August 2015

Empirical Investigation Of The Stakeholders' Understanding Of Information Contained In Score Reports For Large-Scale Testing

Stephen J. Jirka

University of Massachusetts - Amherst

EMPIRICAL INVESTIGATION OF THE STAKEHOLDERS' UNDERSTANDING OF INFORMATION CONTAINED IN SCORE REPORTS FOR LARGE-SCALE TESTING

A Dissertation Presented

by

STEPHEN J. JIRKA

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

DOCTOR OF EDUCATION

May 2015

College of Education

EMPIRICAL INVESTIGATION OF THE STAKEHOLDERS' UNDERSTANDING OF INFORMATION CONTAINED IN SCORE REPORTS FOR LARGE-SCALE TESTING

A Dissertation Presented

by

STEPHEN J. JIRKA

Approved as to style and content by:

_____
Ronald K. Hambleton, Chair

_____
Stephen G. Sireci, Member

_____
Aline Sayer, Member

_____
Christine B. McCormick, Dean
College of Education

ACKNOWLEDGMENTS

I would like to thank my advisor, Ronald K. Hambleton, for his many years of thoughtful guidance and support, and especially with his boundless patience. I would also like to extend my gratitude to the members of my committee, Stephen G. Sireci and Aline G. Sayer, for their helpful comments and suggestions on all stages of this project, even after it had taken longer to complete than originally intended.

I wish to express my appreciation to all the individuals who volunteered their participation in this project. A special thanks to my wife for her efforts in recruiting participants and all her additional support as the whole process was finally coming to completion. Finally, a big thank you to all those whose support and friendship helped me to stay focused on this project and who have provided me with the encouragement to keep on keeping on.

ABSTRACT

EMPIRICAL INVESTIGATION OF THE STAKEHOLDERS' UNDERSTANDING OF INFORMATION CONTAINED IN SCORE REPORTS FOR LARGE-SCALE TESTING

MAY 2015

STEPHEN J. JIRKA, B.S., TEXAS A&M UNIVERSITY

M.A., ST. MARY'S UNIVERSITY

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

Recent legislation (e.g. NCLB) has highlighted the need to administer more large-scale assessments to more grade levels and on more occasions. The results from these assessments are presented in score reports in various formats to many audiences, including teachers, students, school administrators, and parents. Because these assessments will be used as one of several evaluations of student achievement, the score reports should be designed and developed considering the needs of the recipients to ensure understanding. The purpose of this study was to demonstrate why this study is important and relevant, summarize some of the pertinent literature that pertains to the reporting of students' assessment results, and complete a project that examined the impact of various features of score reports on the effectiveness of communicating the results to various audiences.

Three separate studies contributed to the overall research by progressively providing information on the use and interpretation of information from score reports by different stakeholders. Study 1 focused on one-on-one interviews with a set of stakeholders to obtain more in-depth information about how persons are able to process the information they receive and are able to answer questions that they have before receiving the reports. It was also used as a pilot study for the other two studies. Study 2 used focus groups to offer feedback on several different types of data display used to convey test results. Focus groups consisting of stakeholders were shown a series of graphs that had been modified to either comply or not comply with several guidelines for data display that had been proposed by several researchers. Study 3 used a survey where sets of graphs were compared, one with original and the other with improved displays, and answers to questions based on the displays was given. A summative score based on dichotomously scored items was used to perform a t-test between the control and experimental group. While no significant difference was found, additional insight was gained into how various stakeholder groups used the information in reports and what they valued. Future studies will take into account the growing use of electronic distribution of assessment results.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

1.1 Background

Recent legislation (e.g. No Child Left Behind (NCLB)) (2002) has brought into the spotlight the need to administer more large-scale assessments to more grade levels and on more occasions. Additionally, the results from these assessments are indicators of students' performance in academic subject areas and are of interest to many audiences including educators, the general public, legislators, and state education officials. However, reporting test results to any stakeholder group is challenging because of the need to consider the density and accuracy of the information to be communicated to a particular group of stakeholders (Jorgensen, 2005; Ryan, 2006).

