>129</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Examinee Score Report A</td>
<td>50</td>
</tr>
<tr>
<td>4.2</td>
<td>Understanding Your Examinee Score Report for Score Report A</td>
<td>52</td>
</tr>
<tr>
<td>4.3</td>
<td>Examinee Score Report B</td>
<td>55</td>
</tr>
<tr>
<td>4.4</td>
<td>Understanding Your Examinee Score Report for Score Report B</td>
<td>56</td>
</tr>
<tr>
<td>4.5</td>
<td>Examinee Score Report C</td>
<td>59</td>
</tr>
<tr>
<td>4.6</td>
<td>Understanding Your Examinee Score Report for Score Report C</td>
<td>61</td>
</tr>
<tr>
<td>4.7</td>
<td>Individual Questionnaire: Examinee Score Report A</td>
<td>63</td>
</tr>
<tr>
<td>4.8</td>
<td>Individual Questionnaire: Examinee Score Report B</td>
<td>64</td>
</tr>
<tr>
<td>4.9</td>
<td>Individual Questionnaire: Examinee Score Report C</td>
<td>65</td>
</tr>
<tr>
<td>4.10</td>
<td>Examinee Score Report D</td>
<td>116</td>
</tr>
<tr>
<td>4.11</td>
<td>Understanding Your Examinee Score Report for Score Report D</td>
<td>117</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1 Background

Educational testing is not a new concept to the general public. With the No Child Left Behind Act of 2001, testing in education has taken a front-row seat in many public forums including the media, educational funding decisions, policy implementation, and for some states, high-school graduation requirements for students. The emphasis of this public exposure, however, has primarily been in the K–12 public school arena. Annually, millions of students in public schools are administered their particular state’s student test.

Testing for credentialing—teacher certification testing in particular—has less exposure to the general public than K–12 student testing does, yet it shares many of the same issues that are faced by those who develop, validate, administer, and interpret the results of public school assessments. Great efforts are made within individual states to comply with national requirements of having “highly qualified” educators. However, despite these efforts, the question remains as to whether the certification examination results are being clearly conveyed to examinees and other stakeholders.

1.1.1 Statement of the Problem

The reporting of examinee scores on educational tests is at times misunderstood, misinterpreted, and potentially confusing to examinees and others who must use the resulting scores to make a decision. As Hambleton (2002) states, “Over the years, I have been struck by the contrast between the efforts and successes in producing sound
technical assessments, drawing samples, administering the assessments, and analyzing
the assessment data, and the efforts and success in disseminating the assessment results”
(p. 193). The effort and research devoted by testing companies to accurate and defensible
test development, administration and scoring procedures have not always then been
extended and applied to the actual reporting of results.

The issues with score report interpretation and understanding are not unique to
public school student testing and share common traits with teacher certification testing.
For public school examinees, much of the focus is on understanding where a student may
lie in the spectrum of performance-level descriptors such as “Needs Improvement,”
“Basic,” or “Proficient” in addition to analyses that look at performance over time (e.g.,
vertical score scale reporting over multiple years and grades). For teacher certification
testing, the focus is on whether or not an examinee has met a specified criterion and has
“passed” the examination.

Given the different uses and purposes of score reports, this study aimed to gather
information from the field on score report design, the level of comprehension in
interpreting score reports, and feedback regarding score reporting elements desired by
stakeholders in a teacher certification testing environment.

1.1.2 Purpose of the Study

The purpose of this study was twofold: First, this study aimed to create different
score report designs based on a review of the literature on educational score reporting and
design. Second, this study offered an opportunity to gather information from the field by
conducting interviews and focus groups in order to gauge a level of comprehension in
interpreting score report results; and soliciting feedback from educators, educational testing professionals, and students regarding what they may wish to see on a score report for a teacher certification testing program. It is understood that ideally an extension of this study would include feedback obtained from actual teacher certification examinees. While this study does not utilize this feedback group, the feasibility of this and suggestions for examinee groups/subgroups are discussed further in Chapter 5.

While the primary focus of this study was based in the context of teacher certification, the review incorporated literature on score reporting in K–12 student testing in that some of the design principles, terminology, and issues faced are common to both teacher certification and student testing. For example, testing agencies making decisions on score reporting design for student testing must keep in mind the level of understanding of stakeholders. Terminology used must not be foreign or so difficult to navigate that the reader is left wondering what steps should be taken next. In some states, a student’s score places them in a descriptive category (e.g., Basic, Proficient, Advanced) and it must be clearly communicated whether the particular student’s category is sufficient for meeting state requirements such as high school diploma/graduation requirements.

Similarly, score report design in teacher certification must clearly communicate whether the examinee passed and has met the assessment requirement for the particular state in which they are seeking a job. While there is certainly less volume of literature on teacher certification score reporting, by incorporating studies and findings in the K-12 arena, this study approached the teacher certification score report design with as wide a research-based foundation as possible.
Through applying the literature review findings, three score reports were designed, including an explanatory page for each score report. Each score report and explanatory page incorporated various components of the findings in several ways: From design elements such as use of color, font size, graphics size, typeface and layout, to more statistical and technical-based elements such as level of reporting, statistics reported, and supporting glossary/interpretive material. The three score reports designed were of varying substance and depth, reflecting different ways to present the same results of a non-passing teacher certification examinee.

The mock score reports designed were presented first to two groups of individuals:

- 16 educators (public school and college/university) across various subject matter areas throughout the United States, and
- 6 assessment staff involved in designing, interpreting, or advising on score reporting and scoring procedures.

Participants were given the opportunity to review each mock score report and corresponding explanatory page, individually answer some questions regarding the reports, then discuss the reports and in general offer communicative guidance regarding what aspects of the mock score reports may or may not be fully understood. Comprehension-type questions were asked in order to assess whether the participants were able to interpret what was presented and whether they could infer what might be done next to help the sample examinee. These groups proved crucial in gathering a wealth of real-world, field-generated information regarding score report interpretation, misinterpretation, and utility. The feedback and focus group findings and

recommendations were analyzed to present as a conclusion a research-based proposed score reporting design for a sample teacher certification test. Using these findings, the final model score report was designed. A final focus group was conducted, consisting of ten university doctoral education students who were knowledgeable in areas of measurement and psychometrics. These students were first presented with the initial three score reports, and then the final score report.

1.2 Teacher Certification Testing

Teacher certification testing is the act of requiring candidates who are interested in serving as an educator in public schools to demonstrate competency in the knowledge and skills for the area(s) in which they wish to teach. Most recently, the demonstration of competency through a state-approved assessment is rooted in the need to demonstrate compliance with the No Child Left Behind Act of 2001 (as outlined below). While the specific requirements may vary from state to state, ultimately, the outcome is the same—an educator who is interested in teaching must obtain full teacher certification in his or her state. Typically this requires a demonstration of having met that state’s criterion of performance that is deemed acceptable for an “entry-level” educator in the particular subject matter area.

1.2.1 Basis and Purpose of Teacher Certification

Teacher certification testing did not begin with the No Child Left Behind Act. In fact, the legalities of certifying teachers and establishing teacher preparation programs date back to the colonial days.
LaBue (1960) provides an overview of the history of teacher certification pre-1960, dividing the history up into four chronological periods, as outlined below.

**Colonial period to 1789.** During this period, religion as well as local government played an important role in the determination of teacher licensure. While there was some concern regarding the licensing of teachers, there was little to no basis of concern regarding those teachers’ education or knowledge and skills. At times, the granting of licensure included an oath of loyalty by the teacher toward their government and country.

**1789 to 1860.** In this timeframe, movements such as suffrage, industrialism, nationalism, and the labor movement served to expand the idea of public school systems. Additionally, first attempts at teacher education and the formation of “normal schools” helped shape the future of teacher certification programs. The formation of state school systems also led to district-based levels of decision making in matters of certification.

**1860 to 1910.** This time period saw an increase in number of school enrollments, as well as “the establishment of teachers colleges, and the beginning of schools of education and departments of education in universities and liberal arts colleges” (p. 147). Near the end of this period, the groundwork for teacher certification as known currently had been established.

**1910 to 1960.** The final period from LaBue’s overview reviews the years since 1910, which saw multiple developments that have influenced present-day teacher certification and movement toward standards and assessments. The effect of world events on the profession of teaching is prevalent throughout this period.

While World War I saw a shortage of qualified educators, the depression of the 1930s saw an overabundance of teachers looking for employment. As LaBue outlines,
“this oversupply of teachers made possible the continuing rise in minimum standards and a corollary increase in formal teacher-preparation requirements” (p. 161).

With World War II, the nation again experienced a severe shortage of teachers resulting in the establishment of emergency certificates being issued as necessary, with the rationale as “it is better to maintain and even raise requirements while issuing emergency certificates to fill the unemployment gaps in the public schools, rather than to generally lower all standards to attract more teachers” (LaBue, 1960, p. 163).

The expansion of American education, especially by 1960, saw an increase in student enrollments and thus a demand for educators. The work of voluntary associations and the formation of the American Association of Colleges for Teacher Education (AACTE), as well as the National Council for Accreditation of Teacher Education (NCATE) served to establish a governing body on the accreditation of teacher education institutions. Additionally, the use of proficiency examinations in place of only requisite preparation program coursework gained support (LaBue, 1960, p. 171).

1960-1980s. During the 1960s, “most states certified prospective teachers on the basis of successful completion of a teacher education program of study” (Rubinstein, McDonough, & Allan, 1986, p. 18), however change in the 1970s brought a shift towards the use of assessment to support competency-based education. It was during this time that a number of states implemented changes to the policies and practices in multiple phases of the teacher certification programs: admission, completion, first year of incumbency in a classroom position, and during certification renewal (Rubinstein, McDonough, & Allan, 1986, p. 18). Between 1977 and 1981, a number of states required that program graduates pass a test sponsored by the state in order to earn certification.
Graduating from an accredited educator preparation program was no longer an acceptable sole criterion for students; demonstration of competency in subject matter knowledge was also required.

1990s to Present. In 2001 a reauthorization of the Elementary and Secondary Act of 1965 occurred, called the No Child Left Behind Act of 2001, or NCLB. NCLB’s focus on standards-based education and accountability is still in effect. The impact to teacher certification was pronounced in that the law required states to demonstrate how all teachers of core academic subjects in the classroom would be designated as “highly qualified.” Per NCLB, highly qualified “is determined by three essential criteria: (1) attaining a bachelor’s degree or better in the subject taught; (2) obtaining full state teacher certification; and (3) demonstrating knowledge in the subjects taught” (*Highly Qualified Teachers for Every Child*, 2006). With funding attached to the implementation of this demonstration of highly qualified, teacher certification testing became incremental for states in providing evidence of how their state would meet the demands and requirements of NCLB.

1.3 Score Reporting

Just as assessment and teacher certification are important components of education and accountability, the reporting of scores for these initiatives is equally critical. Given the importance of score reporting and that scores may serve different functions, there is a need to ensure clarity in the information communicated.
1.3.1 Importance of Score Reporting

Score reporting is a crucial aspect of educational testing. Without the score report, results would not be conveyed in a standardized way and may be left open to interpretation. The *Standards for Educational and Psychological Testing* (1999) contain a number of references to the importance of reporting scores and providing accompanying interpretive information:

Standard 5.10—When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, and the precision of the scores, common misinterpretations of test scores, and how scores will be used. (p. 65)

Standard 6.3—The rationale for the test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified. (p. 68)

Standard 8.8—When score reporting includes assigning individuals to categories, the categories should be chosen carefully and described precisely...(p. 88)

In addition to being a crucial component of testing programs, the interpretation and communication of scores are job responsibilities for any educator, especially those involved in the remediation of students or the explaining of test results to parents, students, or others. As outlined in the *Standards for Teacher Competence in Educational Assessment of Students* (American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990), “[t]he teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods” (p. 5).
Plake, Impara, & Fager (1993) looked at the results of a teacher assessment to measure the competency areas described in the *Standards for Teacher Competence*. Through this study they found “nearly 85% of the teachers indicated that they were at least somewhat interested in becoming more proficient in interpreting test scores and in student assessment in general” (p. 12).

In order to interpret results, the educator must have an underlying foundation in basic statistics and measurement concepts. This is an underlying foundation present in current pedagogical standards. As outlined in the National Board for Professional Teaching Standards for Generalist educators across all grade levels (2001), assessment is a critical portion of educator expectations and what an educator should be able to know and do. The appropriate use of assessment, interpretation of results, and communication to stakeholders are expected skills for the educator.

1.3.2 Psychometric Considerations in Score Reporting

Score reports serve the function of reporting the examinee performance on an assessment. Considerations must be made in designing a score report, and those considerations are not solely based on what may communicate the clearest message. The testing purpose, or the construct being measured, is an important consideration. Is the assessment looking to place a student from among other students, or is it designed to indicate a level of proficiency or mastery? Is the assessment result based on a pre-determined criterion, or based on the performance of other examinees? These are important considerations when determining what information will be reported on an assessment score report. Other psychometric considerations include the level of reporting
1.3.3 Common Questions About Score Reports

Regardless of the psychometric considerations that may go into designing the results that are provided back to examinees, it is understood there may be questions from examinees that do not pass. Some examples of questions that might be asked by an examinee receiving a score report for a teacher certification examination are:

- “How many questions do I need to correctly answer in order to pass?”
- “I did not pass, how many questions did I miss by?”
- “Why can’t I just see the questions I missed?”
- “What is a scaled score?”
- “I know I did not pass, but where should I focus my studies in order to pass next time?”
- “What do you mean there is ‘error’ in every assessment?”

While these questions and the answers to them may vary from testing program to program, they are meant as examples to illustrate that some questions that arise out of score reporting are common to almost any educational testing program and can be applicable to both the K–12 student and teacher certification testing settings. Test publishers and organizations are aware of these types of questions, and typically attempt to help explain some of these concepts by publishing *Frequently Asked Questions* or other interpretive documents that accompany a score report or testing program.
1.3.4 Functions of Score Reporting

As described by Kolen (2006) the functions of a score report go beyond just a means to communicate an examinee’s performance:

Scores reported to test users are one of the most visible components of educational testing programs. Some reported scores may reflect an examinee’s standing relative to different reference groups, other scores might reflect performance relative to standards set by subject matter experts, and still others might reflect performance on subparts of a test. The reason that multiple scores are often reported for an educational test is that such tests often have many purposes.” (p. 155)

Accordingly, score reports can serve various functions as outlined below.

**Results.** Primarily, a score report serves to provide the result of an assessment. Simply speaking, this information may include a raw score, the number of questions an examinee correctly or incorrectly answered. In other cases, test results may be conveyed as the percent correct out of the total percentage.

**Context.** Score reports may provide context for results by including information about other test takers’ performance (such as those provided for norm-referenced tests, or an average performance of test takers from that date) or the performance of an examinee in relation to an established criterion or standard (such as those provided for criterion-referenced tests).

**Categorization.** Through the use of scaling, score reports can help put all test takers on the same scale as well as assist in the interpretation by governing bodies (such as licensure or credentialing commissions) by placing an entire battery or program of tests on the same reporting scale regardless of the varying subject matter areas or test designs. Descriptors may be used to help place students in categories (such as Basic,
Proficient, and Advanced) in order to provide a roadmap for where the examinee might be headed. These descriptors may also assist with reporting group results and movement of students from one descriptor to the next.

**Diagnostic.** A score report may provide information that would help indicate where areas of deficiency may appear within an examinee’s demonstration of subject matter knowledge. This might be represented as an indication of what areas of strength or weakness were perceived through the items that the examinee correctly or incorrectly answered. This may help outline what content areas need improvement or what aspects of a response may have been lacking.

1.3.5 Issues in Score Reporting

As important as score reporting is, misconceptions abound, as do misinterpretations.

**Too much information.** Score reports that attempt to squeeze a lot of information into a compressed reporting space may ultimately have the opposite effect—too much information that results in not enough information actually understood. The important pieces of information (e.g., how the examinee performed, whether the examinee passed, whether there are specific areas of concentration suggested for examinees who do not pass) may be buried if presented along with an overwhelming amount of extraneous information.

Additionally, an attempt to provide a lot of information may actually be detrimental if the information provided goes against some fundamental measurement principles, that is providing subscores on very small sets of items. These small sets of
items rarely have enough reliability to make a judgment on whether they in fact are areas of weakness needing great attention. If these subscores are reported, one might also ignore other subareas in which they apparently performing well, even though those subscores were a small set of items an not robust enough to clearly say the examinee has mastered that area.

**Not enough information.** In contrast, some testing programs score reports may provide only a limited amount of information. While examinees and other stakeholders may desire an in-depth analysis of the examinee performance, that may not be useful information for a variety of reasons. As mentioned, reporting on a small subset of items may not be reliable or informative in that a general statement regarding examinee proficiency on that subset of content is not based on a robust quantity of items. Reporting a raw score of correct and incorrect responses may also be desired, but not always meaningful or informative. What if a raw score performance on a section of a test is high, but that section is weighted lower in comparison to other sections on the same test? Not enough contextual information may result in accurate information presented on the score report but an inaccurate interpretation of what the results are stating.

**Reporting of raw scores.** One purpose of an assessment (regardless of type) is to have an examinee demonstrate something in response to a command or question, and to report back information on that demonstration. In its simplest form, score reporting tells you “right” and “wrong” (e.g., “out of 20 questions, you got 10 correct.”). Again, in its simplest form each question is “worth” the same amount, so a performance of 10 out of 20 would be 50% correct.
Confusion may also arise when multiple assessments are administered over time. In these cases, score reporting may become less straightforward. This is primarily due to the process of equating, or when a later test form is adjusted to be “equal” to a base (earlier administered) form. An examinee’s performance of 10 correct out of 20 on a test one day might be the same numerical performance on a future test, but that future test may be of a greater difficulty. Tests rarely can be built at precisely the same difficulty from test form to test form, while still maintaining all other test form blueprinting needs and program rules, and thus the same performance on the same “test” but on different forms may actually be very different.

Example: Jane takes a 40-item test in September and gets 20 items correct. Jane then takes the same 40-item test in October, but the test is a different form and is slightly more difficult. Jane still gets 20 items correct. Should Jane be rewarded for getting 20 items correct on the harder form (October), or should she be treated the same as when she got 20 correct on the easier form (September)? The same holds true for the score that is reported—if it can be shown that a demonstration of 20 correct on the hard test is in fact a demonstration that Jane has exhibited a higher level of knowledge and skills than the 20
correct performance from September, then Jane’s 20 correct in October is worth more than the September performance. The score report, then, should reflect this difference.

The full story is not reflected in a raw score. This issue lies at the root of many reporting misunderstandings—why 20 correct on one test form is not the same as 20 correct on another test form. Harris (2003) stated the following about the raw score, “[F]or most tests, the raw score is the fundamental score. Ironically, the raw score is seldom the score on which decisions are based; for many tests, it may not even be reported” (p. 3).
CHAPTER 2
REVIEW OF LITERATURE

2.1 Introduction

Despite what it may seem, there is a growing body of literature on score reporting. The literature varies widely from score report design, program-specific issues with score reporting, determining the finiteness of reporting scores (e.g., reporting sub scores), defining multiple proficiency descriptors (e.g., what does it mean to be “Basic”), making score reports more useful, and the proper/improper use of test results.

Predominantly the literature focuses on the K–12 public school arena, which is understandable given the wide audience being called upon to interpret and make meaning of assessment performance in public schools. Literature on score reporting in credentialing and teacher licensure is much less substantial; however, fundamental score reporting concepts in the K–12 arena are still applicable and warrant attention.

The literature review presented in this chapter will focus on publications that discuss elements of score report design such as language, terminology, layout, misconceptions, and other areas that may contribute to consideration in the design of the mock score reports discussed in subsequent chapters. Guidelines for preparing tables and figures, the display of data, use of subscores and scales are presented. A mention of statistics and data display in mainstream society is also covered, illustrating the applicability beyond just the field of education.

Concluding the literature review is an overview of teacher certification testing reporting, use of scores, and an examination of score reports and supporting information currently available for selected teacher certification examination programs.
2.2 Common Issues In Score Reporting

Score reporting is not perfect, nor is there a single answer to address potential problems. As summarized by Aschbacher and Herman (1991) in looking at state assessment reports,

… some state reports are fairly comprehensive and well used, but many could be far more effective and informative. They tend to be very dry, limited in scope, poorly organized from the reader’s point of view, and ineffectively formatted, with little to capture or focus the reader’s attention. Further, many reports fail to include sufficient explanation or documentation to prevent reader misunderstandings of the meaning of achievement results. And finally, most reports do not relate test scores to important contextual information about student behavior, attitudes, performance on other measures, the learning environment, or community variables (p. 3).

Given the varying nature of potential issues, it is important to point out some common groups of possibilities.

2.2.1 Language and Terminology

Hambleton (2002) presents an overview on score reporting for the National Assessment of Educational Progress (NAEP), conducting a study that obtained feedback from stakeholders on understandability of the score reports. Findings uncovered “one of the problems in understanding the text was due to the use of some statistical jargon” and participants voiced a desire for interpretive guidelines or a “glossary of basic terms” (p. 198). An interesting observation was that the phrase “statistically significant difference” did not have clear meaning to participants—some tried to quantify it in terms of actual percentage point differences, while more commonly the term was thought to be “big and important differences” (pp. 199–200).
The concerns raised in this study are prevalent today, in that there is a desire to receive information, but a need to understand its interpretation: “ways need to be found to balance statistical rigor in reporting with the informational needs, time constraints, and quantitative literacy of intended audiences” (Hambleton, 2002, p. 200).

Accessibility to information contained in score reports hinges upon the ability to understand the language and terminology used in the report. As outlined by the Council of Chief State School Officers (CCSSO) in describing the design of report cards for effective accountability reporting, “…it is important to think about the audience. Certainly, with a wide variety of stakeholders who will be motivated to look at a report card, the document must contain a wide range of information…that is presented in a clear and easy to read format, free of jargon (both verbal and numerical), and points to sources of other relevant information” (Forte Fast, Blank, Potts, & Williams, 2002).

### 2.2.2 Context

In a chapter on reporting and interpreting test results, Harris (2003) gives an overview of norm-referenced terminology and score reporting elements, as well as an overview of derived score terms and elements. These contexts are important in understanding some fundamental differences between reporting in a norm-referenced examination framework (where an examinee’s performance is being characterized as compared to other examinees) and a criterion-referenced framework (where an examinee’s performance is characterized in relation to a set criterion). Harris cautions against interpreting scores “in a vacuum” and offers some potential considerations in interpreting any examination score—such as speededness, administration conditions, and
content (p. 7). Building on that idea, Harris provides an overview on test content as the most important consideration in interpreting scores and some valuable information on factors that help assist with deriving meaning from scores.

These ideas are echoed by the *Standards for Educational and Psychological Testing* (1999):

Test scores, per se, are not readily interpreted without other information, such as norms or standards, indications of measurement error, and descriptions of test content. Just as a temperature of 50° in January is warm for Minnesota and cool for Florida, a test score of 50 is not meaningful without some context (p. 62).

Along the same lines, context of what a specific score or performance actually means is crucial in interpretation. With various states using differing performance indices, it is important to know what “Basic,” “Proficient,” and “Advanced” translate to on an examinee level.

Context is also a guiding force for the type of statistic presented, or the way in which it is presented. As Hambleton (2002) points out,

One of the ways of making statistical results more meaningful to intended audiences is to report the results by connecting them to numbers that may be better understood than test scores and test score scales. For example, To relieve the concern many persons had about flying after the TWA crash a few years ago, the airlines reported that there is a single plane crash every 2,000,000 flights. In case the safety of air travel was still not clear, the airlines reported that a person could expect to fly every day for the next 700 years without an accident. Probably some people felt more confident after hearing these statistics reported in this way. Knowing that the probability of being in a plane crash is less than .0000005 may not be so meaningful. (p. 194)

This idea of finding different ways to effectively communicate the same statistical information was a driving force underlying the design of the score reports for this study.
2.2.3 Design and Presentation of Data

Addressing the specific element of graphic design, Hambleton (2002) presents an introduction of a multi-step instructional module for “the design of figures and tables and specific guidelines for preparing tables and figures” (p. 201) for presentation in score reports. These steps are as follows:

1. Keep presentation clear, simple, and uncluttered.
2. Ensure that the graph is able to stand alone.
3. Ensure that the text complements and supports the graph.
4. Plan the graphical presentation.
5. No form of graph is more effective in all respects than all other forms (however some comparisons of bar charts, circles, squares, cubes, lines and color are mentioned).

(pp. 201–202)

Brown (2001) also looked to examine reporting preferences of teachers and parents for the New Zealand numeracy and literacy curricula (for students in years 5 to 7) and found that in communicating results to stakeholders, the “judicious use of graphical communication principles” is beneficial (p. 4). “These principles have to do with designing reporting mechanisms in accordance with the kind of information, the purpose of the information, and the audience being addressed” (p. 4).

Brown also made use of color in reporting scores and performance with the green display “suggestive of a green traffic light, signifies that these are areas in which the teacher can confidently give the student more work and which should no longer dominate instructional content or time” as opposed to use of red that “clearly signals that this is a
danger area that the teacher needs to investigate to determine the nature of the gap in the student’s achievement” (p. 8). These same principles have also been used in business displays of data, or “dashboard” displays in order to quickly identify risk areas or areas that need to be monitored carefully.

Another study (Wainer, Hambleton, & Meara, 1999) concluded that many of the issues associated with the display of data might be addressed by a more careful consideration during the design phase and subsequent field testing of the reports to identify strengths and weaknesses.

2.2.4 Subscores and Scales

The reporting of subscores and their utility are not without debate. Subscores can be defined as a score provided on a subdivision of content within a total test. Subscores may be presented as scaled scores, raw score performance, or even as performance indices (visual “estimates” of one’s performance on each subsection of content).

Puhan, Sinharay, Haberman, and Larkin (2008) suggested methods of reporting subscores and also asked whether reporting those subscores would provide any additional information than the total score. The authors point out:

…the desire to receive subscores at the examinee and/or institutional level is even stronger in certification and licensure testing because a small difference in the total score of these tests can make a difference in the pass or fail status of the examinees. Therefore the general consensus seems to be that examinees attending remedial training may get a slight edge (i.e., improving the totalscore) by improving on subcontent areas where they may be weaker (p. 1).

In the study, Puhan, Sinharay, Haberman, and Larkin examined eight certification tests and found that the subscores reported may not be important as they did not provide useful
information beyond that of the total test score. Attention was paid to the need to clearly define subscores prior to their implementation, and the authors asserted, “subscores, if defined in raw score units, are not directly comparable across different forms of the test. This finding is also true of augmented subscores. Therefore an important issue with reporting subscores, but for individuals and institutions, is that subscores have to be equated and/or scaled for comparability” (p. 29).

As pointed out by Zieky (2002), “one concern is that subscores based on relatively few items tend to be unreliable, and the reliability of the difference between a pair of subscores goes down as the correlation between them goes up. Since content areas of a test tend to be highly correlated, the reliability of the difference between subscores tends to be low. Be careful not to report unreliable diagnostic information.” (p. 9).

As created by Mislevy (1998) and reinforced by Hambleton (2002), the concept of market-basket score reporting also illustrates the importance of scales and interpretation. The price of a market basket of food can be a measure of economic change in that if it is reported over time it can display whether food prices are increasing or decreasing and at what rate for this fixed set of groceries. As Hambleton describes, “the quality of education might be monitored by reporting performance of a national sample of students on the market basket of items each year. Certainly many policymakers seem to desire a single, clear index about the quality of education, much like the Consumer Price Index” (p. 202).

When scaling scores to a common metric or providing scaled scores, the choice of scale is also important. If scaled scores are placed on a 1–100 scale, the interpretation may be that the scaled score is actually a percent correct.
2.2.5 Overall Score Report Weaknesses and Strengths

Aschbacher and Herman (1991) looked at state assessment results used at that time, and formulated specific guidelines for effective reporting of assessment results as follows:

1. Know the audience and the purpose.
2. Keep it simple.
3. Be clear, accurate, comprehensive, and balanced.
4. Use techniques to capture and focus the reader’s attention.
5. Suit format to purpose.

(pp. 5-12).

The suggestions were each accompanied with detailed situational descriptions of how to best implement these guidelines, including an overview of implementing different graphics and some pros and cons of different graphic formats and best application.

Similarly, an important study by Goodman and Hambleton (2004) not only looked at current practices of score reporting and identified a number of weaknesses prevalent in score reporting, but also suggested some important recommendations for future score reporting design. These recommendations include the following:

- Include all information essential for proper interpretation.
- Include detailed information about the assessment and score results in a separate interpretive guide.
- Personalize the student score reports and interpretive guide.
- Include an easy to read narrative summary of the student’s results at the beginning of the student score report.
• Identify some things parents can do to help their child improve.

• Include sample questions in the interpretive guide that illustrate the types of achievements represented by each performance level.

• Include a reproduction of student score reports in the interpretive guide to clearly explain the various elements of the reports.

(pp. 208–210)

Overall, a clear score report is essential to the quality of any testing program. As Allalouf (2007) indicates, the reporting and documenting of scores are key steps included in the multi-step quality control process of scoring, equating, and reporting. Allalouf states “reporting must be done in such a way that people understand the meaning of their scores” with a quality control mistake occurring when “examinees do not understand the meaning of their scores,” and the resulting quality control process may be to “use focus groups of examinees to construct a meaningful explanation of the score report” (p. 41).

2.2.6 Presentation of Data and Statistics in Other Forums

The display of data is not only important in education. In our society, it is hard to get through a single day without hearing some form of statistic or figure. Between news broadcasts conducting impromptu polls of viewers, or USA Today®-style charts and graphs that aim to simply communicate statistical information, statistics are a prevalent part of our society. Partly this is due to the marketing demand for “proving” something to be better, preferred, or true. The same design principles of clarity and understanding that are to be extended to the reporting of assessment results are echoed in other non-education areas of our marketplace.
A current example of this mainstream use of statistics and data display is seen in the Academy Award® winning documentary *An Inconvenient Truth*. The use of statistics, overlaying of information and identification of trends and projections are foundations of the film. By incorporating these aspects in a simple way through vivid line and bar graphs, this complex data and analysis comes through in a clear and comprehensible way. While there is much debate over the message presented, the successful reliance on graphical, visual, and statistical displays cannot be overlooked.

2.3 Teacher Certification Testing

Score reporting in teacher certification testing differs from reporting in the K–12 arena in a few key areas, mainly in how scores are typically reported and used as well as who uses the scores.

2.3.1 The Reporting and Use of Scores

Teacher certification tests are primarily criterion-referenced tests, in that in order to pass, an examinee must demonstrate that they have met the criterion. In the context of setting passing standards, the criterion may be described as that point at which the prospective educator, or candidate, has demonstrated the minimum requisite knowledge and skills in order to perform the job of an educator in that field.

Because of this criterion-based standard that is set, typically scores are reported as either “Pass” or “Not Pass.” If the candidate’s status is “Not Pass,” a scaled score will be provided. While not all state teacher certification testing programs are alike, some states choose not to report a scaled score if an examinee has passed. Reasons for this may
include a concern that passing scaled scores might be used for purposes other than the testing program intent. For example, a district that is looking to hire a new educator may have two job applicants, one with a passing scaled score higher than the other. If they have both passed, should a higher passing score be considered as a “better” hiring quality? In essence, both applicants have met the criterion; therefore, their test performance should be weighed equally. Hence, to preclude the use of scaled scores for purposes other than the demonstration of meeting the state testing standard, examinees may only be informed that they have “Passed” the test. However, as stated in the Standards for Educational and Psychological Testing (1999), “Nevertheless, candidates who fail are likely to profit from information about the areas in which their performance was especially weak” (p. 157). Therefore, while details for passing examinees may not be warranted or appropriate, some performance detail for candidates who do not pass is important.

Score usage in teacher certification testing is primarily for the purposes of demonstrating to a state entity that a specific criterion has been met in order to satisfy requirements of obtaining a teacher license. Prospective job applicants may also use scores to demonstrate that they have “passed the test” and are merely awaiting their final certification approval from the state. Additionally, teachers who are already on the job may wish to add on additional subject matter areas for which they would like to teach. For example, an elementary school educator may wish to add on a certificate or endorsement that covers early childhood years such as preschool age children. The use of scores in this scenario would be to broaden the range of content and age ranges the educator is eligible to instruct.
Audience is a key difference between score reporting in teacher certification testing and public school testing. As opposed to K–12 assessment results, primarily the most interested party in teacher certification score reporting are the examinees themselves (not their parents). Additionally, valuable score reporting information may be shared with teacher preparation institutions in order to provide quantitative information on how their examinees are performing across the various subject matter areas. This institution score report feedback may be valuable in program instruction and curriculum development. If a college or university finds that a large number of examinees are not passing a Secondary Mathematics teacher certification test, the reports may be used to help pinpoint whether all the secondary mathematics content is shown to be problematic, or only certain areas such as Calculus or Trigonometry. States are also interested in these program results, in that decisions regarding approval to offer educator preparation in certain subject matter areas may depend on the performance of examinees from that program.

Another important component of teacher certification score reporting is the level at which scores are reported. Scores may be reported as an overall Pass or Fail, but information could be provided that would help a failing examinee understand what area(s) of the performance were not satisfactory. Typically a teacher certification test will have major areas of content (e.g., test category, content category, objectives). Examinees may be provided with information that illustrates how they performed on each of the major areas of test content, thus allowing the examinee to focus on the areas of need when studying and preparing to retest.
2.3.2 Score Reporting in Current Teacher Certification Testing Programs

Primarily there are two major testing organizations that deliver teacher certification tests in the United States: Educational Testing Service (ETS), and the Evaluation Systems group of Pearson (formerly National Evaluation Systems). Each of these organizations are involved in teacher certification at the national and/or state levels. For context, following are brief summaries of the level of reporting and score report detail available for selected teacher certification programs.

National Programs. Teacher certification on a national level is currently served primarily through The Praxis Series™, an ETS program. A Praxis™ score report indicates the following performance information (Educational Testing Service, 2009):

- the examinee’s scaled score
- the range of possible scaled scores
- the raw points earned and available in each content category (a subset of items on the test)
- average performance range on the test for both scaled scores and raw scores (the range of scores earned by the middle 50 percent of a group of examinees on that test form)

Explanatory information is also provided for Praxis™ score reports, indicating important information about the raw points, and a description of their conversion to scaled scores.

Another national program for teacher certification is the National Board for Professional Teaching Standards (NBPTS) program. As described by NBPTS:

National Board Certification is an advanced teaching credential. It complements, but does not replace, a state’s teacher license. It is valid for 10 years, and renewal candidates must begin the renewal process during their eighth or ninth years as NBCTs.
National Board Certification is achieved upon successful completion of a voluntary assessment program designed to recognize effective and accomplished teachers who meet high standards based on what teachers should know and be able to do. National Board Certification is available nationwide for most preK–12 teachers.

As part of the certification process, candidates complete 10 assessments that are reviewed by trained teachers in their certificate areas. The assessments include four portfolio entries that feature teaching practice and six constructed response exercises that assess content knowledge.

(National Board for Professional Teaching Standards, 2010)

Score reporting for the NBPTS program reflects its complexity, importance, and effort. Given the portfolio basis of this assessment, detailed information about the scoring rubrics and assessor evaluation of evidence are provided to examinees. An NBPTS score report includes the following information by entry or exercise name:

- Raw Exercise Score (RES)
- Weight (E)
- Weighted Scaled Score (RES * W)

For those entries or exercises that do not meet the performance standard for certification, an indication is given so the examinee understands this portion may be retaken. The weighted scaled scores are summed, and a uniform constant is added to the score to produce a total weighted scaled score. The details of this process are explained (with examples) in the online scoring guide for each field (National Board for Professional Teaching Standards, 2009). NBPTS also provides a Retake Worksheet designed to help the examinee in measuring the impact of a score improvement on overall Total Weighted Scale Score. This worksheet is paired with an online retake calculator that is designed to
help the examinee in understanding the score they would need to receive in order to obtain certification.

Additionally there is a relatively new national teacher certification program currently under development by the Evaluation Systems group of Pearson: the *National Evaluation Series™*. While the specific score report information that will be provided to examinees is not publicly available at this time, the program web site indicates that score information will include detailed feedback (Pearson Education, 2009c).

National teacher certification tests pose unique challenges in score reporting in that the goal is for the portability of results from state to state. Therefore, the scales used and reporting information must be the same regardless of where the test is taken, and for what state the results are to be used. Additionally, examinees must be informed of their performance along the entire spectrum of the score reporting scale, in that a passing score in one state may not be passing in another state. Along those lines, the interpretation of performance can at times be very detailed and complex, as illustrated by the matrix provided for *Praxis™* that outlines passing scores by state (Educational Testing Service, 2010).

**State (Custom) Programs.** In contrast to national teacher certification programs, a state certification program is designed to assess content particular to that state’s curriculum, and to provide specific information on whether the examinee has demonstrated the necessary content knowledge and skills for that state’s certification requirements. Therefore, state programs may differ in how scores are reported, what level of reporting and detail is provided, and whether specific passing information is presented.
Following are some examples of large state programs and their score report information and explanatory materials.

Score reports for the California Subject Examinations for Teachers® (CSET®) include “your passing status and, if you did not pass, your total subtest score…the reverse side of your score report contains diagnostic information for each subtest taken to provide you with information about your areas of strength and weakness in each subtest section” (Pearson Education, 2008a). Information about weighting of multiple choice and constructed response sections of the test is also provided.

Score reports for the Massachusetts Tests for Educator Licensure® (MTEL®) include “whether or not you met the qualifying score, your total test score if you did not meet the qualifying score, and your general performance on each subarea or section of the test” (Pearson Education, 2009a). Additionally, detail is provided for subareas of the test, including a range of the number of test items that are included on the test. A check-mark display helps focus examinees to those areas where they did well, and those areas that may need attention. It is explained that while an examinee cannot pass or fail an individual subarea or section, the performance information may help to target the examinee’s understanding of where they got most, many, some, or few of the test questions correct.

For the New York State Teacher Certification Examinations™ (NYSTCE™) program, candidates are provided with scaled scores at the total test level, and at the subarea level, as well as the range of number of questions in each subarea. It is explained to examinees that, “the total test score is based on the total number of test questions answered correctly…each multiple-choice question counts the same toward the total
score. There is no penalty for wrong answers” (Pearson Education, 2009b). The explanatory materials provided with the score report (Pearson Education, 2007) include a snapshot of a sample score report that shows an examinee how to read and use the detail that is provided by subarea. It points out the highest and lowest performing subareas for the examinee, and directs the examinee to the preparation guide and other resources available for the testing program. The explanation included aims to ensure the clarity of interpretation for the scaled score and subarea performance reporting by indicating that “your total test score is not the average of your subarea performance results. Because subareas of the test may contain different numbers of questions, it is not possible to average your performance results across subareas to arrive at your total test score” (Pearson Education, 2007). As illustrated later, this concept is important in that this can be a potential misunderstanding of test scores on the examinee’s part.

For the Michigan Test for Teacher Certification (MTTC), examinees are provided with passing information for the total test, the range of questions presented on the test by subarea, and for each subarea a performance index. The performance index provides “an indication of performance in each subarea of the test…this information will help you understand your areas of strength and/or weakness; you do not “pass” individual subareas” (Pearson Education, 2008b). The performance index uses “+” to indicate the degree of strength in each subarea, ranging from few questions correct, to most questions correct. A sample test score report is provided in the explanatory materials, illustrating the different information reported (total test scaled score, minimum passing score needed, performance index) and a narrative example on where this sample examinee should probably focus their studies prior to retesting.
Summary. While it is clear that there is no one “gold standard” of score reporting, it seems apparent that the level of detail reported and information provided can be influenced by the test structure, test length, state policies, and the nature of the testing program and population. For some low-incidence tests (e.g., Latin, Russian) it may not be advisable to report detail of the examinee population (e.g., median performance). Additionally, a test that is divided into a subtest structure, may already be finite enough that further reporting by subarea would result in providing unreliable information because of the limited number of test questions.

Preparation Materials. Regardless of the level of reporting from state to state, or nationally, it should not be overlooked that preparation materials may also include helpful information to assist an examinee in interpreting their score or understanding the content being assessed. As with the variance of teacher certification programs offered, states vary in the amount of preparation materials or resources available for examinees and faculty members. Regardless, there are numerous resources available online or in print to assist with examinee preparation for these teacher certification examinations. Whether offered through the testing organization, third-party vendors, or other educational entities, these resources may not only assist with content preparation, but also with understanding how scores are calculated and where deficiencies may exist for candidates who do not pass an examination. These resources may include expanded study guides, practice tests (paper-based or online/interactive), links to other state or national resources (student standards, national standards), or detailed test content information such as “tests at a glance”, test profiles, and the test framework. At a
minimum, all teacher certification programs provide detailed information on the content eligible to be tested for a particular certificate.