The main audiences for these state and national assessment results are educators and state education personnel, including teachers, school administrators, and parents (Ryan, 2006). The students themselves may also be interested. Keeping in mind that these assessments will be used as one of several evaluations of student achievement, the score reports should be designed and developed considering the above individuals. Furthermore, both the use and understanding of these reports should be kept in mind.

The true value of score reports lies in the match between the report, the intended audience, and the anticipated use. As noted earlier, score reports are prepared for parents, teachers, administrators, state officials, and of course the students themselves (Ryan, 2006). A

score report prepared for one particular audience may be helpful for them, but not for a different audience, and so the purposes must be kept in mind at the design stage. In general, student score reports have two broad purposes. The first is informing about a student's progress, the efficacy of instruction, and the curriculum. The second is providing information needed by local, state, and national officials for accountability programs. All of these purposes and audiences must be kept in mind when designing and developing score reports and a reporting system in general (Goodman & Hambleton, 2004; Ryan, 2006).

1.2 Statement of the Problem and Its Significance

Because of the previously mentioned recent legislation, the results of large-scale testing have become more important, and the ability to effectively disseminate and communicate these results to a variety of audiences has become more important, accordingly. Examining what the current best practices are and what improvements may be made must also be an important aspect of any testing program. This leads to the problem and issue that were addressed in this study. After all of the time and effort has been put into designing, administering, and statistically analyzing these very important large-scale assessments, an equal amount of time must be put into properly disseminating the results to the diverse stakeholders. Unfortunately, this is not always the case. As a result, reports are often seen as an opportunity to present as much information in as small a space as possible (Goodman & Hambleton, 2004). Not much thought appears to have been put into exactly what information is needed (or conversely not needed) by specific groups and how that information can best be presented though this situation appears to be

rapidly improving. This study addressed this important issue through the use of a comparison study.

1.3 Purpose of the Study

The purpose of this study was to explore different methods for displaying test data (scores) that will be utilized by various stakeholders in large-scale testing by carrying out three studies: a study involving one-on-one interviews, a study utilizing focus groups, and a study using a survey for comparisons. In short, these studies formed the foundation of this dissertation and aimed to confirm or expand existing guidelines. The study investigated two questions: (1) How are the stakeholders for score reports able to glean the information they need from these reports and apply it for the intended purposes? (2) What new guidelines, if any, can results from the current study for the reporting of large-scale assessment results?

1.4 Outline of the Study

What follows in Chapter 3 is a description of three studies to investigate how effective the various elements of score reports are for conveying the necessary information to the various stakeholder groups. In Chapter 2 a review of the existing literature in this area, including relevant standards for reporting assessment results, studies addressing score reporting for large-scale assessment results, national assessment results, ways of augmenting displays with additional information, and some summary studies, are provided.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, a review of previous research concerning the visual display of information from assessment results is provided. The chapter begins with a review of the relevant standards that pertain to the reporting of assessment results. Second, several important studies on reporting large-scale assessment results are described before describing relevant studies on reporting national assessment results. Third, summary reviews and guidelines for score reporting are described before ways of augmenting displays with additional information are described. Finally, the major findings gleaned from the review of the literature are summarized, and the link between the purposes of the study and the literature review is addressed.

As addressed earlier, recent legislation (e.g. NCLB) (2002) has brought into the spotlight the need to administer more large-scale assessments to more grade levels and more frequently. The results from these assessments are indicators of students' performance in academic subject areas and are of interest to many audiences including educators, the general public, legislators, and state education officials. However, reporting test results to any stakeholder group is challenging because of the need to consider the density and accuracy of the information to be communicated (Jorgensen, 2005; Ryan, 2006).

The main audiences of these state and national assessment results are educators and state education personnel, including teachers, school administrators, and parents (Ryan, 2006). Keeping in mind that these assessments will be used as one of several evaluations of student achievement, the score reports should be designed and developed considering the above individuals. Both the use and understanding of these reports should be kept in mind.