2.4 Literature Review and Connection to Score Report Design

Given the literature review findings, design suggestions, contextual information needed, and current material in the field, it was important to apply these ideas when designing the sample score reports, and use the interviews and focus groups as a way to dig deeper and obtain qualitative feedback with regard to specific choices made on the score reports. By doing this, the goals of this study were realized in that more than just preferences were obtained from the participants.
CHAPTER 3

METHODOLOGY

3.1 Introduction

The methodology to this study involved four key steps:

- the design of the sample teacher certification testing score reports and explanatory materials;
- the conducting of interviews and focus groups to review and provide feedback on the sample score reports;
- the processing and review of interview and focus group feedback with a resulting application toward an improved teacher certification score report, and lastly
- the conducting of a student focus group to review the resulting model score report and offer student feedback

3.2 Research Questions

There are a number of research questions that were asked during the course of this study. Primarily, the questions attempted to gather information on score reporting strengths, weaknesses, and comprehension, as follows:

- Are there commonly misunderstood components of score reporting?
- What are ways in which data can best be presented in order to inform instruction for teacher certification candidates who have not passed?
- Are there commonly desired elements of score reporting that are not being realized?
• Is more information on a score report helpful, or is it only more confusing for those who need to interpret it?
• Should feedback provided on score reports be dependent on audience?
• Given feedback, what are some fundamental elements of score reporting that are effective? What are some fundamental elements that are not effective?

3.3 Sample Score Report Design

The first key step in this study involved the design of the sample score reports. Information was gathered from the field and literature on what is currently being reported on score reports, teacher certification reports in particular. This information from sample score reports, literature, score report interpretation guides, and various public sources helped to frame what is currently reported and to lend some insight to the questions that arose regarding commonly used terminology, uncommon terms, as well as information desired by stakeholders.

Following this review of the literature, three sample teacher certification test score reports were designed for an area common to educator preparation, Fundamental Skills: Mathematics. The score reports presented hypothetical information of a not passing examinee performance on basic mathematics subject matter knowledge, as divided among four areas of subcontent, or learning objectives. The score reports incorporated aspects of the research findings, and each were a sample score report for the same examinee’s information. The sample score reports were not affiliated with any actual teacher certification testing program, nor linked to any actual examinee score, and all
participants and focus groups saw the same sample score reports. Internally and externally, the reports were identified using neutral identifiers of A, B, C, and D.

The concept of using multiple sample score reports helped to provide comparative information from methodology to methodology. For example, one score report makes no use of color, whereas another report incorporates some design elements that makes use of graphics and color. Two of the reports contained contextual performance information in relation to other passing examinees, while the third did not.

Given that the design of the score reports themselves are in essence “results” of this study, their creation, questions developed to assess them, and corresponding explanatory materials are outlined and presented in Chapter Four.

3.4 Interviews and Focus Groups - Research Design

There are a number of components to this research design, primarily involving the qualifications of participants, the conducting of interviews and focus group sessions, documenting of dialogue and discussion, and collection and processing of all feedback. Lastly an application of the feedback was made through suggested potential revisions and improvements that could be made to the score reports, and applying those to designing a final fourth score report.

In preparation for the focus groups and interviews, a form outlining Consent for Voluntary Participation was designed and Human Subjects Review Board approval was obtained. The approved consent form can be found in Appendix A, and a signed consent form was obtained for each participant in the study.
3.4.1 Participants

Focus group and interview participants included educators, educational testing organization personnel, and doctoral students as follows:

- 16 educators (public school or college/university) currently or recently practicing across various subject matter areas throughout the United States
- 6 assessment staff involved in designing, interpreting, or advising on score reporting and scoring procedures
- 10 university doctoral education students, knowledgeable in measurement concepts and psychometrics

Each participant was asked to respond to basic background questions, and to provide informed content.

The participants and groups involved in a way represent a continuum of information. First, educator participants were selected in order to hear from educators currently working with potential future educators, or knowledgeable of assessment results and their interpretation. The justification being that given the purpose of teacher certification and the knowledge of these educators, some would have extensive experience in using assessment results in their day to day work in the academic environment, while others may have limited but still a basis for some use of assessment results. Some educators had been involved in education for over 15 years, while others had just recently taken and passed their certification examination.

The assessment staff were an integral component of the feedback process in that it would help to display whether comments shared by educators were similar to comments shared by assessment staff, or if a different lens might be applied when reviewing the
score reports and providing feedback. All of these assessment staff had experience in applying measurement concepts, designing program material, or explaining test results or concepts to examinees and other stakeholders.

Lastly, the students seemed a logical ending point, given that education students are the ones receiving individual score reports for the teacher certification program. To be able to walk through the entire process with the students, and then present score report information designed based on informed feedback (Score Report D) helped to take the study one step further towards an application of feedback obtained through the interview and focus group process.

### 3.4.2 Interview/Focus Group Feedback

Using the sample score reports, focus group participants were asked to respond to specific questions that elicited information regarding their level of understanding and interpretation of each sample score report. Questions types included the following:

- participant demographic questions (e.g., How many years of experience do you have as an educator?)
- identification questions (e.g., Did the examinee portrayed in this score report Pass the exam?)
- calculation questions (e.g., In learning objective 001, what number of items did the examinee get correct?)
- comprehension questions (e.g., Do you understand what is meant by the term “confidence band”?)
• interpretation and application questions (e.g., Based on the information provided, what advice would you give to this examinee if they were preparing to retest?)

• preference questions (e.g., Would the use of raw scores or an N correct be more understandable/meaningful for you? Did the use of color in the performance chart enhance or hinder your interpretation of the examinee’s performance?).

After gathering this individual feedback at the beginning of introducing each score report, a dialogue took place with the participants in order to discuss what aspects of the reports were useful, least useful, misunderstood, and to elicit a comparative viewpoint of one score report against another. All report interviews and focus groups were recorded (audio and Web presentations) in order to allow a revisiting of the feedback upon reflection.

Although it may have been ideal to conduct each educator interview in person, this was not physically possible. Instead, each educator interview was conducted live via the internet and telephone. Given the widespread use of web-based meetings in the business world, and distance/web-enhanced learning in education, it seemed only appropriate to make use of this medium and involve participants from across the nation, from different universities and public schools.

3.4.3 Data Collection

The data collection step in the study occurred as interviews and focus groups were conducted. Responses to specific questions were captured on questionnaires particular to
each score report. Discussion was also held on each score report with questions and answers recorded and noted. All questionnaires and interview questions are found in Chapter 4. Quantitative responses such as those to pre-designated questions regarding interpretation or understanding of the sample reports were entered electronically into a table, including the key-entry of any open responses. The qualitative discussion and outcomes of the focus groups were summarized, and emerging trends and comparisons are highlighted and presented.

3.4.4 Data Analysis

A processing and qualitative analysis of the feedback and responses received occurred after the educator interviews and testing professionals focus group. This summary presents primarily qualitative information on reactions to the model score reports. The information obtained was reviewed for response differences and findings of interest. Through the discussion and interviews, feedback was asked of participants of a comparison of the score reports and feedback was obtained regarding likes and dislikes across all three reports.

All findings and results in this study were reported anonymously and, where possible, in the aggregate as group data. Some comparison of feedback as based on demographic question responses or commentary was presented. For example, a comparison was made of comments by educators working with elementary students against those educators working with upper-level grades for a specific component of Score Report C.
All open responses were reviewed for trends (e.g., “I did not understand the concept of imprecision,” “Confidence bands are difficult to interpret”) in order to draw comparisons in interpretation or qualitative review. Additionally, some feedback on the process itself (use of Web-based meetings, presentation of questions and reports) was recorded and noted.

3.4.5 Application of Results and Findings

The final step in the methodology involved the design of a fourth Score Report (Score Report D). By synthesizing all information obtained through the interviews and focus group feedback, the fourth sample score report was a qualitative-informed design effort, for which doctoral students provided further feedback and direction.
4.1 Introduction to the Results

The results obtained in this study are numerous given the feedback provided by participants. Given the progression of activities involved in obtaining the results, the materials presented in this chapter follow the progression of the study itself.

First, the design of Score Reports A, B, C and their corresponding explanatory information pages is presented. Next, the design of the questionnaires for which respondents would provide their answers is described. Following the design of the score reports and questionnaires, the educator interviews and the feedback obtained through this population are presented ordered by Score Report. After this, the focus group with educational testing professionals is presented, along with the feedback obtained through this population ordered by Score Report. Using the feedback obtained from the educator interviews and the educational testing professionals group, Score Report D was designed. The description of this design process and resulting score report is presented next. Lastly, the focus group with University Doctoral students is described, along with the feedback obtained from this population ordered by score report, including the feedback obtained on Score Report D.

4.2 Designing the Score Reports and Explanatory Information

The fundamental cornerstone of the study was the design of the score reports. It was important to consider different ways of presenting the same information. After
considering whether the data should vary from report to report, the decision was made to keep the data standard across the reports. That way the same data would be presented for the examinee but in different ways. In some cases the examinee data would include confidence intervals, in others it would not but would include comparative information. In essence, the same performance would be presented, but communicated differently through different statistics used. After deciding this, there was a question too as to whether the examinee should be not passing or passing. Based on the stated purpose of this study, it was important for this examinee to be not passing.

The score report design began with determining the teacher certification test (subject matter) the score reports would represent. Given that basic skills are fairly universally known and understood by educators and education students, the sample test used in this study was modeled on this subject matter. Typically, basic skills comprise Reading, Writing, Mathematics, and at times, Technology. The test used in the score reports was named “Fundamental Skills,” so as not to give an impression that this was part of an actual testing program. The concentration of subject matter within these skills was decided to be Mathematics, as it would be segmented and distinct to allow for discussion across the different learning objectives from the participants. It was not necessary for the field to be Mathematics, it could just have easily been Reading, Writing, or a specific subject matter area such as science, but Mathematics lent itself well to distinct learning objectives and categories of content.

In defining the specific learning objectives that would be reported within the score reports, a number of current basic skills teacher certification test frameworks were consulted (e.g., Praxis I: Pre-Professional Skills Test, Georgia’s GACE program,
California’s CBEST program) and four main learning objectives were determined to encompass the knowledge and skills that would need to be demonstrated within this fundamental skills mathematics test framework. The four learning objectives crafted were 001 Number Properties and Operations; 002 Computation and Problem Solving; 003 Statistical Concepts, Data Analysis and Interpretation; and 004 Measurement Concepts and Principles of Geometry.

The test design was created next. First, an experimental design of an equally weighted test was explored, with 12 items in each learning objective. While this design had merit in that it would allow for an equal comparison of performance across learning objectives, equally comparable learning objectives may have limited the study and the questions that could be asked of participants. Because unequal weights across reporting categories is common in score reporting, this study adopted that model in order to explore whether this would influence feedback received, and whether that would be a challenge for interpretation. By having a different number of questions (or score points) in each learning objective, participants were able to be asked about relative weight and a better assessment could be made as to whether it was understood by participants that each learning objective did not have the same number of questions.

With the test design determined, and an unequal weight of each learning objective in relation to the other, a determination of the total number of items was made. Given the goal of reporting detailed information by learning objective, a determination was made to have at least 40 items on the sample test in order to have learning objectives vary in the number of items each would have on the test, resulting in a range of 8 to 14 questions per learning objective. These quantities were based on the need to have enough items to
assess knowledge and skills in these mathematical areas, and enough items for the learning-objective level reporting. For example, the Praxis™ does not report average performance if there are fewer than 8 items in a topic section (Educational Testing Service, 2009).

Learning Objective 001 was given the greatest weight, slightly less weight was given to Learning Objective 002, the least weight to Learning Objective 003 and the remainder went to Learning Objective 004. The result was a 45-question test with four learning objectives. The test was then divided among the learning objectives as follows:

- Learning Objective 001: 14 questions, 31% of the test
- Learning Objective 002: 12 questions, 27% of the test
- Learning Objective 003: 8 questions, 18% of the test
- Learning Objective 004: 11 questions, 24% of the test

After determining the test design, the examinee’s individual performance was created. With four learning objectives, and a desire to ask direct comprehension questions of the participants, the examinee performance varied from having very good performance on two learning objectives, poor performance on one objective, and mediocre performance on the remaining objective.

In addition to designing each score report, it was important to include an explanatory page for each score report. Most teacher certification testing programs either include this information with the score report, or make it available online. Some examples include “Interpreting Your Praxis™ Examinee Score Report” (Educational Testing Service, 2009), or “How to Read Your Score Report” for the New York State Teacher Certification Examinations (Pearson Education, 2007). These materials have a
dual purpose: to provide an explanation of the components included on the score report and to present other important program information in relation to test scores. In designing the “Understanding Your Score Report” pages, a consideration was made for what was presented data-wise on the score report and what an examinee who did not pass may need to know (e.g., retesting information).

4.2.1 Score Report A and Explanatory Information

In designing Score Report A, consideration was given to simplicity of presentation and information. At the most basic level, the administration date and examinee ID were placed at the top of the score report, followed by the examinee mailing information located in a traditional “mailer” format that would allow for the address to show through a windowed envelope if the report were to be mailed.

The test name is featured prominently in the middle center, followed by score information. Because the certification exam reports scores as scaled, “Your Score” is listed first, followed by a statement indicating what minimum scaled score would be necessary to pass. Then listed directly below “Your Score” is the examinee’s passing status.

As mentioned earlier, a goal was to provide a detailed performance explanation at the learning objective level, but have varied types of data presented across score reports A, B, and C. For Score Report A, the following pieces of information were included to describe the learning objective performance:

- number of questions for each learning objective
- examinee’s % correct for each learning objective
• learning objective number and name

• total number of questions on the test

After presenting the information above in a chart format, a bar graph was developed that visually showed the examinee’s percent correct out of 100% for each learning objective. Next to that bar graph, a Test Design cylinder chart was presented. The purpose of the cylinder chart was to visually display how the test was divided up by objective. The percent of test for each objective was listed in the corresponding cylinder piece, with the cylinder ordered in ascending objectives 1 through 4. Visually, the cylinder chart was designed to show how the entire test is distributed across objectives, and therefore a percent of test designation was present on the left-hand side of the cylinder. Figure 4.1 shows the final Examinee Score Report A used in the interviews and focus groups.
Figure 4.1 Examinee Score Report A

In designing the “Understanding Your Examinee Score Report” material for Score Report A, each performance reporting component of Score Report A was described as detailed below:
• defined what “Your Score” means in terms of being scaled on a range of 100 to 300.
• defined what a passing score would be.
• broke apart the information presented in the performance chart and defined each piece (N questions, % Correct, Learning Objectives).
• defined what the bar chart and the test design were designed to show.
• included information that the examinee may need regarding retaking the test, since they had not passed.

Figure 4.2 shows the final Understanding Your Examinee Score Report for Score Report A.
EXAMINEE SCORE REPORT A

UNDERSTANDING YOUR EXAMINEE SCORE REPORT

Overview. This score report provides your test results for the teacher certification test that you took on the test administration date indicated on your report.

Your Score. You will find your total test score for the test listed. This score is based on the number of items you answered correctly and is converted to a scale that ranges from 100 to 300.

Passing Score. A score of 200 is the passing score for the exam. Obtain a score of 200 on an exam, or higher, and you will pass. A score lower than 200 is designated as “Not Passed.”

Performance Information. On your score report, you will find information regarding your performance on each Learning Objective on the test. This information can help you identify your relative strengths and weaknesses.

N Questions. This is the number of questions on the test for each learning objective.

% Correct. This is the percent of questions you answered correctly for each learning objective.

Learning Objectives. These are the learning objectives for the Fundamental Skills: Mathematics test. For more information on these objectives, please visit: http://www.teacherscertificationdissertation.com

Your Performance. This chart shows you the percent of questions you answered correctly for each learning objective.

Test Design. This chart shows you the % of questions from each objective that make up the entire test.

Retaking the Test. You may retake the test at any future test administration by following the same registration procedures you completed for previous test administrations.

Please keep in mind that minor fluctuations in scores are expected over repeated testing. While large differences in percent correct across Learning Objectives may indicate relative strengths and weaknesses, small differences in performance may be due to these minor fluctuations in performance that are typical of multiple test administrations.

Figure 4.2 Understanding Your Examinee Score Report for Score Report A
4.2.2 Score Report B and Explanatory Information

In designing Score Report B, a visual departure from Score Report A was made, as well as providing some additional performance-related information, and not including some performance information that was present on Score Report A. Examinee information was placed at the bottom, resulting in the examinee seeing their performance first when reading from the top down. The test name was still prominently placed at the top center; however, changes were made in how the scaled score and passing status information were presented in Score Report B—the passing status was placed top left, followed by the minimum passing scaled score, and the “Your Score” designation. All were placed on the same visual line. A visual organization change was made in the chart outlining the examinee’s performance by placing the learning objective name in the first column, rather than the performance/statistical information first. Another visual change was the addition of color in the “Your Performance” bar graph. The color tied into the dashboard technique of using red, yellow, and green as indicators of concern, caution, and clearance. A key or legend was also developed to explain the color indicators and how they were defined. In defining this key for the bar graph, performances of less than 50% correct were marked by using a red bar, performances between 50 and 70% correct were marked by using a yellow bar, and examinee performance above 70% correct received a green bar. These thresholds were chosen as arbitrary cutpoints for the three levels of attention mentioned. The visual enhancements and additions in Score Report B were based on the design principles echoed by Hambleton (2002), Brown (2001), which called for the use of graphics and their ability to stand alone, while also connecting that information to meaningful numbers (e.g., the median reference group).
Statistically, additional performance information was provided on Score Report B. While no number of questions was provided as was present on Score Report A, the examinee’s percent correct by objective was presented. In order to provide a different level of performance statistic, and a contextual frame of reference, the median percent correct of passing examinees was added. In addition to providing the percent in the table, the median percent was also represented in the bar graph by an “X”. The median statistic was chosen over the mean in that the median would be less susceptible to large variations due to outliers in performance. By supplying this additional frame of reference for the examinee in this hypothetical score report, the goal was that an examinee’s performance could be interpreted within a larger context of understanding how passing examinees performed, beyond just a scaled score of 200. This would enhance the discussion with educators and educator testing professionals, and open up discussion on whether statistics of other passers are helpful. In reporting this statistic, the entire passing cohort was represented, and no differentiations were made between just barely passing, or just above passing. While it was understood that the statistic included everyone with a total scaled score between 200 and 300, it was clear that breaking this statistic out into separate passing groups would likely not be possible with some teacher certification tests that have few examinees. By including all examinees who passed, the largest number of examinees possible were represented in this performance statistic. No test design information was provided on this score report. Figure 4.3 shows the final Examinee Score Report B used in the interviews and focus groups.
Figure 4.3 Examinee Score Report B

The “Understanding Your Examinee Score Report” material for Score Report B was designed to describe each performance reporting component of the Score Report B. Any component in Score Report B that was also in A had the same description. Additional descriptions were necessary to provide for “Median % Correct of Passing
Examinees” and for the key that explained the color indicators on the bar chart. In defining the median percent correct of passing examinees on the explanatory page, it was emphasized that the statistic was the middle value of percent correct for passing examinees. There was concern that median might be misinterpreted for mean or average, therefore the statistic was reinforced in the explanatory report. Figure 4.4 shows the final “Understanding Your Examinee Score Report” for Score Report B.

![Image](https://example.com/image.png)

**Figure 4.4 Understanding Your Examinee Score Report for Score Report B**
4.2.3 Score Report C and Explanatory Information

Score Report C was also designed to be visually and statistically different than Score Reports A and B. Visually, in Score Report C, examinee information was placed back at the top of the report, including administration date and examinee ID, as well as the mailing address. Summary performance information was visually reorganized by placing “Your Score” and “Minimum Passing Scaled Score” above the performance chart and listing the passing status below the performance chart. Within the performance table, no changes were made between Score Reports A and C. The table included the number of questions for each learning objective, the examinee’s percent correct for each learning objective, the learning objective number and name, and the total number of questions on the test. Visually, the percent correct by learning objective was presented in a vertical fashion and used a graphic rather than a bar. A pencil icon was chosen to serve as a visual means to fill in the graph up to the performance level. Partial pencils had to be used to round out the upper end of the chart. While it was understood the visual may not have been “precise” in its presentation, the idea was to create a chart that might be more visually pleasing and “approachable” to audiences who may be intimidated by charts and graphs. The use of the pencil icon was an example of using graphics to make score reports more appealing to educators.

In contrast to this accessible “pencil” chart, an additional chart was placed on the score report to reflect new statistical information. A chart showing “Performance Confidence and Median Performance of Passing Examinees” was added. The chart shows vertically for each objective the relative 90% confidence interval, where the examinee’s score is the middle point, and the upper and lower thresholds of the band
represent where the examinee’s score may lie if testing again without any additional instruction. The confidence interval width varied by learning objective. Additionally, an “X” was placed in the chart to indicate where the median percent correct of passing examinees was. While this information was not provided in the table up above with the examinee’s performance, the same statistic from Score Report B was represented in C, but in the chart only. Figure 4.5 shows the final Examinee Score Report C used in the interviews and focus groups.
In designing the “Understanding Your Examinee Score Report” material for Score Report C, each performance reporting component of Score Report C was described. Any component in Score Reports A or B that was also presented in C had the same description. The only additional descriptions provided for Score Report C were for
“Performance Confidence” and “90% Confidence Interval.” In describing the performance confidence, an attempt was made to explain as simply as possible why there is a performance confidence, what the band means, what the width of the band means, and what is meant by “90%” confidence interval. Figure 4.6 shows the final “Understanding Your Examinee Score Report” for Score Report C.
EXAMINEE SCORE REPORT C

UNDERSTANDING YOUR EXAMINEE SCORE REPORT

Overview. This score report provides your test results for the teacher certification test that you took on the test administration date indicated on your report.