Again, the true value of score reports lies in the match between the report, the intended audience, and the anticipated use. As noted earlier, student score reports are prepared for parents, teachers, administrators, state officials, and of course the students themselves. A score report prepared for one particular audience may be helpful for them, but not necessarily for a different audience. In general, score reports have two broad purposes. The first is informing about a student's progress, the efficacy of instruction, and the curriculum. The second is providing information needed by local, state, and national officials for accountability programs. All of these purposes and audiences must be kept in mind when designing and developing score reports and a reporting system in general (Goodman & Hambleton, 2004; Ryan, 2006).

The purpose of this chapter is to summarize some of the relevant literature that pertains to the reporting of students' assessment results. To be more specific, this chapter explores the meaning and the medium of score reports in detail, attempts to summarize what has been learned about the benefits and disadvantages of including or excluding certain

content, and examines the impact of various formats on the effectiveness of communicating to various audiences by describing relevant studies.

This literature review is composed of four parts. It first discusses the relevant standards that should be kept in mind when designing and implementing score reporting. Next, studies on reporting of large-scale assessment results are presented, followed by studies on reporting national assessment results such as NAEP (National Assessment of Education Progress). Next, studies describing ways of augmenting score reports with additional information are presented. Finally, several very good summary studies that include important guidelines are described before some general conclusions are drawn.

2.2 Standards Related to Educational and Psychological Testing

The *Standards for Educational and Psychological Testing* (AERA et al., 1999) and updated and expanded in 2014 (AERA et al., 2014) is the major document related to all aspects of testing and addresses standards related to test scoring and reporting of students' performance on the tests. This document is the primary source of professional and technical standards that guide most aspects of testing. At least thirteen standards contained in *The Standards* are relevant to reporting student scores and are listed below:

2.2.1 Standard 6.10

When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used.

2.2.2 Standard 6.12

When group-level information is obtained by aggregating the results of partial tests taken by individuals, evidence of validity and reliability/precision should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals without appropriate evidence to support the interpretations for intended uses.

2.2.3 Standard 6.14

Organizations that maintain individually identifiable test score information should develop a clear set of policy guidelines on the duration of retention of an individual's records and on the availability and use over time of such data for research or other purposes. The policy should be documented and available to the test taker. Test users should maintain appropriate data security, which should include administrative, technical, and physical protections.

2.2.4 Standard 6.15

When individual test data are retained, both the test protocol and any written report should also be preserved in some form.

2.2.5 Standard 6.16

Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores and pertinent ancillary information.

2.2.6 Standard 12.1

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described by those who mandate the tests. It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test developer and/or user.

2.2.7 Standard 12.8

When test results contribute substantially to decisions about student promotion or graduation, evidence should be provided that students have had an opportunity to learn the content and skills measured by the test.

2.2.8 Standard 12.9

Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on alternative forms of the test or be provided with technically sound alternatives to demonstrate mastery of the same skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow students to obtain the relevant instructional experiences.

2.2.9 Standard 12.10

In educational settings, a decision or characterization that will have major impact on a student should take into consideration not just scores from a single test but other relevant information.

2.2.10 Standard 12.11

When difference or growth scores are used for individual students, such scores should be clearly defined, and evidence of their validity, reliability/precision, and fairness should be reported.

2.2.11 Standard 12.13

When test scores are intended to be used as part of the process for making decisions about educational placement, promotion, implementation of individualized educational programs, or provision of services for English language learners, then empirical evidence documenting the relationship among particular test scores, the instructional programs, and desired student outcomes should be provided. When adequate empirical evidence is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the students.

2.2.12 Standard 12.15

Those responsible for educational testing programs should take appropriate steps to verify that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified.

2.2.13 Standard 12.18

In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.

There are several common themes that can be gathered from the above standards. To begin with, it is the responsibility of those who develop and administer the assessments and who produce the score reports to ensure that there is valid interpretation and use of the test results. Moreover, valid interpretations of the results are shared with the different audiences that are given the results. Secondly, great care and caution should be taken when the data are presented in groups. The validity, comparability, and reliability must be established, and, if not, cautionary statements must be included. Finally, it is especially important to carefully consider the technical merits of reporting gain scores, since recent legislation (e.g. NAEP) makes this standard very important. While these gain scores are not reported for individual students, they are reported to show Adequate Yearly Progress (AYP), which is a main component of NCLB. Several of the above guidelines (especially 5.10 and 13.14) will be explored more in-depth in Chapter 3 when the methodology of this study is presented.