Your Score. You will find your total test score for the test listed. This score is based on the number of items you answered correctly and is converted to a scale that ranges from 100 to 300.

Passing Score. A score of 200 is the passing score for the exams. Obtain a score of 200 on an exam, or higher, and you will pass. A score lower than 200 is designated as “Not Passed.”

Performance Information. On your score report, you will find information regarding your performance on each Learning Objective on the test. This information can help you identify your relative strengths and weaknesses.

   N Questions. This is the number of questions on the test for each learning objective.
   Your % Correct. This is the percent of questions you answered correctly for each learning objective.
   Learning Objectives. These are the learning objectives for the Fundamental Skills: Mathematics test. For more information on these objectives, please visit http://www.teacherscertificationdissertation.com.

Your Performance. This chart shows you the percent of questions you answered correctly for each learning objective.

Performance Confidence and Median Performance of Passing Examinees. This chart shows you the performance confidence (90% confidence interval) for performance on each objective.

Performance Confidence. All tests may have some source of error, resulting in an observed score that may change slightly with retesting. The confidence band provides an interval over which the examinee would score if he or she were to retake the examination without any additional instruction. Reliability describes the consistency of the measurement obtained. The wider the band, the lower the reliability of scores, and the greater possible variation in score may be observed.

90% Confidence Interval. This means that if you tested again and computed the confidence intervals each time, the confidence intervals you get will include your score 90% of the time. In that sense, there is a 90% chance that any specific confidence interval actually contains the observed score.

Median % Correct of Passing Examinees. This is the median (middle) percent of questions answered correctly by passing examinees, for each learning objective. If the X lies within your performance confidence, then your score may be expected to meet the median performance of passing examinees.

Retesting the Test. You may retake the test at any future test administration by following the same registration procedures you completed for previous test administrations.

Please keep in mind that minor fluctuations in scores are expected over repeated testing. While large differences in percent correct across Learning Objectives may indicate relative strengths and weaknesses, small differences in performance may be due to these minor fluctuations in performance that are typical of multiple test administrations.

Figure 4.6 Understanding Your Examinee Score Report for Score Report C
4.3 Designing The Individual Questionnaires

Following the creation of the three score reports and their explanatory pages, individual questionnaires were designed for each report. The questionnaires were each designed to assess participant’s comprehension, identification and opinion. The goal was to have participants individually respond to the questions, and then to facilitate a discussion of the report with the participant(s). That way, participants were focusing on the materials at hand in order to comprehend the report, rather than clarifying the information as part of a discussion and then answering.

For Score Reports A, B and C some common comprehension questions were asked, such as the date of the test administration, titles of objectives, the passing status and the strong/weak learning objectives. Common opinion questions were also asked, such as whether in general the participant understood the performance information communicated in the report and what suggestions the participant might make to enhance the performance information communicated in the report.

For Score Report A, specific questions were crafted to ask about the cylinder chart, and the objective weights. For Score Report B, specific questions were crafted to ask about identifying the examinee’s performance in relation to the median percent correct of passing examinees. For Score Report C, specific questions were crafted to ask about the confidence bands, their meaning, and the examinee’s performance in relation to the median percent correct of passing examinees. Figures 4.7, 4.8 and 4.9 show the final individual questionnaires for Score Reports A, B and C respectively.
SCORE REPORT A

Individual Activity

For the first part of the focus group, specific questions will be asked regarding Score Report A. Prior to any group discussion, please take some time to answer the following specific content-related questions. If you are unsure of an answer, please leave the response blank. We will review the answers and discuss the reports specifics after everyone has finished.

Score Report A

1. The report reflects data from what administration date?

2. How many examinees are represented on this score report?

3. Did this examinee pass the examination?

4. What is the title of learning objective 002?

5. In learning objective 004, how many total items did the examinee see?

6. Does each learning objective have the same number of questions?

7. From the information presented, can you tell if one learning objective counts more or less than another?

8. In learning objective 001, what number of items did the examinee get correct?

9. In what learning objective did the examinee perform best?

10. In what learning objective did the examinee perform worst?

11. What do you think the cylinder chart represents?

12. In general, did you understand the information communicated in Score Report A?

13. What suggestions would you make to better enhance the performance information communicated in Score Report A?
SCORE REPORT B

Individual Activity

For this part of the focus group, specific questions will be asked regarding Score Report B. Prior to any group discussion, please take some time to answer the following specific content-related questions. If you are unsure of an answer, please leave the response blank.
We will review the answers and discuss the reports specifics after everyone has finished.

Score Report B

1. The report reflects data from what administration date?

2. How many examinees are represented on this score report?

3. Did this examinee pass the examination?

4. From the information presented, can you identify the total number of items assessing learning objective 004?

5. For each learning objective, indicate whether the examinee is performing below, at, or above the median performance of passing examinees, by circling the appropriate descriptor:

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Examinee Performance in Relation to Median % Correct of Passing Examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Below</td>
</tr>
<tr>
<td>002</td>
<td>Below</td>
</tr>
<tr>
<td>003</td>
<td>Below</td>
</tr>
<tr>
<td>004</td>
<td>Below</td>
</tr>
</tbody>
</table>

6. From the information presented, can you tell what percentage of the total number of items the examinee gets correct?

7. In what learning objective did the examinee perform best?

8. In what learning objective did the examinee perform worst?

9. In general, did you understand the information communicated in Score Report B?

10. What suggestions would you make to better enhance the performance information communicated in Score Report B?

Figure 4.8 Individual Questionnaire: Examinee Score Report B
SCORE REPORT C

Individual Activity

For this part of the focus group, specific questions will be asked regarding Score Report C. Prior to any group discussion, please take some time to answer the following specific content-related questions. If you are unsure of an answer, please leave the response blank. We will review the answers and discuss the reports specifics after everyone has finished.

Score Report C

1. The report reflects data from what administration date?

2. How many examinees are represented on this score report?

3. Did this examinee pass the examination?

4. From the information presented, can you tell if one learning objective is weighted more or less than another?

5. From the information presented, can you tell what percentage of items on the total test the examinee gets correct?

6. In what learning objective might you advise this examinee to study before retesting?

7. In what learning objective did the examinee perform best?

8. In what learning objective is the performance confidence band the widest?

9. Why is one confidence band wider than another?

10. For each learning objective, indicate whether the examinee is performing below, at, or above the median performance of passing examinees, by circling the appropriate descriptor:

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Performance in relation to Median % Correct of Passing Examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Below At Above</td>
</tr>
<tr>
<td>002</td>
<td>Below At Above</td>
</tr>
<tr>
<td>003</td>
<td>Below At Above</td>
</tr>
<tr>
<td>004</td>
<td>Below At Above</td>
</tr>
</tbody>
</table>

11. In general, did you understand the performance information communicated in Score Report C?

12. What suggestions would you make to better enhance the performance information communicated in Score Report C?
4.4 Educator Interviews

As mentioned previously, following the design of the reports and explanatory materials, obtaining educator feedback and information on the reports was a main goal of the study. Educator interviews were designed to solicit information and produce a dialogue that might better reveal what “worked” and what “didn’t work” with each report, and how various pieces of the report were received.

A total of 16 educator interviews were conducted. Educators volunteered to participate and were current or recent public school, faculty, or district-based educators. Interviews were conducted individually with each educator and took approximately 60 to 90 minutes each.

4.4.1 Overview

Educator interviews were originally conceived of as a large-group format, with multiple participants at one time providing feedback and having a dialogue about the score report. After some failed attempts to gather multiple educators at once for the discussion, the focus shifted to individual interviews. One benefit of this model over a group setting was there was an increased sense of participation from each interviewee, in that directly responses to every question were obtained from every participant. If this study had relied on a group setting, on the other hand, there may have been some vocal participants, but likely there would have been some participants who would not speak up at all on some of the questions. As each interview date and time was confirmed, written consent for voluntary participation was obtained. As each interview was scheduled, a confirmation e-mail with connection information for WebEx (a Web-based meeting
service) was sent to the participant. Using WebEx to conduct the educator interviews allowed for each participant to see the score report materials, respond to the individual questions and discuss the reports via telephone.

Each educator interview began with the participant seeing a “welcome” on the WebEx screen, thanking them for participating in the study and providing them with the details for phoning into the teleconference. After connecting to the teleconference, introductions were made, and the purpose of the study was described to the participant. Participants were asked if there were any questions prior to beginning.

Next, each participant was asked to respond to six background questions, in order to describe the participants in the study. Participants were reminded that their individual responses or comments would not be used anywhere with their name. Using the WebEx system, the background questions were made available for the participant to respond to using a polling feature. The polling feature allows you to create surveys or poll questions prior to the WebEx session and then open the questions up to participants during the meeting to “poll the audience.” Polling was used as a way for participants to respond to the specific background questions and score report questionnaires. After participants finished responding to a poll, they submit their answers, and their replies were transmitted instantly for data capture. The responses were saved as “Poll Results.”

After receiving the poll results for background questions, Score Report A was displayed for the participant along with the corresponding Understanding Your Examinee Score Report. Participants could page between the score report and the explanatory page, however they could not see Score Reports B or C. After ensuring that the participant was able to see the score report and explanatory page, the poll questions for Score Report A
(Score Report A Questionnaire) were released to the participants. Each participant typed in their responses to the questions and submitted them when finished. Participants were given 10 minutes to complete this poll; however, they could submit their responses before the 10 minutes expired. Following the receipt of the poll responses, the score report was discussed with the participant. At that time, no changes could be made to the responses submitted. The questions asked during this process of discussion are summarized in section 4.4.3 of this chapter. At the end of the discussion, each participant was asked if they had any further comments to share on the score report before moving to the next one.

The process described above for Score Report A was then repeated for Score Reports B and C. During each score report questionnaire, participants were not able to page between score reports, only between the specific report in question and the corresponding score report questionnaire, and during each questionnaire participants could not modify their responses once submitted.

After discussing Score Report C, participants were asked to compare all three score reports, and were given the electronic permissions to page between each report in order to do so. Participants were asked questions regarding preference, clarity, raw scores, and methods for presenting data or describing imprecision. While no poll was used to capture these responses, the interview allowed for opinions, preferences and final thoughts to be shared. At the end of this discussion, each participant was asked if he or she had any further comments to share on any of the score reports before concluding the interview. Participants were then thanked and they disconnected from the meeting. For
each interview, all poll results were saved and each meeting was recorded (audio and Web-video of documents presented) with the proceedings saved electronically.

4.4.2 Participant Demographics

Participants were first asked to respond to background questions. Table 4.1 shows the demographics of the educator participants in this study.

Table 4.1. Demographics of Educator Participants (Interviews)

<table>
<thead>
<tr>
<th>1. Gender</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>13</td>
<td>81%</td>
</tr>
<tr>
<td>Male</td>
<td>3</td>
<td>19%</td>
</tr>
<tr>
<td>No Answer</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Ethnicity</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian/Pacific Islander</td>
<td>2</td>
<td>13%</td>
</tr>
<tr>
<td>Black, not of Hispanic origin</td>
<td>1</td>
<td>6%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Native American</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>White, not of Hispanic origin</td>
<td>12</td>
<td>75%</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>6%</td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Years of experience working in education</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–3 years</td>
<td>1</td>
<td>6%</td>
</tr>
<tr>
<td>4–10 years</td>
<td>3</td>
<td>19%</td>
</tr>
<tr>
<td>11–15 years</td>
<td>3</td>
<td>19%</td>
</tr>
<tr>
<td>over 15 years</td>
<td>9</td>
<td>56%</td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>100%</td>
</tr>
</tbody>
</table>
Table 4.1. (continued)

<table>
<thead>
<tr>
<th>4. Job Description</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public School Educator</td>
<td>7</td>
<td>44%</td>
</tr>
<tr>
<td>District Superintendent/Administrator</td>
<td>1</td>
<td>6%</td>
</tr>
<tr>
<td>College/University Faculty</td>
<td>6</td>
<td>37%</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>13%</td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>100%</td>
</tr>
</tbody>
</table>

5. From question #4, please specify the field or content area in which you work (e.g., Mathematics).

Responses:
- Band
- Early Childhood
- Elementary Education
- Foundations of Education
- Music
- Music
- Physical Education, K–12 Pedagogy
- School Guidance
- School Psychologist (background in special education)
- Music Education
- Second Grade Teacher—All Subjects
- Science
- Science Education
- Special Education
- Special Education

15 responses, 1 no response

6. How would you categorize your level of experience with educational tests and statistics?

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Minimal experience</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Moderate experience and use throughout the academic year</td>
<td>8</td>
<td>50%</td>
</tr>
<tr>
<td>Ongoing experience and use throughout the academic year</td>
<td>8</td>
<td>50%</td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>100%</td>
</tr>
</tbody>
</table>

4.4.3 Score Report A Feedback

In looking at the individual educator interview responses to Score Report A, all participants were able to correctly respond to the comprehension/identification questions listed below along with the correct response:
• The report reflects data from what administration date? (February 1, 2010)

• How many examinees are represented on this score report? (One)

• Did this examinee pass the examination? (No)

• What is the title of Learning Objective 002? (Computation and Problem Solving)

• In Learning Objective 004, how many total items did the examinee see? (11)

• Does each learning objective have the same number of questions? (No)

• In what learning objective did the examinee perform best? (Learning Objective 001: Number Properties and Operations)

• In what learning objective did the examinee perform worst? (Learning Objective 004: Measurement Concepts and Principles of Geometry)

When asked, “From the information presented, can you tell if one learning objective counts more or less than another,” 11 participants indicated yes they could tell. However, other participants indicated no, with one participant indicating that while he/she could see there was a different number of questions presented for each learning objective, he/she could only assume that each question is weighted equally, and thus could not confidently state whether one learning objective counts more or less than another.

Participants were then asked “In Learning Objective 001, what number of items did the examinee get correct?” While getting the answer required a computation of calculating 86% correct of 14 items (12 items correct) some participants did not do the
calculation and reiterated “86% correct.” Of those participants who did the calculation, all but two correctly stated 12 items.

Participants were asked to indicate what they thought the cylinder chart represented. While most respondents answered either exactly or paraphrased what was listed on the explanatory page (“the % of questions from each objective that make up the entire test”), some indicated “learning objectives” and “% incorrect.” Most respondents made some reference to weight in that the Test Design chart could help you see the relative weight of each learning objective in relation to the entire test.

In addition to comprehension questions, participants were asked for their opinions on clarity and suggested improvements. In response to the question “In general, did you understand the information communicated in Score Report A?” all participants indicated “Yes”, or some variation (e.g., “pretty much” or “I think so”).

In response to the question “What suggestions would you make to better enhance the performance information communicated in Score Report A?” some participants indicated “none”, while others indicated the following:

- I thought it was clear
- I think this is easily understandable.
- I would explicitly communicate that each objective is weighted differently to determine the final score.
- Scale on cylinder chart is confusing
- Omit the bar graph since it repeats the initial table; make the cylinder chart description more specific - is it the percent of the test or the percent it counts
- Better understanding of the graphs
• I like it with the added page of information. My only comment would be to try
and put that information (for each section) with that section on the examinee score
report form.

• Put test design below performance, not besides. Gives equal weight. On same
visual line. Equal weight?

• Name the # of items student got correct out of possible correct

• No Changes Needed - Easy to Understand

• In addition to the percentage correct for each learning objective, I like seeing how
many questions I got correct for each objective

• Identify how much questions are worth, be them weighted specifically by
objective, or evenly weighed. It would also help to see how my % correct on each
objective equated to my total score of 180.

• The Test Design cylinder is a little confusing. Take a second to comprehend how
it is related. I would like to see how many questions the student got right for each
learning objective.

• Color coordination of objectives

  During the course of the interview, the following standard questions/opinions
were asked in order to get similar types and depth of feedback from participants.

Following are questions that were asked during the educator interviews in discussing
Score Report A.

• Was it clear from the report whether the examinee passed?

• Was it clear from the score report what learning objective was the weakest?

• Did the bar graph help or hinder your interpretation of the performance?
• Did the cylinder chart help or hinder your understanding of the test design? Do you understand what it represents?

• What additional information could be added to this score report to help in remediation for this examinee?

• What advice might you give to an examinee who received this score report?

• Can you think of any additional information that we could give to candidates who do not pass that might help them prepare to pass the test on a second occasion?

    Through the course of discussion, it was clear that all educators interviewed understood that the examinee did not pass, with one suggestion to bold the font for passing status, to draw more attention to it. It was also clear that learning objective 004 was the weakest performing, although one participant initially responded that the weakest was “learning objective 003 because of the 18%,” and then immediately realized that was the test design percent, not examinee performance.

    Many educators felt that although the bar graph helped with their interpretation of the performance, although some indicated they did not even look at it. Across the board, however, educators acknowledged that having both the percent correct in the performance table and the bar graph with the same information would likely serve both visual and numerical interpretation needs and preferences. One participant indicated that the bar graph helped and that’s “what we are used to looking at all the time.” One trend in response was that it was difficult initially to connect that the bar graph was merely a reiteration of the performance information already presented. In order to make this connection, some educators suggested color coding the learning objective names so it would be clear that if learning objective 001 was presented as blue in the performance
chart, it would also be presented as blue in the bar graph, and in the test design. The color coding, as it was suggested, would help to tie the three pieces of information together since learning objectives are represented in all three.

Regarding the cylinder chart, there were mixed reviews. Most educators indicated that at first they were confused by the information, or it took a moment to understand it. Essentially, at first glance the test design was less clear, but once understood many educators indicated that the relative weight or proportion for each learning objective was important information to know. Others indicated that since the information was test-specific, and not performance-related they were unsure it belonged on the score report, especially being on the same visual line as the your performance bar graph which made it seem even more connected to actual performance. Those who thought it was useful indicated it was helpful to know the examinee did very well in learning objective 001 and that was the largest portion of the test. For some, the display connected the weight or proportion to the examinee performance, which was helpful (although one educator pointed out that the information could be figured out from the performance chart). A few educators indicated the % of test rule along the left hand side of the cylinder chart made it more confusing. Another educator indicated the learning objectives were in opposite order (4 to 1, descending) than the performance chart, and another indicated they were labeled as “1” instead of “001”. In summary, the cylinder chart presented useful information, but there was question as to whether this information would be useful to the examinee, and how or where it should be placed and defined.

When educators were asked about additional information that could be added to the score report to help the examinee, many educators indicated they would like to see a
connection between the examinee performance and what is covered by the learning objectives, or a finer breakdown of the content and some remedial materials. The link on the backer was mentioned and most thought that if this could lead the examinee to the preparation materials or instructional tools available, that would be sufficient. Educators indicated they would like to see the raw score information, including the number of questions missed by learning objective, or the number correct, and an indication of the overall % correct on the test. Some educators indicated too that information on the backer could be brought to the front to help explain what was presented, such as the definition of the test design or the your performance bar graph.

Educators were asked what advice they would give an examinee receiving this score report, and almost all indicated to study learning objectives 004 and 002, including other suggestions such as enrolling in a basic math class, or referencing the texts and resources the examinee would have.

Other concluding remarks from educators on this score report included:

- It would be interesting to know if they completed the section. If I had a student who worked too slowly, it might be interesting to know if there were blanks.
- The link to additional information is the best thing. Thinking back to other exams, students struggle with where to go for information…helping students to understand where their weaknesses are is so important.
- It doesn't tell me how the other students did. I might be concerned about learning objective 004, but if I knew everyone had trouble in that area then I might not be as concerned.
• Include a sample question for each of the learning objectives so that the examinee preparing to retest would have an idea of what to expect.

• Supply the actual test questions the examinee missed, or if that is not possible be as specific as possible regarding the content I did poorly on.

• Certainly directing the examinee to a sample item wouldn't hurt anything. I do like that the report is clean, there is not a lot on it. Simplicity is the strength.

In summary, educators generally felt the presentation in Score Report A was clean and simple, although additional explanation or guidance would assist in connecting the “Your Performance” bar graph to the percent correct performance information in the chart, and to disconnect the Test Design as it was not related to examinee performance. Placement of score information was clear, however the request to link the learning objectives to either sample items assessing those objectives, or further detailed information on the learning objective demonstrated the need to have guidance beyond the report itself, and perhaps additional context to better understand more about the examinee performance.