Searches on websites and in the technical literature seem to suggest that test publishers haven't explicitly stated that they follow specific guidelines given by the above standards as they report their test results. No reference is given to the specific guidelines, but the test publishers state somewhere that they follow general guidelines. According to Thomas Brooks who has been involved in work like this at Harcourt Assessment (personal communication, October 15, 2007), inclusion of details such as error bands was done to adhere to *The Standards* (AERA et al., 1999, 2014). It is hard to evaluate how well the publishers and states are following these guidelines, but after looking at several newer example reports from different testing organizations (e.g. Pearson Educational

Measurement), they seem to be making an attempt to follow them by including more contextual information, error bands, and highlighting results. Allalouf (2007) had some additional guidelines for reporting as a part of the whole assessment process.

2.3 Studies on Reporting Large-Scale Assessment Results

There have been several studies that have examined the interpretations of state-wide or large-scale assessment results. These studies concentrated on student-level results, typically using focus groups. The importance of the use of focus groups will become evident throughout this dissertation.

An early study in this area was done by Impara, Divine, Bruce, Liverman, and Gay (1991). Impara and his colleagues investigated the extent that teachers in one state were able to interpret student-level results on a standardized state assessment and to what extent interpretive information provided on the reverse side of the student score report helped improve teacher understanding. While many teachers provided reasonable interpretations of information contained on the score reports, some types of information were misunderstood by large numbers of teachers. As noted by Impara, "areas of weakness related to scale and normal curve equivalents (NCE) scores; the percentile band performance profile, interpreting grade-equivalent scores; and the norm-group number correct on the skills chart; which provides the average number correct by the national norm group and the number correct by student." Unfortunately, regardless of the availability of interpretive information, most of the teachers in the study could not properly interpret percentile band performance profiles, something that is commonly provided in score reports (Goodman & Hambleton, 2004).

This study also noted that including interpretative information helped address many, but not all of the difficulties teachers had in interpreting the other scores. Problems still remained even when interpretive material was provided. For example, even with interpretive material, teachers still did not understand the meaning of percentile bands that overlapped. In addition to recommending more research on teachers' understanding of student score reports, Impara and his colleagues suggested that some problems in score interpretation might disappear if score reports contained only instructionally relevant information. For instance, they recommended removing rarely used scores such as the NCE to make the reports less intimidating for teachers and parents. Again, it is often just as important to know what to leave out as what to leave in, and this will be described more in the Results and Discussion chapters.

Written before the more widespread use of assessment, the National Educational Governance Panel (NEGP) (1998) provided a good source of information for states on how they could report statewide and individual student assessment results in better ways. This report recommended keeping four questions in mind in order to achieve a balance between providing too much and too little information:

1. How did my student do?
2. What types of skills or knowledge does his or her performance reflect?
3. How did my child perform in comparison to other students in the school, district, state, and if available, the nation?
4. What can I do to help my child improve?

The NEGP (1998) suggested providing an interpretive scoring guide to help answer these questions and to give parents information that will not be able to be included on the report themselves. NEGP did emphasize that parents should be informed prior to assessments, something that is not always done, though in recent years parents more often are.

Other ideas provided by NEGP (1998) to help states report results included encouraging parents to contact their child's teacher for information about their test results, encouraging parents to ask questions about the particular educational policies, emphasizing the importance of looking at different sources of information, and providing examples of student work and questions that illustrate what students know and should be able to do.

As has become typical of these types of studies and the present study, information gathered by a small focus group was also included in this report. Eleven parents were shown score reports from several commercial test publishers and asked to make comments. The parents liked to have explanations of what the scores meant included in the reports, such as subtest scores and a description of what skills were assessed by the test, and they liked inclusion of what can be done to help improve their child's scores. Things the parents did not like included reports that were too technical, included too much jargon, and had fonts that were too small and hard to read.

As described in Ryan (2006), Forte Fast and Tucker (2001) presented a four stage process that was used while reviewing and redesigning the reporting system for the Connecticut Mastery Test program (CMT). The first stage was a review of the current state reports and the relevant state and federal reporting requirements. Stage two extended this review to other states. Stage three was the use of focus groups around the state that included teachers, administrators, and parents. Information from all of the above sources was used to redesign the student, classroom, school, district, and state level reports during stage four.

An important aspect of this study that can be applied to future studies regarding score reports was the use of focus groups composed of different stakeholders. Briefly, a focus group is a form of research in which a group of people are asked about his or her attitude towards something, whether this is a product, service, concept, or idea. Questions are asked in an interactive group setting (typically 8-10 members) where participants are free to talk with other group members (Fern, 2001; Krueger & Casey, 2000). The facilitator's typical role is to elicit comments and solicit questions and information, not to sway panelists one way or another. As it was employed in the current research project, focus group methodology will be described in more detail in Chapter 3, but the point to keep in mind here is that all of the intended audiences of the particular types of reports were asked to respond to questions concerning the reports they were likely to read. Remember, the expectations of different groups must be considered when examining the effectiveness of different score reports, as the needs will differ significantly. The parents looked at the individual student level reports, while the teachers and administrators

reviewed the classroom and district level reports. Higher level administrators looked at the school, district, and statewide reports.

Again, the focus group methodology was a very important aspect of this study. There is demonstrated value in asking intended audiences about reports they are most likely to use. The different groups involved in this study had many suggestions about the reports and involved the following categories of report design features:

- Format features
- Graphical displays
- Numerical displays
- Normative information
- Detail and specificity
- Support materials
- Glossary of terms
- Directions to supplementary information
- Easily reproduced materials.

2.4 Studies on Reporting National Assessment Results

Hambleton and Slater's (1995) research on the National Assessment of Educational Progress (NAEP) executive summary reports further highlighted the problems policy makers and educators have in understanding reports of assessment. These researchers proposed to find out exactly what the intended users of NAEP reports understood and were able to get out of the reports that they looked at. As reported by Hambleton (1997,

2002), this study gathered together administrators and policy-makers to look at some of the reports produced by NAEP and discuss what they liked, did not like, and what they understood from the reports and displays. Once again we see the use of a focus group that presented a number of data displays to the panelists and asked them a series of questions to gauge their understanding of these displays. Additional questions probed for the panelists' opinions of the displays and how they might be improved.

Despite the fact that most of the sample had one or more statistics or testing courses in their background, many interviewees had forgotten a lot of the statistical and measurement information they had known at one time, if they had known at all. Problems were found with the statistical jargon or terminology used (e.g., proficient, statistically significant, standard error, cutpoint, scale score), the construction of tables (e.g., cumulative columns exceeding 100%, too much detail, statistical symbols like greater than and less than signs, ordering of material, footnotes), and the design of graphs (e.g., over-complexity, novelty) (Hambleton & Slater, 1997).

Wainer, Hambleton, and Meara (1999) followed up the work of the Hambleton and Slater (1995) study and also extended some of the work by Wainer (1996, 1997a, 1997b). This study sought to take some displays from the current reporting of NAEP that were problematic, revise them according to some of the emerging data-display guidelines, and to administer them both to educators and policy-makers.

In this study, questions were generated that were believed to be the sorts of questions that policymakers and educators might want to answer. Some examples of these types of questions might be as follows:

1. What was the general direction of results between 1992 and 1994?
2. Which region showed the greatest decline in performance for the 12th graders from 1992 to 1994?
3. In 1994, which region of the country had the lowest average reading proficiency at all three grade levels?
4. What is the ranking of the regions from best to worst in terms of average reading proficiency for grade 12 in 1994?
5. Which of the four regions is most typical of the U.S. results?
6. Which regions for 8th graders in 1994 performed better than the average for all of the United States?

The answers to these questions given by the panelists when using both the problematic and updated displays were the central interest of the study. Additionally, times to respond were tracked for some questions as a possible factor of interest. The results of this study were very interesting. For some displays, the revision made little difference. For others, the responses were often faster and more accurate for the revised displays. Finally, it should be noted that the preferences of the panelists were not associated with the accuracy of interpretations.