4.4.4 Score Report B Feedback

In looking at the educator interview responses to Score Report B, all participants were able to correctly respond to the comprehension/identification questions listed below with their correct responses:

• The report reflects data from what administration date? (February 1, 2010)

• Did this examinee pass the examination? (No)
• In what Learning Objective did the examinee perform best (001, based on % correct, not N of items)

• In what Learning Objective did the examinee perform worst (004, based on % correct, not N of items)

Respondents did not answer all other questions identically. When asked how many examinees were represented in the score report, the intent was to get a response of “one” since this was a single examinee’s individual score report. However, the point was made that because median percent correct of passing examinees was included in the report, that data was for multiple examinees, therefore did the score report now represent only a single examinee?

Interviewees were also asked if they could identify the total number of items that assessed Learning Objective 004. All but one educator indicated “No” since there was no N of questions provided, but one indicated “just the percentage, 36%.” This is inaccurate as 36% of the test was not assessing learning objective 004. The examinee only got 36% of the learning objective 004 questions correct – you still cannot tell the total number of items assessing learning objective 004.

Because Score Report B included performance information of a reference group (passing examinees) it seemed comprehension questions regarding the examinee’s performance in relation to the reference group were warranted. For each learning objective, respondents were asked to indicate whether the examinee is performing below, at, or above the median performance of passing examinees. The intent was all respondents would indicate 001 and 003 were above the reference group, 002 and 004 were below the reference group, and no learning objectives were at that performance
level of the reference group. Two educators incorrectly responded about objective 002, one educator incorrectly responded about learning objective 003, and on learning objective 004 there was one response left blank and one incorrect. A potential rationale for these incorrect identifications is discussed later in this section.

Interviewees were asked whether from the information presented they could tell what percentage of the total number of items the examinee got correct. While the intent was that no one could not identify that total percent correct since no N of questions per learning objective was given, only % correct, one educator mistakenly indicated “Yes, by dividing the examinee score (180) by the total points available as listed on the backer (300).” This highlights a misconception of scaled scores, that somehow a scaled score represents a specific percentage correct within an examination area. While it is true that a scaled score begins by mapping a raw score performance (or % correct) to a point on the 100 to 300 scale, that does not mean that a scaled score of 180 out of 300 means the examinee got 60% of the test questions correct.

In addition to comprehension questions, participants were asked for their opinions on clarity and suggested improvements. In response to the question “In general, did you understand the information communicated in Score Report B?” all participants indicated “Yes.” In response to the question “What suggestions would you make to better enhance the performance information communicated in Score Report B,” some participants indicated either “nothing,” “looks good”, or the following:

- Overall percentage correct

- Inform test takers of the number of questions in each section & the weight of each section
• Add explicit information about the weight of each objective.

• Knowing the number of questions and number of questions right

• Better layout of information

• I would add the total number of items for each objective too.

• I would suggest having information about the number of questions in each learning objective included, as well as how much weight is put upon each objective.

• Use Grayscale Bar Chart and Simplify Key

• Put name and date at the top of the report. I really like seeing the median scores!

• Add the number of test questions for each learning objective

• The information at the bottom is ordinarily found at the top. Being at the bottom a student will overlook it.

• I would like to see an overall % correct, as compared to the median percentage of passing examinees.

• The Xs in the Your Performance chart should represent the test takers score. The different colors used in the graph and in the key required additional time to understand.

• Include number of questions for each learning objective category

During the course of the interview, standard questions/opinions were asked in order to get similar types and depth of feedback from participants.
Following are questions that were asked during the educator interviews in discussing Score Report B.

- Was it clear from the report whether the examinee passed?
- Was it clear from the score report what learning objective was the weakest?
- Did the chart help or hinder your interpretation of the performance?
- Did the use of color enhance interpretation of performance?
- Did the addition of performance information of passing examinees help with your interpretation?
- Would you prefer to provide a mean or a median? Should the term “average” be used instead given it is widely understood what that means?
- Do you think the choice of reference group was appropriate, or what do you think would be a good reference group? Everyone? Just barely passing? Just above barely passing? All Passing?
- What additional information could be added to this score report to help in remediation for this examinee?
- What advice might you give to an examinee who received this score report?
- What were you hoping to see on this report that would add to its utility?
- How would you compare Score Report B to Score Report A?
  - Are there aspects of Score Report A you liked better than Score Report B, or vice-versa?
  - Is there performance information in Score Report A that you found useful and did NOT see in Score Report B?
Through the discussion process, it was clear that many educators liked the clarity and information presented in score report B. It was clear to educators that the examinee did not pass, and most respondents indicated they liked the placement of the passing information in the top left corner. One educator commented that the examinee information was important to place at the top, especially if retaining a file of multiple score reports where it would be important to quickly see the administration date when flipping through a file. The educator also commented that the examinee ID would be helpful to have placed at the top as it is often needed or referred to when contacting the testing organization for customer support or questions.

Respondents indicated it was clear that objective 004 was the weakest, but when asked about the bar graph and use of color there was no overwhelming opinion that by using red, objective 004 was more easily identified as the weakest area. When discussing the use of color, some educators indicated it did not help them, and they thought it somehow related to the objective number, not performance. Once they read over the key, they found the understanding was clear, but upon first glance the red, yellow and green did not immediately strike home as danger, caution and ok. Some educators commented that the key and color were redundant information to present.

Interview discussion also touched on the median percent correct of passing examinees. While most educators welcomed this information and found it provided a useful context, some misinterpretations occurred with the bar graph. In designing the bar graph, the intent was for the bar to represent the examinee’s performance. It also seemed important to include the median percent correct performance in order to visually compare the examinee’s performance to the reference group. Unfortunately, by using “x” to
designate the median percent correct performance of passing examinees, a number of educators at first mistakenly thought the “x” represented the examinee’s performance, not the reference group. Because the eye was immediately drawn to the “x” an incorrect conclusion may have been made. This may account for the incorrect responses when asked if the examinee’s performance was below, at, or above the median percent correct performance. In the performance chart it is clear which number is higher or lower, but in the graph it took some interpretation and close reading of the key to fully understand what was represented. With regards to the usefulness of providing the statistic, most participants found it useful to provide as a context. One educator pointed out that through the context it is clear the examinee’s performance on learning objective 002 is closer to that of the passers, and therefore the passers also had difficulty on this learning objective. Therefore, they might advise the examinee to focus on learning objective 004 given the contextual information provided.

Other questions asked about the median percent correct of passing examinees focused on the statistic used and the reference group. Most educators indicated that median statistic was appropriate, and a few indicated that those preparing to become educators would need to be knowledgeable of mean, median, and mode. While educators agreed that an average or mean might be more widely understood or approachable, they indicated median was the best statistic to use. One educator indicated you could provide all three statistics, in order to show examinees the differences across each one. After explaining that the reference group included all passing examinees, from those who just barely passed all the way up to the “superstars” educators were asked is this group seemed appropriate. Overwhelmingly, the educators agreed that all passing seemed
appropriate. In fact, two educators drew the connection to the introductory explanation indicating the teacher certification examinations were criterion referenced and one must meet the criterion to pass. Since there are no varying “degrees” of passing, they indicated it seemed correct then to include all passing examinees in the reference group.

Interestingly, no questions were asked by educators about how many examinees were represented in this passing examinee group, or from which administration the data came. These questions were raised in the focus groups, but educators may have assumed the data represented the examinees who took the test at that administration. Given the current use of computer-based testing and continuous testing dates throughout the year, it seems presumptuous to assume that reference data on a score report for those other than the examinee come from the same administration date. Looking ahead, this was another point that needed to be clarified in developing score report D.

When asked to compare score report B to the report A, in terms of information given, and advice that might differ, most educators indicated they missed seeing the number of questions, and liked how that was provided in score report A. Some responded that the ordering of information in the performance table on score report B was clearer – first listing the learning objective, then the performance information helped to frame the information right up front. Educators were split between liking A more than B or vice versa, but all indicated that having the number of questions on A was helpful. In terms of advice, only one or two educators hinted at not emphasizing learning objective 002 quite as much since it was in the yellow and the reference information showed passing examinees were also performing at a low percent correct. Educators did not overwhelmingly indicate the advice would differ between the two score reports.
In summary, educators also liked score report B, although there was information they had seen included on score report A that they wanted to see on score report B. Use of color in the bar graph was not a major point of conversation, and while educators seemed to visually like it, no one indicated it was necessary in order to understand the level of performance. Only a few educators were taken aback by seeing the examinee information at the bottom of the page, but for those that did they indicated they liked it better at the top. All educators who commented on the placement of the passing status indicated they liked where it was placed. Regarding the explanatory information page, educators indicated they understood the description of median and the reference group. Overall there were some important pieces of score report B that seemed to have merit for future use.

4.4.5 Score Report C Questionnaire Results

Score Report C was an attempt to present another level of information and context, but proved outright to be educator’s least favorite score report. Some aspects of Score Report C were widely understood in that all participants were able to correctly respond to the comprehension/identification questions listed below with their correct responses:

- The report reflects data from what administration date? (February 1, 2010)
- Did this examinee pass the examination? (No)
- In what learning objective did the examinee perform best (001, based on % correct, not N of items)
As with score report B, when asked how many examinees were represented on the score report, all but one educator indicated “one”, with the remaining educator pointing out that the median percent correct of passing examinees represented a group of examinees, not just one. When asked what learning objective the educator might advise the examinee to study before retesting, all but two educators indicated learning objective 004. The remaining two educators advised both learning objectives 004 and 002.

When asked “from the information presented, can you tell if one learning objective is weighted more or less than another” the intent was for respondents to indicate that yes, you could tell based on the number of items associated with each learning objective. Only 7 educators indicated yes, one commenting that only “if you assume the number of questions affects weight”, and another commenting “you can figure it out but it would be nice to have this information on the report.” The nine remaining educators indicated no. Because the test design chart was not present, it was not immediately clear that the weights or proportions of each learning objective could be obtained by looking at the number of questions per learning objective.

When asked whether “from the information presented, can you tell what percentage of items on the total test the examinee gets correct” the intent was for responders to indicate that yes you can figure it out. By looking at the percent correct for each learning objective and the total number of questions, you could figure out the total percent correct of the total number of questions. Six educators indicated that yes they could tell, or that it could be figured out and calculated from the information presented. The remaining ten educators indicated that no, the percentage of items could not be derived from the information present on the score report. One of these educators
indicated that you couldn’t calculate it because it was not clear whether each item counted the same. Additionally, another educator indicated that you could derive it from knowing the total scaled score was 180. Again, this is a misconception that 180 (examinee’s scaled score)/300 (total scaled score) = examinee’s total percent correct on the examination.

Given the addition of confidence bands on score report C, and an overview provided on the explanatory page, it seemed prudent to ask educators about which learning objective band was the widest, and why one confidence band may be wider than another. With respect to which confidence band was the widest, 11 educators indicated learning objective 002, and two educators indicated learning objective 001. Two other educators indicated learning objective 004, with one indicating they chose it because the examinee was least likely to have the same results if they retested. Based on these responses, there may have been some ambiguity in how the confidence bands and the median performance “x” graph were interpreted. Some may have misread the x as the examinee’s score. Others may have misinterpreted the x as where the score might be if the examinee retested. Based on some of these misconceptions, and some encountered in score report B, it does seem that when “x” is presented on a graph or chart, the eye is drawn to it and the assumption is that it represents the examinee performance. Using x to represent anything may be misleading by design.

Educator responses to the question of why one confidence band is wider than another are listed below (three educators indicated they did not know why):

- difference between examinees score and likelihood of a similar score if taken again
• difference in reliability in testing each objective using given questions

• scores with bigger error bars are less reliable than those with smaller error bars.

• reliability

• more variability in performance

• there is less chance of an error for this testing.

• to show an interval over which the examinee would score if he or she were to retake the exam without any additional instruction.

• the reliability of the scores varies.

• the objective does not provide a very reliable score.

• because the range of the passing examinees was wider

• greater fluctuation in answers correct

• other examinee scores were more widely distributed

• the wider the band, the lower the reliability of the scores, and the greater possible variation in score may be observed upon retest.

Based on these responses, it seems clear that some educators correctly identified the relationship between the confidence band width and the reliability of the learning objective. Some responses, however, indicate that educators interpreted a relationship between the confidence band width and the performance of passing examinees, or the variability in performance across the learning objective. These are not correct inferences, in that there could be a wide range of scores within the learning objective, but if they are correlated to the overall performance of those examinees then the learning objective may be highly reliable.
Educators were also asked to indicate for each learning objective whether the examinee’s expected performance would be below, at, or above the median performance of passing examinees. Table 4.2 shows the distribution of responses to this question.

Table 4.2. Distribution of Educator Responses to Examinee Expected Performance

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Expected Performance in relation to Median % Correct of Passing Examinees</th>
<th>(N responses)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Below</td>
<td>At</td>
</tr>
<tr>
<td>001</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>002</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>003</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>004</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

The table results indicated that while most educators understood where the examinee’s expected performance (values within the band) would lie in relation to the “x”, median performance of passing examinees, it appears some mistakenly thought the “x” was the examinee score and they indicated where that “x” was in relation to the band (rather than the other way around). As mentioned previously, this was likely a misinterpretation of the graph.

In response to the question “What suggestions would you make to better enhance the performance information communicated in Score Report C” two participants indicated “none” and the remaining participants indicated:

- Clearer explanation of Performance confidence
• Use solid bars instead of the pencils and explicitly indicate and explain how scores are weighted by objective.

• Performance confidence level is confusing; narrative.

• The graph with the pencils seems too cute, so I’d eliminate it or use bars if you feel the graph is helpful; the performance confidence graph probably would not be helpful to the examinee in this format – B is easier to read.

• Better method of communicating information to examinee.

• I would keep the information about the confidence intervals here. I may even try to put a little bit of that information above that section on the report.

• Move Performance Confidence info to below pencil chart.

• I think the information about the performance confidence should be eliminated.

• This was very confusing, most examinees wouldn’t understand the confidence bands.

• Take away the graph with pencils. It makes the page busier and is not needed when all of that information is presented in other ways on the page.

• I found this more confusing because my eyes were drawn from one graph to another.

• Eliminate performance confidence graphs... I think I understand what it means, but I’m not sure why I need to know what it’s telling me. I do not like the pencil bar graphs. I wish it said clearly if I had passed or not at the top of the page.
• The pencil graph was imprecise, I instead looked at the table above. The confidence information was confusing.

• Clarity of confidence band graphics enhanced somehow?

Overall, ten interviewees indicated that in general they understood the performance information communicated in score report C. However, the remaining six educators indicated:

• generally, but not as clearly as A & B
• yes, but I had to think about it
• little more challenging to understand
• some of the information was clear and other information was not
• sort of…
• took a little longer to figure out how to read it but after reading the backer a few times helped. I did not use the information on the pencil graph at all

In general there was more disagreement about the clarity and utility of Score Report C than with score reports A and B.

During the course of the interview, standard questions/opinions were asked in order to get similar types and depth of feedback from participants.

Following are questions that were asked during the educator interviews in discussing Score Report C.

• Was it clear from the report whether the examinee passed?
• Was it clear from the score report what learning objective was the weakest?
• Did the “Your Performance” chart help or hinder your interpretation of the performance?
• Did the “Performance Confidence and Median Performance” chart help or hinder your interpretation of the performance?

• Do you understand the reason for the confidence bands to be different in width?

• Did the addition of median performance information help with your interpretation.

• What additional information could be added to this score report to help in remediation for this examinee?

• What advice might you give to an examinee who received this score report?

• What were you hoping to see on this report that would add to its utility?

• How would you compare Score Report C to Score Report B?
  
  o Are there aspects of Score Report C you liked better than Score Report B, or vice-versa?

  o Is there performance information in Score Report B that you found useful and did NOT see in Score Report C?

Through this dialogue, it became clear that while score report C presented some information directly, such as passing status, percent correct by learning objective and the number of questions per learning objective, educators had some opinions regarding the utility of the report. In particular, the Your Performance chart with the pencils was characterized as “too cutesy”, “bad”, and “not useful for all potential educators.” Although some educators liked the chart initially, they agreed that it might not be as widely accepted at all teaching levels and for all tests. One educator pointed out that it in a way the graph seemed out of character and potentially demeaned the seriousness of the
teaching profession, indicating that a medical doctor would not receive a score report with stethoscopes included in the graph, why should a teacher receive a score report with pencils in the graph?

As with prior score reports, a request was made for the raw scores, number correct and overall percent correct. Placement of the passing score status was requested to be moved to the top. In discussing the performance confidence chart, a number of educators indicated the examinee would not need to know that level of detail and information, and stated it was the least useful information across the three score reports. Interestingly, though, some noted that as an advisor the confidence interval might influence the advice they would give to an examinee, in that they would more fully understand which areas could potentially shift the most, or that the examinee might be encouraged to see the confidence band included the median percent correct for all passing examinees. Overall, in comparing score report C to others, many educators immediately indicated they preferred score reports A or B to score report C.

4.4.6 Other Feedback From Educator Interviews

After being presented with all score reports, educators were then asked to consider all three reports. A dialogue took place regarding all three reports, especially in response to the following questions:

- Would you have preferred to see the total number of questions the examinee got correct by objective (the raw score) or the percent correct?
- Are there other preferred methods of displaying examinee results that we have not talked about here? (what and why preferred)
Can you suggest any other ways to communicate the concept of imprecision regarding the learning objective scores? Do you think it is important to communicate imprecision information?

What information was *most* useful across all three score reports? For a “failing” examinee, what information would be the most important to help them prepare to pass the exam?

What information was *least* useful across all three score reports?

Did one score report stand out as more detailed or informative, or clearest in its meaning?

Did one score report stand out as least useful or informative?

Most respondents indicated they understood the use of % correct, but they would also like to see raw the score, or would have preferred to see the raw score. In addition to raw score, some educators stated they would want to see not only the number of questions correct, but also an indicator or gauge of how many incorrect.

In discussing other ways of presenting data, most educators indicated they could not think of an additional method of presenting the information. Of those with suggestions, one indicated that a pie chart would be a better way to represent the test design. One asked whether a bell curve would better illustrate how the examinee’s performance was in relation to the median percent correct of passing examinees. Another suggested if there were a way to represent the examinee’s total performance on the test using color, that might be useful. Some educators asked whether text descriptions might serve better than graphs or statistics. For example saying “For learning objective one, your total percent correct was 86.” Educators also indicated that any additional
information that could be provided would be welcomed, such as sample items or more specific information about the learning objectives.

As mentioned previously, educators were somewhat split over whether imprecision information was important to communicate to the examinee. However, most indicated that what was provided by way of explanation was sufficient.

The most useful information across the three score reports seemed to be either the % correct by learning objective (numerical) or the median % correct of passing examinees (numerical). Information that was least useful across the three score reports was the pencil graph.

When asked which score report stood out as being the most informative or clear, educators were split between A and B. But almost every educator indicated score report C was least useful, due to the pencil chart and the questionable utility to the examinee of confidence bands and imprecision information. Overall there appeared to be a clarity and clean aspect to A and B, while C included information that educators were not convinced the examinee would need in order to understand where their strengths and weaknesses were.

Additional comments received from educators included some positive remarks about the use of the Web-meeting, and the method in which the materials were presented and accessed for the interviews. Some educators saw this as a valuable tool they might make use of in their own academic environment.
4.5 Focus Group With Educational Testing Professionals

In order to incorporate feedback from individuals directly involved in the design, explanation, or otherwise impacted by score reporting, professionals from an organization that develops, administers and scores educational assessments were asked to participate in a focus group to review and obtain feedback on Score Reports A, B, and C.

4.5.1 Overview

The focus group was held in one single session, which lasted approximately one and a half hours. A total of six professionals participated in the session. Participants were told that the score reports had been reviewed by individual educators through an interview process, and also informed that university education students would be seeing a report designed using that feedback and the feedback from their focus group.

Participants saw each score report separately, and answered the individual questions by filling out a questionnaire for each score report. Once everyone completed the individual questionnaire, a dialogue was had regarding each score report and the reports as a whole.

4.5.2 Participant Demographics

Participants were invited to be part of the focus group, and each participant had experiences in designing, interpreting, or explaining score reports. The participants cut across different areas of the testing organization, and each brought a unique perspective
to the reports and their interpretation and utility. The demographics of the group are found in Table 4.3.

Table 4.3. Demographics of Educational Testing Professionals (Focus Group)

<table>
<thead>
<tr>
<th>Years of experience working in education</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–3 years</td>
<td>1</td>
<td>17%</td>
</tr>
<tr>
<td>4–10 years</td>
<td>1</td>
<td>17%</td>
</tr>
<tr>
<td>11–15 years</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>over 15 years</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>100%</td>
</tr>
</tbody>
</table>

Job Description within Testing Organization
Psychometrician
Customer Service
Customer Service
Data Analyst
Marketing Director
Management

<table>
<thead>
<tr>
<th>How would you categorize your level of experience with educational tests and statistics?</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Minimal experience</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Moderate experience and use throughout the academic year</td>
<td>1</td>
<td>17%</td>
</tr>
<tr>
<td>Ongoing experience and use throughout the academic year</td>
<td>5</td>
<td>83%</td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>100%</td>
</tr>
</tbody>
</table>

4.5.3 Score Report A Feedback

In response to Score Report A, all participants were able to correctly respond to the comprehension/identification questions listed below along with the correct response:

- The report reflects data from what administration date? (February 1, 2010)
- How many examinees are represented on this score report? (One)
- Did this examinee pass the examination? (No)
What is the title of learning objective 002? (Computation and Problem Solving)

In learning objective 004, how many total items did the examinee see? (11)

Does each learning objective have the same number of questions? (No)

In what learning objective did the examinee perform best? (learning objective 001: Number Properties and Operations)

When asked, “From the information presented, can you tell if one learning objective counts more or less than another” all participants indicated yes they could tell. However, one participant indicated they could only tell if the relative number of items divided by the total items equaled the weight that was implied.

The group was asked “In learning objective 001, what number of items did the examinee get correct?” To obtain the correct response, a computation of calculating 86% correct of 14 items (12 items correct) was necessary. Four of six respondents indicated “12”, one indicated “4,” and one reiterated “86% correct.” It was unclear how four was derived by one of the participants. Additionally, when asked “in what learning objective did the examinee perform worst” all respondents but one correctly indicated 004. The other responder indicated 002 – which was an area of attention, but not as poor as 004. When asked what they thought the cylinder chart represented, responses referenced the weight of each objective, or the proportion of the test that each learning objective accounted for.

In addition to comprehension questions, participants were asked for their opinions on clarity and suggested improvements. In response to the question “In general, did you
understand the information communicated in Score Report A?” all participants indicated “Yes.”

In response to the question “What suggestions would you make to better enhance the performance information communicated in Score Report A,” responding participants indicated or asked the following:

- Consistency in labeling of learning objectives (001 vs. 1). Does examinee care about test design? Is that graphic helpful? Change label “N Questions” to “# of” or “No. of” Will examinees know “your performance” chart corresponds with “% correct” column?

- What are minor fluctuations? Different s of questions for different test forms? Reword “retaking the test” in backer.

- The test design chart is unclear in its purpose at an initial glance. Perhaps the explanation as to its purpose can be placed on the front of the sheet, instead of the backer. The performance chart is slightly repetitive but does allow for a different visual perspective regarding one’s performance. Instead of just stating % correct next to the number of questions, you could list the exact number of test questions they got correct as well, but this could lead to the temptation to average one’s score. What percentage do they need to pass? Do we want to provide this information or an approximation?

- Perhaps put “Your Score” and “Minimum Passing Score” together (aligned, and passing status off to right.)
• Can anyone retake, or should this be directed to nonpassers? Perhaps add more explanation about the “Your Performance” graph – to focus additional preparation as needed. Second sentence in “Retaking the Test” may be confusing.

Participants found that it was clear from the score report that the examinee did not pass, and that it was clear learning objective four was the weakest. In discussing the bar graph, some interesting points were raised. It was asked whether perhaps because the graph was disconnected from the information listed in the chart above whether that would cause a comprehension issue. An examinee may find this potential disconnect between two pieces of information, even though the statistic was the same. It was suggested that the explanatory text for the bar graph be brought to the front of the report to help clarify what it represented, or to use the same heading as in the chart above (“% Correct”). Another source of confusion may also have been the use of the term “N” as a heading for the number of questions. The focus group indicated this was more “psychometric talk” and could be clarified by saying number of questions. This raised the issue as to whether an examinee might misinterpret this to be the number of questions they got correct versus the number of questions on the test.

With regards to the test design, the group asked whether an examinee would really be concerned with the test design. As educational testing professionals, it was indicated that an examinee should know the test design going into the exam, since this information is made available. So, while the group understood this information, they felt it was not important in the context of the score report, as examinees would already have it or could find it elsewhere. If it did need to be included, a suggestion was made to put the design with the “understanding your examinee score report” text. The issue seen with
this, however, is that each explanation page would then be customized to a test, rather
than to a testing program. For a testing program with few tests, this may not be
problematic. However, for a program with numerous tests (some teacher certification
programs may have as many as 30, 40, or 50 individual tests) having a customized
explanatory page might be a potential issue for quality control, in that one would need to
ensure the proper explanatory page was matched with the proper test score report.

Another suggestion was made to represent the test design as a pie chart, and to
incorporate the bar graph into the performance table up top somehow, to tie the results
and bar graph together.

As is illustrated from the feedback provided individually, participants had some
interesting questions about the nature of the Fundamental Skills: Mathematics test and its
score report. One question asked was whether field tested (i.e., non-scorable) items
would be reported on the score report. The response was no, while the test likely would
have some items that were on the test for purposes of gathering information on their
psychometric qualities, those items would not be part of the score reporting information
since they did not count towards an examinee’s score. Given this, an examinee may see
the “45” listed no the score report as the number of questions but might remember that
they actually took a 50+ item test. This may prompt further questions from the examinee.

Therefore it was suggested that the examinee be informed that additional questions may
have appeared on their test form, but those were there for experimental purposes and did
not count towards the examinee’s score. A place to explain this may be in the
explanatory page.
Another question asked whether once an examinee passed if a scaled score would be reported. While this is not handled consistently across teacher certification testing programs, the intent for these mock score reports was for scaled scores to be reported along the entire scale of 100 to 300, and if an examinee got a score of 200 or higher they would still be presented with their total scaled score.

Participants also commented on the text in the explanatory materials. There was some concern that discussing “fluctuations” in scores might be confusing to examinees, additionally it might encourage retesting on the examinee’s end since it was stated there are fluctuations expected. There was a question as to whether “fluctuations” should be defined more clearly in the materials, or deleted altogether. It was asked whether the “retaking the test” section of the explanatory materials might only appear for failing examinees. Some other formatting and line spacing suggestions were made to the explanatory page for ease of reading, which would affect score reports A, B, and C. In general, score report A seemed well-received, but some critical questions were asked, and observations made.

4.5.4 Score Report B Feedback

With Score Report B, there was some consistency of responses to the individual questions. The testing professionals all correctly responded to the administration date reflected in the report, whether the examinee passed, whether the total number of items assessing learning objective 004 could be identified, in which learning objective the examinee did best and worst, and for each learning objective the testing professionals
correctly indicated whether the examinee was performing below, at, or above the median performance of passing examinees.

In response to the question asking how many examinees were represented on the score report, the testing professionals also pointed out that while the score report reflected one examinee, the median information involved multiple examinees. When asked whether one could tell what percentage of the total number of items the examinee got correct, all but one respondent indicated No. Based on discussion, it seemed clear that the group understood you could not see the total percent correct, nor the total number of items on the test, but could see percent correct by learning objective.

In general, the focus group found score report B understandable, although indicating that the median may not be well-understood, and noting the “numbers needed to pin down the information were missing.” When asked to provide suggestions to enhance score report B, responses were as follows:

- Not as visually appealing, could be fixed by moving address and date to top of page. Place box around graph and key to delineate from other sections/information. In backer or on website provide an example and explanation of numbers. Need more info regarding color coding? Ex. 70% and above suggests passing performance? Font of status and score should be larger, information of greatest interest.

- Test date, etc. at top. No median % correct (not needed). Performance bar graph should not include median % correct. “Fluctuations” on backer, what does this mean? Performance bar graph, include w/ “Your % correct”
• Add N items tested per objective. Change the math symbols in key to text, not symbols.

• Some questions that may arise: The median information is confusing and unnecessary. If you received 68% in subarea 4 for example (or the median % in each subarea) do you pass the test? If an examinee received the exact % of questions correct in the median column, do they pass? What is each subarea worth toward your final score? Did you fail because subarea 4 was worth 75% of your total score? The colors are not necessary. If you achieve all green do you pass? It’s weird that the personal information is on the bottom. The median % includes all examinees. Is there a minimum you can achieve in order to pass?

• Larger font for “status” line. Test design (relative importance).

• Not include median % correct if this is a Criterion referenced test. Move test date and examinee info to top of report

After completing the individual questionnaires there was some involved discussion regarding score report B. It was clear there was some information such as passing status and weakest performing learning objective that were understood by all. It was requested the text at the top indicating passing status be presented in a larger font. The testing professionals were also not used to seeing the examinee and administration-specific information at the bottom of the report, indicating this seemed strange. However, in discussing the bar graph, and especially the color-coding of red, yellow, and green, much disagreement was raised. While on the one hand, it was eye-catching and an immediate visual cue, there were concerns with interpretation, and some key questions were raised:
• How were thresholds of less than 50, 50 to 70 percent correct, and greater than 70 percent correct established?

• Would those thresholds somehow equate to a passing status?

• Would it be possible for an examinee to get all green bars (above 70%) on each learning objective and still not pass?

• Would the thresholds stay the same on every test, or should they adjust based on difficulty of the exam?

These were valid questions regarding the color-coding of performance. As it was explained to the focus group, the thresholds were chosen as arbitrary points of possible warning, caution, and clearance as many in education may be familiar with. As to their global applicability across all tests in a testing program, there would be concern that some tests with lower passing standards (fewer percent correct required to meet the passing standard) it might be possible for an examinee to be in the yellow on most learning objectives and still pass. And for a test with a higher passing standard (greater percent correct required to meet the passing standard) it might be possible for an examinee to have all green bar performance in each learning objective, yet to still not pass. From a customer service perspective, it was understood that these scenarios might be hard to explain. The question is whether any potential utility of the red, yellow and green outweigh the potential confusion in interpretation when dealing with tests of varying difficulty and performance requirements. Additionally there was concern with using the mathematical symbols for “greater than” and “less than” in the key, and there was a suggestion to spell this out rather than relying on the symbols.
Similarly there were a number of questions raised regarding the addition of median percent correct of passing examinees. Some key questions were:

- Are these examinees from just this administration or from prior/other administrations? Would the data change from administration to administration?
- How many examinees are represented in these data?
- Is it appropriate to provide a reference group on a criterion-referenced test?
- Would it be possible for an examinee to meet the median percent correct for all learning objectives and still not pass?
- If I pass, might I be tempted to retake the test to reach or exceed the median percent correct of passing examinees?
- If I pass, might an employer look to see if I at least met the median percent correct of passing examinees in all learning objectives?

Again, this feedback facilitated a valuable discussion, through which it was clear that an explanation or definition of the population referenced in the statistic was warranted. Additionally, questions such as its appropriateness brought to light that at times a testing industry may hesitate in providing a statistic given it may be perceived to be inappropriate, in contrast to the educators positive reception of receiving this information.

In general a number of questions and concerns were raised about score report B, the appropriateness of color-coding performance, and providing reference group information. There was a notable difference in opinion from educators to the focus group
of testing professionals, given some of the insight and cautions offered by the testing professionals focus group.

4.5.5 Score Report C Feedback

Individual questionnaire responses to score report C indicated that all focus group participants were able to correctly identify examinee passing status, administration date of the test, best performing learning objective, and widest learning objective confidence band. Again, when asked how many examinees were represented on the examinee score report, respondents made reference to the fact that the median performance was likely for more than one examinee, but they were unsure of how many there were and from what administration date they represented.

When asked if it could be determined from the information presented whether one learning objective was weighted more than another, two focus group respondents indicated yes, since the number of items was presented. The remaining respondents either did not provide an answer, or stated “no.” When asked if the percentage of correct items on the total test could be found, four respondents indicated yes, and that it would need to be calculated, while one other respondent stated no and another left the response blank. All respondents in the focus group indicated that they would advise the examinee to study learning objective 004 before retesting, with one respondent adding learning objective 002 to the recommendation.

When asked to indicate whether the examinee’s performance was below, at, or above the median percent correct performance, responses were as follows:
Table 4.4. Distribution of Educational Testing Professionals Responses to Examinee Performance

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Performance in relation to Median % Correct of Passing Examinees</th>
<th>(N responses – testing professionals focus group)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No response</td>
<td>Below</td>
</tr>
<tr>
<td>001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>002</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>003</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>004</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

A note about this question and the responses. For the individual educators, the question was phrased as to whether the examinee’s “expected” performance was below, at, or above that of the median percent correct. For the focus groups (both testing professionals and students) the term “expected” was inadvertently omitted, and therefore the question asked primarily looked at whether the focus group participants could identify where the examinee’s performance was in relation to the median percent correct (as with the same question on score report B). Based on the responses, there may have been some confusion for objectives 002 and 003, given the “expected performance would include (at) the level of the median percent correct of passing examinees. However, the actual performance for 002 was below the median, and 003 was above the median. It seemed that would be a reasonable explanation as to why the group was split between designations for 002 and 003.

In responding to the question regarding why the width of the confidence band may vary, respondents made reference to reliability and understood the relationship between the varying width and the reliability of the learning objective. The item statistics
in learning objective 002 were less reliable than in the others, hence the wider confidence band.

In general, for score report C certain information was understood in how it was communicated, and other information was not. The focus group participants indicated that passing status information was clear, but that performance confidence was a bit confusing. There was uncertainty from one respondent as to whether the performance confidence was related to the median performance of passing examinees. Another respondent reiterated they did not feel that the peer performance (the median percent correct of passing examinees) was relevant.

Suggestions made to enhance the performance information communicated in score report C, and some focus group individual questions follow:

- Prefer bar graph to pencils
- The pencils were confusing and it is difficult to see exactly where you fall within the percentage correct. The top of the pencil in objective 002 falls over 60%, but it is missing its eraser. That’s a small detail that might not be noticed. The information regarding performance confidence is confusing. When reading the explanation, there is mention of “error.” Many examinees may assume that the computer made an error in the scoring process. What is each objective worth toward the final score? The Xs could be considered the examinee’s scores.
- Add middle hash mark to confidence interval band. For this audience, don’t use pencils in graph, better suited for K-12.
• Move status up with examinee’s score. Do not include performance confidence or median performance of passing examinees. What does a ½ pencil mean? Use of “error” in explanation is not good.

• Do not think examinees will understand confidence bands from backer. Hard to define without saying “there is error” and “we’re no 100% sure of your score.” Do not like pencil graphic, prefer horizontal bar presentation. Prefer status info next to score info. Will examinees think “x” on graphic equals their score?

• Do not include confidence interval – too confusing and not helpful to candidates. May cause examinees to focus on these statistics. “Error” in tests. Low reliability of scores vs. identifying areas of weak performance to focus future test prep. Graph depicting confidence interval and median passing scores may be confusing. May be hard to relate it to examinee’s score.

Through discussion of score report C, additional suggestions were made. It was suggested the pencil graph of Your Performance be rotated and presented horizontally. It was also communicated (as was mentioned in the educator interviews) that the “x” in the performance confidence and median performance graph was initially misinterpreted to be the examinee’s score. It was discussed whether “error” should be explained in more clarity in the explanatory page, to which there was mixed feelings about communicating the concept of error, and the possibility some examinees may see the term “error” and think there was something wrong with their test. One suggestion was to remove the entire “Performance Confidence” paragraph from the explanatory material. It was suggested too that “weighting” be discussed in general, but as part of the explanatory page. Explaining, for example, that each item counted for the same value, there was no
penalty for guessing and that learning objectives with more items on the test than others meant those are “weighted” more heavily since they made up a larger proportion of the test.

4.5.6 Other Score Report Feedback

Overall, the focus group was asked to reflect back on the three score reports. The group seemed to feel that while it was clear across all three reports that learning objective 004 was the weakest, score report A stood out as being the clearest and most informative. Information that was found to be most useful was the Your Performance graph in score report A, without any color-coding or arbitrary cutoffs, and the test design so examinees understood how the test was “weighted.” As mentioned, however, the test design was discussed as not necessary to include on the score report as it was not related to performance, but it was valuable information as related to the test design. It was agreed that examinees needed to understand there was some imprecision in test scores and performance, however the focus group was not in agreement as to where and how that information should be communicated.

The group unanimously agreed that the most important information to communicate to a failing examinee was “what they need to do in order to pass.” In discussing the best ways to communicate this, again it was clear that while providing concrete performance information such as raw score, or percent correct may be beneficial for the examinee (instead of providing more general performance indicators), there was still a disconnect between that performance and the scaled score. This was clear in the educator interviews as well. It was asked in the focus group whether an “approximate
“percent correct needed to pass” could ever be provided. Potential issues with this is that
given equating, this value is expected to change slightly. Even with that, though, there is
still no explanation given that shows how that percent correct is scaled. An examinee sees
their % correct performance from one administration date to the next is the same, but
their scaled score changes. The connection between scaled score and the raw score or
percent correct seemed to be a continuing source of confusion and potential room for
examinee distrust or misunderstanding.

4.6 Processing Interview and Group Feedback and Designing Score Report D

Score Report D was meant to be a product representing all the feedback, pros, and
cons received from score reports A, B, and C. After considering the feedback from
educator interview and the focus group with testing organization professionals, some key
points were clear that needed to be addressed in creating Score Report D. Some of the
areas needing to be addressed were visual in nature, others were statistical. The key
points addressed included incorporating the bar graph with the performance data,
removing the use of color in the bar graph but keeping the reference point of the median
group, providing raw score information, providing reference group information (median
of passing examinees), removing references to confidence intervals, and ensuring the test
design was clearly separated from the actual examinee performance information.

Visually, educators commented that the graphs were more difficult to interpret
because they were separated from the table outlining examinee performance. This gave
the impression that the graph represented data that was new, or had not been seen
already, when in reality the bar graph was a visual representation of percent correct
information that had already been provided in the performance table. This separation was due partly to the portrait orientation of Score Reports A, B and C, and partly due to the size of the bar graph. For Score Report D, the orientation was changed to landscape, and the performance bar graph was brought up with the chart of examinee performance. By merging the performance chart and the Your Performance bar graph, the educator issue of connecting visual performance to specific learning objectives was also addressed. With the new orientation, the reader’s line of sight would be drawn across the page and would be presented with all performance information for a specific learning objective on a single visual line. The learning objective number and title were retained in the leftmost column in order to present the context first before the performance information. The test information and passing status were kept at the top, with passing status in the top left line of sight. The test administration date was placed at the uppermost corner of the score report based on feedback indicating that for examinees who retest and have multiple score reports, having the administration date at the top is ideal for filing purposes.

Another visual aspect of Score Reports A, B, and C that was commented on was the use of color. While some educators remarked how the use of color in the bar graph was very helpful in visually drawing the eye toward areas of concern or caution, some educators found it provided no value added. Additionally, through the discussion with testing professionals, some concerns were raised regarding the key and choice of percent correct thresholds for performance and relative color designation. For example, an examinee hypothetically could get all green bars in the graph and not pass the examination if there were a high standard or cut score on the test. Additionally, not every content area may have the same expectations regarding the red, yellow and green
thresholds. That is, it may not be appropriate to use the same color bar key on a Physics teacher certification test as used on a Basic Skills test. For this reason, and no overwhelming reaction to its use, the color performance bars were omitted. Comments received about including the median percent correct of passing examinees in the bar graph indicated that when “X” was used to identify this statistic, it was misinterpreted to be the examinee’s own performance. Therefore, an “X” was used in the bar graph to indicate the examinee’s percent correct (in essence, an “X” was placed at the end of each bar in the bar graph) and an “M” was used to designate where on the bar graph the median percent correct of passing examinees fell. The hope was that confusion between the individual and the passing group performance would be cleared up while still being able to present both in the same graph.

Visually, the cylinder chart used in Score Report A was appreciated but not necessarily the most accessible method for presenting the information regarding test design. Educator feedback indicated that a pie chart might work best for displaying this information, and suggested that it be either placed on the explanatory page (“Understanding Your Examinee Score Report”) or placed in a different visual line since it does not contain actual performance information. For Score Report D, the test design was placed in a location separated from examinee results, and a frame was placed around it to further distinguish it from the examinee data presented. The test design was presented as a pie chart, with each objective shown as a different shade of grey, each listed with the relative percentage of the test.

Statistically, feedback from educators indicated that while percent correct was informative, it was also necessary to indicate how many items were presented in each
objective. Some educators also indicated the importance of providing a raw score or total number correct by objective. In designing Score Report D, all three pieces of information were presented and formatted in a way that the examinee could easily tell how many test questions were missed and how large the learning objective was. A frame was placed around this information, to help visually set off these performance indicators. Additionally, a total test percent correct was added, based on feedback from educators. Finally, the median percent correct of passing examinees was retained and placed after the examinee percent correct and before the bar graph. The final Examinee Score Report D is shown in Figure 4.10.

The added statistics warranted some additional explanations in the “Understanding Your Score Report” document drafted for Score Report D. Text in the document was clarified and elaborated on to include further descriptions of scaled scores and the group comprising the median % correct of passing examinees. For scaled scores, feedback was received from the testing professionals group as to whether examinees would understand that on one occasion a raw score and percent correct may equal a certain scaled score, and on another occasion that scaled score may be different even though the raw score and percent correct have not changed. Again, this is dependent on equating and the difficulty of the test form. For that reason, it was suggested that if raw score information was presented, an explanation regarding changes in scaled scores should be provided. Based on questions about who comprised the group of examinees referenced in the “Median percent correct of passing examinees” a definition of this group was also added to the explanatory page. The final Understanding Your Examinee Score Report for score report D is shown in figure 4.11.
Figure 4.10 Examinee Score Report D
4.7 Conducting the Focus Group with University Doctoral Students

Doctoral students from the Research and Evaluation Methods Program (REMP) at the University of Massachusetts were asked to participate in a focus group in order to get
their feedback on Score Reports A, B, and C, as well as the newly designed Score Report D. Participants were told that the score reports had been reviewed by individual educators through an interview process and testing organization professionals through the focus group. Students were informed that they would be seeing a report designed using that feedback.

4.7.1 Overview

Education students attended voluntarily. They were asked to attend an open session focus group, during which they were informed that a review of sample score reports for a teacher certification test would be carried out. It seemed important to not just describe score reports A, B, and C to the students, but to lead them through a similar process as the prior participants so that score report D could be reviewed in light of the prior score reports reviewed. No tenured faculty attended the sessions, and introductions were made at the beginning of the session. The session had a collegial, yet professional atmosphere and students were willing to speak freely about the reports and provide their opinions.

4.7.2 Demographics

Ten persons participated in the focus group, all were graduate students except one who was an adjunct assistant professor and she was very knowledgeable about score reporting. All the participants were part of the University of Massachusetts REMP program which specializes in psychometrics and other educational measurement concentrations. Responses to demographic questions are found in Table 4.5.
Table 4.5. Demographics of University Education Students/Representatives (Focus Group)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>6</td>
<td>60%</td>
</tr>
<tr>
<td>Male</td>
<td>4</td>
<td>40%</td>
</tr>
<tr>
<td>No Answer</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2. Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>1</td>
<td>10%</td>
</tr>
<tr>
<td>Black, not of Hispanic origin</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Native American</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>White, not of Hispanic origin</td>
<td>9</td>
<td>90%</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3. Student Description</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undergraduate student</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Graduate student</td>
<td>9</td>
<td>90%</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>10%</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Student Description (e.g., Major)**

Responses:

8: Research and Evaluation Methods Program
1: Post-doc

9 responses, 1 no response
Table 4.5. (continued)
4. In relation to #3, which of the following best describes your status as a student?

<table>
<thead>
<tr>
<th>Status</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>First year student</td>
<td>5</td>
<td>50%</td>
</tr>
<tr>
<td>Second year student</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Third year student</td>
<td>1</td>
<td>10%</td>
</tr>
<tr>
<td>Fourth year student</td>
<td>2</td>
<td>20%</td>
</tr>
<tr>
<td>Part-time year student</td>
<td>1</td>
<td>10%</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>10</td>
<td>100%</td>
</tr>
</tbody>
</table>

“Other” Description from #4 (e.g., Major)

Not a student at this time

5. How would you categorize your level of experience with educational tests and statistics?

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Minimal experience</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Moderate experience and use</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>throughout the academic year</td>
<td>1</td>
<td>10%</td>
</tr>
<tr>
<td>Ongoing experience and use</td>
<td>9</td>
<td>90%</td>
</tr>
<tr>
<td>throughout the academic year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>10</td>
<td>100%</td>
</tr>
</tbody>
</table>

4.7.3 Score Report A Feedback

The focus group session with doctoral students began with an introduction of the topic, the purpose of the session, and then an individual review of score report A. From the individual perspective, many responses were in agreement. All students correctly indicated the administration date, the number of examinees represented on the score report, the examinee’s passing status, the title of learning objective 002, the total items
seen in objective 004, that each learning objective did not have the same number of questions, and the examinee’s best and worst performing learning objectives. All students correctly indicated what the cylinder chart represented, referencing percentages of the test, the test blueprint, and the breakdown of the test/weights.

Some questions with differences in individual responses included whether one learning objective counted more than another. While seven students indicated yes, and some referenced the test design or the number of questions as showing that information, the three remaining students indicated that they could not tell if one learning objective counts more or less than another. One respondent indicating “no” also added that “it is possible that they are weighted before being scaled.” Additionally, when asked about the number of items the examinee answered correctly in learning objective 001, all but one student indicated 12 (after doing the calculations). One student indicated “about 8” and that it was not clear from the information presented.

Overall, the doctoral student focus group indicated the information communicated in score report A was understood, one mentioning it was “basically straight forward if you didn’t over think it.” When asked to make any suggestions to better enhance the information communicated, the students provided the following responses:

- Number correct, consistency in terms, purpose of test, score range, error band
- Listing of incorrect items if items are available somewhere for examinees to look at. Listing of the score out of the total possible score points.
- You could include number correct and weighted percentage in the main table
- Label learning objectives as 001…004 on the cylinder chart. Not clear if questions in each learning objective are scored the same, so is one section “worth
more” than another? What types of questions were answered incorrectly (to help prep for the future)?

- Color would improve test design cylinder. Lots of open space – could make things (charts) larger without cramming things. Make the status stand out more.
- It seems quite clear to me. Although, I may want to see how I compared to others taking the same test on the same occasion for curiosity, even while this may not be norm referenced.
- Lots of redundant information, Divide % correct info, # correct as well.
- I didn’t see a place where it said the # of items correct for each learning objective, just percentages. Maybe include this information? Also, describe each learning objective on the second page.
- Number correct per Domain, Norms/Averages for passing scores.
- Maybe the specific benchmark of the content that the examinee got wrong on framework, if there exists such content.

In the focus group discussions that followed the individual feedback, it was clear that the passing information and the weakest objectives were easily identified by the students. In discussing the bar graph, it was clear that some students found it helpful as it was a visual representation of the best and worst performance, “not just the numbers,” while other students indicated they did not even look at it and only looked at the numbers. As to whether to include the graph or not, students indicated that visual learners might benefit from having it, so it would be good to include it.

In discussing the cylinder chart, students indicated it would likely lead to more questions, although they understood the information it communicated. The point was
made that in the performance chart information is ordered left to right: number of questions then the percent correct, but in the two graphs left to right you first get a visual of percent correct, then a visual of number of questions. Flipping the order of the graphs to match the order of presentation in the performance chart was suggested. Another suggestion involved incorporating the two graphs directly into the performance chart by including a visual cylinder next to (or underneath) the number of questions column, and including a horizontal bar next to (or underneath) each percent correct statistics. In short, there needed to be a stronger tie between the information in the performance chart and the visual graphs provided below the performance chart. It was also suggested that confusion may arise by having two sets of percentages – a percent correct, and a percent weight for each objective. The test cylinder could instead list the number of questions out of 45 in a visual way, the same information but keeping it on the number of questions scale rather than percent correct scale. At a minimum, students suggested the labeling of the chart and corresponding graphs should be consistent. If “001” is used in chart, “001” should be used in the test design. If “%” is used in the chart, keep “%” rather than “percentage” in the graph.

Other feedback received on score report A included the suggestion of providing even more information on the learning objectives, and providing the number of questions the examinee got correct. While it was pointed out that the score report explanation would point to the further clarification of content included in the learning objectives, it was asked what other helpful information could be provided to again help focus the examinee’s studies.
4.7.4 Score Report B Feedback

Next the students were presented with score report B. Individual responses to the questionnaire showed that (aside from one response of “yes”), the students correctly identified the administration date. All students responded that the score report reflected only one examinee, but some made mention of the group including other students although it was unknown what those numbers represented were. The students could clearly indicate that the examinee did not pass the examination, and that they were unable to identify the total number of items in learning objective 004.

All students responding correctly indicated for each objective whether the examinee’s performance was below, at, or above the median percent correct of passing examinees. For learning objective 002 and 003, there was one non-responder. All students correctly indicated that the examinee’s best performance was in learning objective 001 (Number properties and Operations) and their worst performance was in learning objective 004 (Measurement Concepts and Principles of Geometry). All students but one correctly indicated that there was no way to tell the total percentage of items the examinee got correct. One student indicated “about 64%” which would be what is calculated if you added up the “Your % correct” for each learning objective and divided by four. However, this is not truly an average as each section has a different length and weight. This happens to be correct in this case (given the sections are very close in length to one another) but is not an accurate way to solve for the total percent of questions correct. In this case the examinee got 29 of 45 items correct (64% correct) but that cannot be derived from score report B. The same raw score performance with different “Your % correct” numbers would yield a different calculation using the student’s process. This common
practice, however, is likely used by examinees receiving a score report. They may believe that since there is no total percent correct provided that if they average the four that’s how they did. This is inaccurate and may lead to potential examinee confusion.

In general, the students indicated they understood the information communicated in score report B, although one student remarked “not as well as A,” and another indicated “yes, except why are the colors of the bars important? Why less than 50%, 50-70% and greater than 70%.” As mentioned previously, the inclusion of these thresholds raised questions about whether 70% correct was considered passing, and if not, why the bars would be green.

Suggestions for enhancing the performance information communicated in score report B included the following:

• Reverse order of address/admin and performance data. Make results of interest more prominent.

• Feel as though the colors are confusing, don’t really need them as percent correct is shown. Probably would develop a different visual graph if wanting to include median scores, confidence intervals, where are the item #s?

• Include numbers of items not just percent. If Learning objective 4 is 2 items I will not worry as much as if it is 50 items.

• Include # of questions for section and # answered correctly. Info at a more focused level for incorrect answers.

• Move admin date, examinee ID and Address-info to top of report. Make note that not all objectives contain the same number of questions. Specify what X is in further detail (this admin, all admins, this test form).
• Status and my score should be presented closer together. I would want to know the total # of people taking the test. I would want to know the confidence bands surrounding my performance.

• Color is super! I would like the number of items in each category and the number correct in each category. Include raw score.

• Include information about how each section was weighted/counted toward total score. What are the points of breaking up the colored percentages this way – is yellow considered a minimum? Also, it would be helpful to know the number of items in each section.

• Total score and total %. N items/domain. Confidence bands? Use colors in charts?

• I think I would still want to know the # of questions representing each objective. The interpretation of scores may differ getting 36% right in 4 items as opposed to 12 items (more reliable). Maybe CI compensates for that?

In discussing score report B as a group, it was immediately pointed out that while clearly the examinee did not pass and learning objective 004 was the weakest, the number of questions was gone. Additionally, while the color was seen as eye-catching, the students did not immediately indicate they understood the red, yellow and green coding was meant to convey a similar message as a traffic light or dashboard. Most important, the thresholds of 50 and 70 percent correct were discussed. What was the real purpose of using these thresholds and having them represented using different colors? Would mixed signals be sent to an examinee and might they think that each learning objective could be passed individually on the test? The potential confusion between these thresholds and a
perceived minimum score needed to pass a “section” of the test was somewhat worrisome. Additionally, the question of whether these thresholds would vary from form to form made it clear that what was intended to be an arbitrary threshold was analyzed further than intended, and an unintended meaning was then associated with the various percentage correct thresholds.

Additionally, in discussing the median percent correct of passing examinees, some of the same questions asked in the testing professionals group were raised. Who were these examinees, how many were there, and when did they take the test? The description in the explanation page was too brief in describing this statistic. The point too was made that these numbers would be expected to vary across different forms of the same examination. If different forms had different average difficulties within each learning objective, then the median percent correct of passing examinees may vary slightly as well. It was suggested that perhaps a total test performance of median percent correct and median number of questions correct for the passing examinees be provided to take away from the individual learning objective emphasis. Students commented the typeface (font size) was a bit small for the score information listed at the top.

Overall while there was a positive reception to the use of color, there were enough questions about its utility that it was clear it may add to confusion or misinterpretation.

4.7.5 Score Report C Feedback

Following the discussion of score report B, score report C was presented to the students. Individual questionnaire responses showed that all students were able to correctly identify the administration date, the passing status, the best performing
objective, the learning objective with the widest confidence band, and that the score report represented only one examinee’s performance (although mention was made regarding the passing examinees group being present, with an unknown number of examinees represented in that group).

In response to the question of whether students could tell if one learning objective was weighted more or less than another, most students indicated no even though the number of items was presented on the report. Again, it had not been made clear on the explanatory page that each question had the same value, and that each learning objective had a different number of questions and therefore a different weight. Two students did indicate they could tell that the weighting was present.

In response to whether it was possible to tell the total percentage of items the examinee got correct, six students indicated no, while the four remaining indicated that you could calculate that from the information presented. For advice to the examinee, the students all indicated they would advise the examinee to study learning objective 004 prior to retesting, and two students indicated they would also advise learning objective 002 to be studied prior to retesting.

When asked about the confidence bands and the varying width between each one, all but one student gave a response. Most responses indicated the students drew the connection between confidence interval width, reliability, and measurement error. Some also mentioned that the width may relate to the number of questions (the greater the number of questions, the greater the reliability should be).

When asked to indicate whether the examinee’s performance was below, at, or above the median percent correct performance, responses were as follows:
Table 4.6. Distribution of Student Responses to Examinee Performance

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Performance in relation to Median % Correct of Passing Examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N responses – student focus group)</td>
</tr>
<tr>
<td></td>
<td>No response</td>
</tr>
<tr>
<td>001</td>
<td>0</td>
</tr>
<tr>
<td>002</td>
<td>0</td>
</tr>
<tr>
<td>003</td>
<td>0</td>
</tr>
<tr>
<td>004</td>
<td>0</td>
</tr>
</tbody>
</table>

As with the focus group of testing organization professionals, the term “expected” was inadvertently omitted, and therefore the question asked primarily looked at whether the focus group participants could identify where the examinee’s performance was in relation to the median percent correct (as with the same question on score report B). Recall that for the individual educators, the question was phrased as to whether the examinee’s “expected” performance was below, at, or above that of the median percent correct. Based on the student responses, there may have been some confusion for objectives 002 and 003, given the “expected performance would include (at) the level of the median percent correct of passing examinees. However, the actual performance for 002 was below the median, and 003 was above the median. Again, the omission of “expected” seemed a reasonable explanation as to why the group was split between designations for 002 and 003. One respondent clarified too that because the actual median percent correct was not provided numerically in score report C, that some responses with regards to below, at, and above may be inferences.

In general all students indicated they understood the performance information communicated in score report C, although comments were made about the pencil graph.
in particular “what do pencils represent? Hard to tell what percent they represent,” and “the pencils kind of remind the examinee of the objectives, but they are very distracting.”

Suggestions made by the students as potential enhancements to the performance information included in score report C are as follows:

- What is the purpose of pencil chart?
- Move the status of passing up with your score. Give the actual median score to accompany the X. Use a different visual for performance graph – back to score report A. Should have individual score on confidence interval graph. Make charts horizontal.
- In the chart with confidence bands, include the examinee’s score. Replace pencil graph with a line graph.
- Don’t like the pencils and pencil parts. Confidence bands may be confusing. More information about incorrect questions.
- Move status up with score and minimum passing score. Either explain the pencils or get rid of them, preferably get rid of them. Make note that confidence bands are explained on the back, maybe use an “*”.
- Take out the pencil graph, too confusing. I liked the chart from score report A better to convey this information. Likewise, I like the confidence band info but would like to see it presented horizontally and not vertically. I like the sandwiched table between your score and status info. Maybe switch the two. I would still want to know the # of people the median performance is based on.
- Put number correct in. The pencil graph is really bad.
• Add number of items answered correctly, not just percentages. The pencil graph is too hard to read.

• Clarification of Table 3 (confidence intervals), add total %.

• Total % of correct.

It was clear from the feedback received and the discussion that followed on score report C that while passing information and weak performance areas were clearly communicated, the “your performance” graph was not well-liked. In using the pencils, an appeal was made to present a performance graph that was less about precision, and more about visually conveying differences in performance using a symbol that might be welcomed by those who recoil from statistics. The questions raised by the students included an inability to accurately read what percentage correct was being communicated for each learning objective. “How much is an eraser worth?” was a question raised when seeing the graph. A suggestion was made to perhaps have a single pencil for each learning objective, with the sharpened tip being placed at the level of percent correct for each. The graph would then show four pencils, each of a different length.

In discussing the confidence intervals and median performance of passing examinees graph, students were not in complete agreement as to whether the confidence interval was important to include for examinees. More than one student indicated the graphs might be better understood if they were presented horizontally rather than vertically. Also, it was mentioned again that the “x” designating the median could be confused with the examinee’s own performance. It was suggested that some visual way of representing the examinee’s score in the center of the confidence band would be a better option for drawing attention to the examinee’s percent correct.
Another visual comment made on the score report was the space separating “your score” and “minimum passing scaled score” – it was suggested these be placed closer together. It was also suggested it might be better to move the passing status designation up to appear with “your score” rather than the bottom of the performance chart. Statistically, it was also requested that the number of questions correct and the total percent correct be provided on score report C.

4.7.6 Other Feedback

In discussing all three score reports with the doctoral students and the professor, it was clear that presenting more information was better, but a clarity in presentation was desired. A raw score and percent correct by learning objective and overall test were desired. Information about imprecision was important, but students were not unanimous in their recommendation of providing it on the score report. Score report C was not a favorite, but some aspects of score report B liked by educators interviewed (median percent correct performance, and color-coded performance thresholds) were questioned by the students.

4.7.7 Score Report D

Following the discussion of reports A, B, and C, score report D was presented to the students and the professor. It was explained that they were the first to see the report and provide feedback on it. Immediately score report D brought a lot of positive feedback, and further suggestions, but overall, it was generally well received. For purposes of the discussion, the participant group was walked through some of the
changes made, and rationales for each. There was no individual questionnaire, only a
group discussion.

The most visible change was noticed immediately; the report was presented in a
landscape visual style. Discussing the uppermost portion of the report, the reason was
given for keeping the administration date and examinee ID at the top, but comments were
made about somehow setting off the passing status and your score, by perhaps boxing
them and putting them together. It was explained why the learning objective titles were
presented first. Reading across the chart, an examinee would see the learning objective
name, specific detail on their performance, reference information (median percent correct
of passing examinees) and then the visual information of performance. In general, the
students liked this approach, and especially liked setting off the your performance
information in a box, with the raw information provided, and total test information
presented in addition to the learning objective level performance. It was suggested that
the median percent correct of passing examinees could be presented instead as a median
number correct.

The “Your Performance” bar graph also generated a number of comments. After
explaining why “x” and “M” were chosen to represent the two statistics of interest, there
was debate as to why “x” was necessary and whether examinees would understand their
score was at the end of the bar. The suggestion was made that in lieu of using “x” and
“M” a graphic could be used, such as an icon of a person or other visual. The concern
was raised that where “x” and “M” overlap it might be hard to discern what was intended.
The suggestion was made that perhaps a line through the M would further help to
distinguish it. Some suggestions were made to reword the key slightly, but overall the
bar graph was well received. It was mentioned to the students that the color and thresholds were removed from the bar graph, and there was no objection. Some students missed not seeing the confidence bands, and suggested ways they could be incorporated into the Your Performance bar graph using a box around the “x”, or by extending the bars to include the confidence bands, or even to include a smaller bar (above each learning objective bar) that would visually show the confidence band in relation to the overall performance. No comment was made about including an overall percent correct bar in the graph, but a question was raised as to whether an overall test confidence band might be useful to include.

The test design represented in score report D was well received. Students immediately identified that the pie chart was not related to the examinee’s individual performance, and that it was clear how the test was divided up. A note was made that the term “key” could be dropped from the explanation in the Test Design by learning objectives pie chart since there was a separate “key” explanation in the explanatory materials that referenced a different chart.

Overall the use of space in score report D seemed well received, although a student did ask whether the examinee information in the bottom right hand corner was filling space. It was explained that the placement was for in the event this report was to be physically mailed out to an examinee.

With regards to the explanatory page, it was explained to the students that the scaled score description and the population comprising the passing examinee group of was defined more clearly.
Students did not feel that the overall lack of color or graphic such as pencils caused the report to be less approachable. The only overall concern expressed was that three scales were used on the score report: scaled scores, raw scores, and percent correct scores. The question was asked as to whether each learning objective could be reported as a scaled score. While not explored fully within the group, this approach 1) reinforces the lack of a connection between raw performance and scaled score performance in communicating results, and 2) would potentially raise other examinee questions, especially if an examinee tried to average their objective scaled score across all four objectives and found it did not equal their total scaled score.
CHAPTER 5
DISCUSSION

5.1 Summary of Findings

This study aimed to explore some commonly misunderstood concepts in examinee score reporting for teacher certification. Through a careful design process, three sample score reports and explanatory pages were developed that incorporated various elements of design, graphics, statistical information, and layout. These score reports were presented individually to sixteen educators, each of whom gave personal feedback on his/her impressions, preferences, and comprehension of the score reports. A focus group was also conducted, convening six professionals from an educational testing organization. The feedback received from this focus group took into consideration issues of a psychometric nature, as well as feedback currently received from examinees through customer service inquiries and other examinee contacts with the organization.

Participant feedback was extensive, but commonalities were present, with distinct caveats. Common themes from educators, such as wanting to see the raw score information (number correct, number incorrect), and finding value in having a reference group against which to compare performance, were considered important. Also considered important was the notion that while measurement error may be a feature of test scores and should not be “hidden,” the participants in this study primarily thought examinees would likely not use measurement error and confidence interval information in preparing to retest.
Feedback from the educational testing professionals included some aspects not mentioned by educators. These aspects included the caution with which raw scores, comparative information, and the concept of measurement error are communicated to examinees. From the testing organization perspective there is a responsibility to report statistics that are meaningful and based on sound quality and quantity, while still providing information to alert an examinee clearly as to what their performance was. Two underlying common themes conveyed by both the educators and educator testing professionals were (1) no clear connection is communicated to examinees in how raw score performance information is related to scaled score performance, and (2) there is a need to direct examinees to materials and study resources available on the testing program and the content included on the test.

These commonalities and caveats were used to design a fourth model score report that was reviewed with doctoral students in educational measurement. The fourth score report used raw score reporting, visuals, and limited comparative information to help communicate examinee performance. Support materials were elaborated to include cautions raised by the educational testing professionals, and address potential misinterpretation by educators or examinees.

While many concerns and questions were addressed from the three score reports, in reviewing the final score report with the doctoral students, it is clear that there are still some considerations that should be made in connecting the measurement scales presented – raw score scale, percent correct scale, and scaled score.

To summarize, while there is no one “gold standard” of score reporting agreed upon unanimously by everyone, in general there is a desire to provide information where
possible at its most basic level (raw score and percent correct), and to accommodate
different comprehension styles by providing performance information in both a statistical
and visual way.

5.2 Conclusions

Through this study, a number of conclusions about score reporting in teacher
certification were made: (1) purely statistical terms and language may be ambiguous to
the reader; (2) narrative performance descriptions may be helpful, especially when
describing a difficult statistical concept such as measurement error; (3) use of color,
while visually pleasing, may be redundant and unnecessary; (4) use of graphs in
communicating performance information is helpful when properly labeled or tied in to
results; (5) confidence intervals were not immediately understood or seen as useful to the
examinee, while the performance of passing examinees provided an important contextual
framework; (6) raw score performance is desired when possible to provide; (7) scaled
scores need more explanation in how they are related to and derived from raw score
performance; and (8) supporting materials or resources should somehow be connected to
score reporting. Many of the conclusions stated could generalize to other areas where
performance information is communicated to stakeholders, in that application of these
conclusions will likely aid in interpretation.

Avoidance of purely statistical terms and language. One conclusion drawn from
this study is that where possible, pure statistical terms or abbreviations should be
avoided. For example, even the use of “N” to represent number was questioned.
Although the ability to interpret assessment statistics was mentioned as a pedagogical
foundation, that is those seeking teacher certification would have this background, a
reliance should not be placed on examinee’s ability to interpret these statistical concepts
such as error or confidence intervals. With explanation, these concepts become clearer,
but as a stand-alone figure they become ambiguous.

**Use of narratives may be helpful.** In a similar vein, mention was made by
participants as to whether narratives could be used to describe those statistical concepts
that might be important to communicate but not easily understood. The best use for this
may be in describing the confidence interval. For example, a notation such as “Please
note that if you tested again tomorrow with no additional preparation, your performance
on this learning objective could be expected to be between 51 and 64 percent correct.”
One potential problem with this, as pointed out by participants, is that by stating this you
may actually be encouraging examinees to retest without preparation, i.e., if someone is
just below the minimum passing scaled score, they may see this as an opportunity to pass
without any additional preparation.

**The specific use of color for this study was found to be redundant and
unnecessary.** The use of color in this study drew immediate reactions from participants.
As one respondent indicated, the use of color made the score report, “more attractive, but
you really had to read to understand what it meant.” With the student group,
interestingly, the color choice (red, yellow, and green) were not immediately seen as
intended in that the participants did not voice an understanding of the dashboard
 technique used, until it was discussed. Additionally, since color was used to express
performance ranges and thresholds, some saw the color as redundant given specific
percent correct was provided in the performance chart. Additionally, some educator
participants thought color would better be used to tie the specific learning objective in the performance chart to its counterpart in the bar graph. In the testing professionals focus group, the use of color raised more questions in that color choice was dependent upon performance, yet the performance thresholds were not related to the specific cutscore of that test or test form. Given the way in which color was implemented and the feedback obtained (both positive and negative), for purposes of this study color did not seem a necessary or worthwhile component to enhance comprehension in communicating results for teacher certification testing, and if used would need explicit instruction as to its meaning.

Use of graphs can be helpful, but type and location of graph should be tied to the purpose. As voiced by participants, some readers comprehend information better when presented visually, while others may ignore visuals and instead go directly to the statistics and performance data presented. However, when presented with a bar graph that reinforces and restates in a visual way information already presented, an added layer of potential comprehension error may be presented. Score Reports A, B, and C all made use of bar graphs to communicate percent correct, however the graph was not tied to the performance chart information which was presented first in all cases. When reviewing these score reports with participants, suggestions were made to tie together the graphical interpretation with the performance chart in order to better express that the graph was merely a restatement of the same information, not new information. Score Report D and the combined statistical performance chart with bar graph representation was well received and more quickly comprehensible. By lining up performance information and
vide, participants were readily understood that there was a relationship among the information.

Additionally, when multiple graphs were presented, if presented on the same visual line there was an impression that the two were somehow related or similar in importance. When presenting information independent of performance, such as test design, keeping that information separate from performance-related graphs or statistics aided in the reader’s interpretation. For communicating a concept such as test design, the pie chart was preferred over the cylinder, partly due to an incorrectly perceived hierarchy of the cylinder chart, and partly due to the ability of the pie chart to effectively communicate all the parts that make up the whole.

Provide performance of passing examinees instead of individual confidence intervals. Through the discussions with educators, students, and testing professionals while confidence intervals would be an additional piece of information for performance context, they do not seem to be necessary in communicating areas of attention where the examinee should study. Instead, the median percent correct of passing examinees seems to help guide the reader towards understanding that although performance may not be stellar in some areas, passing examinees also had difficulty in those same areas. Participants understood the passing information provided, but if using a reference group such as this, an explanation should be provided describing who the reference group is (e.g., administration date(s) represented in the data).

Provide raw score information where possible. Across the board, participants echoed the desire to see raw performance of the examinee – at a minimum the number of items in each area and the examinee percent correct should be provided. Since from this
the examinee could figure out the actual number of items correct and incorrect the question became why not provide that too? By providing this information, the examinee would have a clearer picture of their performance and potential areas of strength and weakness. One drawback to this, however, is if there is ever an error in scoring that needs to be corrected. By reporting raw scores, if the scorability status of an item changes, or if a scoring error is uncovered, the error may be amplified when raw scores are reported. Essentially, it is possible at times for a scoring change to occur and not affect a scaled score, but it would affect a raw score. By providing raw scores, a testing organization would have to embrace the change associated with reporting scorability or other.

**Explanation of scaled scores.** Another conclusion is that even when a basic level of information is provided on a score report, one clear barrier to understanding the total performance and why an examinee has not passed seems to be rooted in the conversion from raw performance to the scaled score performance. All participants understood raw score. They understood that if you get a certain number right out of a certain total number of questions, that you can calculate a percent correct. What all participants were unable to calculate was the scaled score. The major concern with this is that scaled scores began to be treated like raw scores. Participants divided the scaled score (180) by the total possible scaled score points (300) to state that was a percent correct. This is false and misleading.

Therefore, one suggested method of addressing this confusion (which is likely prevalent in any testing program using scaled scores) is to make public the raw to scaled score conversion. If the goal is for the recipient to understand how they did, but place it
on a common metric or scale, the best course may be to give them a key to understanding that common metric. Other large-scale testing programs include this information: MCAS in Massachusetts, the Texas Assessment of Knowledge and Skills (TAKS) in Texas – both student-testing programs. A drawback to this conclusion would be potentially every test form would require a unique raw to scaled score conversion table. While this is not impossible, it may further confuse examinees that do not understand that raw to scale score conversions are different from form to form depending on the difficulty of that form. An examinee may ask why they did not receive a different form and point out their raw score on their form would be passing on a different form of the same test. The College Board (2009) chose to report ranges for their raw to scaled score table, which may addresses some of these concerns, but could raise other questions as to precision and information shared.

Connect score reporting to available support materials and resources. One only needs to conduct an internet search for “exam preparation” and they will see a plethora of resources available – study guides, expanded study guides, practice tests with “authentic” test questions, etc. Examinees can get lost in navigating around these materials. While making sense of these resources is not necessarily the responsibility of the test publisher, there are helpful ways that the publisher could link performance as communicated on a score report to official materials available for the testing program. Whether support materials are free or fee-based, providing an examinee with a direct connection to these resources from the score report would help to focus any retaking preparations. For example, in Score Report A for this study, what if the score report issued was electronic, and the learning objective names were “clickable” and would take you to the public test
framework for that test field and that learning objective? In a way then, some score reporting categories (names of learning objectives, test design, performance of reference groups) could serve as a table of contents allowing the examinee to navigate where they need to go.

5.3 Limitations of the Study

No study is without limitations, and it appears that this study has at least four: (1) score report design; (2) participants; (3) methodology; and (4) extending the conclusions to other non-credentialing areas. Each of the limitations will be considered next.

5.3.1 Score Report Design

The designs of the score reports for this study contribute to some limitations in that they are not necessarily typical of all teacher certification tests. For example, the score reports reported on only four learning objectives. This may have made comprehension on the participant’s end easier. Some teacher certification tests may have 15 or more objectives or content competencies. When faced with a score report containing fifteen different areas of reporting, it may not be as easy to clearly distinguish strengths and weaknesses.

An additional feature of the score reports for this study is that the participant comprehension questions and presentation of information assumed all items counted for the same value and there was no penalty for guessing. If a performance component were added (e.g., an essay, spoken response, analysis of a case study) there may be weighting implications and implications to reporting percent correct scores. A single item could
then count for as many as, say, ten points, and may contribute as much as ten or twenty percent of a total test score, regardless of the full test length. Some student tests may make use of short answer items, or innovative items involving multiple tasks and points available, which may affect the ability to report meaningful performance in terms of raw score or percent correct scores. This added level of complexity could impact on the clarity of performance comprehension as communicated on the score report. Studies of these points could be carried out in the future.

Lastly, as pointed out by the doctoral student participants, the score reports used in this study were for a sample test where an examinee could not pass sections, but had to pass the test as a whole. A score report for tests where sections or subtests may be passed individually may add yet another level of interpretation on the examinee’s end, and may not be understood as clearly as providing a single scaled score for the test as a whole. Further investigation of this point seems desirable because there are a number of credentialing agencies where candidate pass/fail decisions are based on pass/fail performance at the subtest level on a number of subtests.

5.3.2 Participants

As with any survey, interview, or focus group activity it is preferable to have as much confidence as possible in the ratings and feedback obtained from participants. In selecting these samples of educators and assessment personnel it is understood that while the selection encompasses members of the relevant population (i.e., a selection of educators who are involved in interpreting individual examinee score reports, reviewing individual results, and offering instruction or formulating plans of instruction based on
testing results), the participants primarily are not the persons who will ultimately be the end user. The doctoral students used in this study are all students with measurement expertise. Because they have experience in measurement, statistics, and an understanding of scaled scores, they are not representative of all education students. Getting a larger group of students at differing points of their education career, and all seeking teacher certification, would represent a more focused group of feedback participants, and would also be the audience interpreting score reports for this purpose. It is not clear what the impact of using recipients of the score reports would have been on the findings, but it simply wasn’t possible to use them in this study—they simply weren’t available for the study. Clearly this was a shortcoming of the study that will not be repeated in any follow-up studies. They may be the best group for providing relevant information about score reports and report redesign.

5.3.3 Methodology

Because the methodology for this study was consistently applied across sixteen interviews and two focus groups, there is a question as to whether any order effect of the presentation of the reports may have been a factor in preferences and opinion, especially if a larger number of participants were involved. Were order effect important, it may have influenced some of the results from the study. In working with participants, Score Report A was always shown and discussed first, followed by B, then C. If there was any perceived “rank order” on the participant’s end, they may have incorrectly thought that A was intended to be the best, or vice-versa.
Additionally, given the web-based forum for the educator interviews, there was no opportunity for the educator participants to print out and mark-up or write on the score reports given to them. It is unknown whether this would have benefited the participants as all educators were willing to share commentary about the visual presentation and layout, but some participants in the two focus groups did make a point of conveying feedback by actually drawing or marking directly on the score reports.

5.3.4 Conclusions

The conclusions reached in this study are based on the focus and feedback being in the area of score reporting for a credentialing program, that is, more specifically, teacher certification testing. That’s not to say the conclusions are not applicable to other areas of educational testing and reporting of scores (e.g., conclusions regarding color, graphs, statistical language, use of narratives), however there are differences that should be noted.

Unlike student testing, the most important piece of information to convey to an examinee taking a credentialing examination is whether or not the examinee passed. Partly this is due to the criterion-referenced nature of licensing and certification. Given this purpose, the nature of score reporting in teacher certification is different than that used for student testing. In score reporting for student testing, it is critical to show what performance level a student has achieved, but there are typically multiple performance levels that can be achieved. Most state K–12 educational testing programs are not based on a single “Pass”/“Did Not Pass” criterion, but instead use multiple performance indices such as “Basic,” “Proficient,” and “Advanced.” While performance narratives may be
useful in teacher certification testing, a caution was ensuring narratives would not encourage retesting without some preparation. For student testing, this is less of an issue, as students test according to a pre-determined schedule, and retakers are typically only those students in the upper grades retaking in order to meet graduation requirements.

With regards to the conclusions of the need to clearly communicate how a scaled score is reached, and the request for guiding materials that direct students to further materials available to help in understanding the content being assessed there are some differences in application between student testing and credentialing. Some student testing programs already communicate this information: The Massachusetts Comprehensive Assessment System (MCAS) incorporates this feature into their student assessment program (2010a), as well as providing raw to scale score conversion tables for public consumption (2010b). Typically the population seeking teacher certification has a university background, and may have a better sense of measurement and assessment concepts than the population reading student score reports, who are parents. There may be comprehension issues faced with the parent audience that are not faced in the audience receiving and interpreting teacher certification score reports, therefore the support materials offered such as those for MCAS may be out of that need to help explain results to parents.

5.4 Directions for Further Research

This study provides some key points of score report comprehension and utility through the lens of credentialing tests. The teacher certification field would benefit from further studies that include education students, or actual examinees. By including education students, a more authentic response may be gained from respondents. Another
approach would be to survey examinees who are retaking a teacher certification test and inquire as to what they found as the most useful tool in remediation and guiding their studies. Is the score report being used when looking towards retaking the exam; is it tossed in a drawer, filed away, and of little to no value to the retaker; or does the utility of the score report lie somewhere in the middle? These are excellent questions for follow-up study.

Future research could also look at whether retakers who receive diagnostic feedback, such as subarea/learning objective-level reporting do better upon retesting than retaker examinees who only receive a total scaled score and no subarea information.

The conclusions and feedback received essentially point towards the decision that information beyond just a scaled score can be useful for examinees in preparing to retake an exam and understand areas needing attention, although too much information (color-coding performance, providing confidence intervals) may actually make the information beyond the examinee’s understanding. Given the variables previously stated (who is seeing the report, what type of learner are they) a testing organization may find it worthwhile to explore offering examinees the opportunity to “build their own” score report.

In a way, this type of service is currently available to institutions of higher education, or client states. State or University staff are able to directly import results into their own systems and essentially mine the data how they would like. There may be a customer service advantage in providing a service where an examinee could access their scores online and decide themselves whether to display or suppress certain statistics such as their raw score, the relevant median percent correct of passing examinees, and perhaps
the examinee’s own raw to scale score conversion table for their particular test form. By allowing examinees to pick the data themselves, an organization would provide a tool for examinees to see how the information may be presented and used, but decide for themselves whether they wish to use it. Explanatory materials would be dependent on the level and amount of information chosen by the examinee on his/her report. Targeted links to specific testing materials provided by the program could also be embedded as part of the report, given most of the materials are likely available online as well.

An expansion of this could include a faculty tool that by using the same interactive principles (and with the examinee’s permission) would allow a faculty member or guiding counselor at the educator preparation program to see different levels of reporting information, including confidence intervals. Essentially the score report could become audience-based and serve an individual’s needs whether they passed or failed, while providing as much psychometrically sound information as possible, and allowing the end user to make data driven decisions based on what is displayed.

By making use of a web-based reporting tool such as the one described, examinees could have the flexibility of choosing their options when assembling their score report, including options that may currently not even be within the examinee’s control, such as:

- Color coding learning objectives or major areas of content in order to see how information is connected among common content areas.
- Choosing to print the score report in portrait or landscape orientation.
- Allowing the examinee to move charts or graphs and where they are placed in the report, or suppress them entirely.
There would be a need for some consistency in the reporting, this is not to imply that reporting would be a “free-for-all.” Consistently placing examinee identifying information, scaled score information, and test name would likely be welcomed by those individuals working with the score report (so as to avoid having them hunt around for this information on every score report that is shown to them).

In summary, while the idea of a score report will likely be forever engraved in our mind as an official, printed document complete with watermark and perforations, the digital age presents an opportunity to offer customized, interactive reporting that best enhances examinees’ comprehension of their performance on an assessment - which is, after all, the fundamental reason for reporting scores.
APPENDIX: CONSENT FOR VOLUNTARY PARTICIPATION

MEMO

SCORE REPORTING IN TEACHER CERTIFICATION TESTING

CONSENT FOR VOLUNTARY PARTICIPATION

I volunteer to participate in this qualitative study conducted by Heather S. Klesch through the University of Massachusetts at Amherst, as supported by the Evaluation Systems group of Pearson, and understand that:

1. I will participate in focus groups led by Heather S. Klesch, which may include a guided discussion among participants, as well as specific questions regarding sample score reports for a teacher certification examinee.
2. The questions I will be answering address my interpretation and opinion of sample score reports for a teacher certification examinee. I understand that the primary purpose of this research is to gather feedback that may effectively contribute to score report design and considerations. The purpose of the research may be considered non-threatening and low-stakes for all participants.
3. The focus groups will be recorded to facilitate analysis of the findings.
4. My name will not be used, nor will I be identified personally, in any way or at any time. I understand it will be necessary to identify participants in the dissertation by position and focus group affiliation (e.g., a public school educator from the state of Florida); however, no names will be attached to specific comments that may be quoted.
5. I understand this study is not affiliated with the Florida Teacher Certification Examinations (FTCE), or National Evaluation Series™ (NES®) programs, nor is this study endorsed by the Florida Department of Education (FLDOE). Additionally, this study is not directed to the district, institution, or state from which I am affiliated.
6. I may withdraw from part or all of this study at any time.
7. I am free to participate or not to participate without prejudice.
8. I have the right to review material prior to the final oral exam or other publication.
9. I understand that findings from this study will be included in Heather S. Klesch's doctoral dissertation and may also be included in manuscripts submitted to professional journals for publication. Results published will be reported in aggregate form, and individuals will not be identified.

________________________  __________________________
Researcher’s Signature  Participant’s Signature

________________________  __________________________
Date  Date


guides: Review of current practices and suggestions for future research. *Applied
Measurement in Education, 17*(2), 145–220.

educational assessments. In R. P. Phelps (Ed.), *Defending standardized testing*
(pp. 91–110). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scales
and reports more understandable? In R. W. Lissitz & W. D. Schafer (Eds.),
*Assessment in educational reform* (pp. 192-205). Boston, MA: Allyn & Bacon.

understandable to policy makers and educators?* Paper presented at the meeting

Harris, D. J. (2003). Reporting and interpreting test results. In J. E. Wall & G. R. Walz
(Eds.), *Measuring Up: Assessment Issues for Teachers, Counselors, and
Administrators.* (pp. 53-64). Austin, TX: PRO-ED.


and Macmillan.

of Teacher Education, 11*(2), 147–172.

Massachusetts Department of Education. (2010a). Massachusetts Comprehensive
http://www.doe.mass.edu/mcas/testitems.html

Massachusetts Department of Education. (2010b). Massachusetts Comprehensive
Assessment System: Results. Accessed March 13, 2010 at
http://www.doe.mass.edu/mcas/results.html


interpreting your scores. Retrieved February 18, 2010 from