Wainer and his colleagues came up with the following two conclusions. First, policy makers and educators had considerable difficulty with a number of the original score report displays. Secondly, many of the problems associated with the data displays might be overcome with more careful consideration during the design phase and with field tests to identify strengths and weaknesses.

In an earlier study with a similar respondent profile, Henry (1993) found that graphic displays were a reasonable option for information-rich displays in that the level of understanding, though poorer than table-based information, was still reasonable. He reasoned that practice with graphical forms would enhance ease of understanding and argued that comparison information must be in close proximity to ensure accuracy of comprehension. Another important graphical principle supported by Henry's findings was the importance of making sure empty space is not used to communicate information. According to him, gaps do not necessarily convey quantities accurately to the viewer (Henry, 1993).

Two recent studies involving focus groups looking at NAEP data displays were undertaken by researchers from the University of Massachusetts. The first study (Zenisky et al. 2006a, Zenisky et al. 2009) focused on reading test results, and its purpose was to evaluate the use and understanding of several of the common strategies for communicating NAEP results using a focus group of state reading content specialists to discuss several common displays used to communicate results from the reading assessment. The group of eight Reading Specialists from multiple states reviewed fifteen

NAEP data displays from recent reports, the NAEP Question Tool, and the NAEP Data Explorer and discussed their understandings and impressions. Some of the findings included the need for clarification in footnotes, legends, and keys, as well as simplification of some displays to minimize clutter. In addition, participants sought additional information about the practical meaning of some displays, especially when statistically significant test results were presented. The authors suggested two possible future directions for research, including the use of one-on-one conversations with users of the NAEP website (often called "think-aloud" studies) about selected data displays and using suggestions from the focus group to redesign some displays for tryout with focus groups. These directions were taken into account for the current research project.

The second study (Zenisky et al. 2006b, Zenisky et al. 2009) focused on results from the mathematics test from NAEP. As part of a larger evaluation of the utility of NAEP score reports, this focus group composed of mathematics curriculum leaders from across the United States was held to explore the extent to which different NAEP data displays had meaning and usefulness. The authors noted that the most important finding was that even educators with quantitative skills experienced some difficulty with many of the common NAEP score reports. The focus group made several suggestions for revising the layout of several data displays, particularly with respect to footnotes and arrangement of keys/legends within figures. It is worth pointing out that this study again demonstrated the advantages of the focus group methodology for gaining insights about score reports and the importance of either revising the NAEP score reports to make them more user-friendly and/or the need for more explanatory materials for persons using these reports.

The authors cautioned that a focus group of eight is not a sufficient basis for initiating major report changes, but it does suggest the need for substantially more research.

Overall, the findings from Zenisky and her colleagues illuminated specific aspects of the score reports that might best be revisited in a more experimental setting. Particularly, how do stakeholders with different levels of quantitative skills interpret and utilize the information in score reports differently? Can the arrangement of keys and legends in the graphs truly help or hinder the information the different stakeholders garner from the reports? Does inclusion of explanatory materials such as a glossary or other interpretive materials help? Can a reduction in the "clutter" of displays also help in the interpretation and understanding of these displays? These questions will be addressed by manipulation of these particular aspects in the present study. Details on the study are offered in the next chapter.

2.5 Augmenting Displays with Additional Information

Several methods for providing additional, contextual (collateral) information to the display of basic performance scores have been proposed and studied over the years. It is hoped that by providing this information, additional insight can be gained into the examinee's performance. Several of these methods will be described here.

2.5.1 Market-Basket Approach

In 1998, Mislevy described an innovative method for score reporting called the market-basket approach. The idea originated from the use of market basket reporting used to explain economic changes over time as reflected by the consumer price index. The price

of a basket of food (with known and fixed grocery items) is reported each month to provide the public with a single, easy to understand measure of economic change. In an educational setting there could be a collection of test items and performance task items that measure important educational outcomes. The collection of assessment material would reflect diverse item formats, difficulty levels, cognitive levels within a subject area, and any other dimension of interest. By reporting the performance of a national sample of students on the market basket of items each year, the quality of education might be monitored over time using a single, clear index that can be easily understood. Via the use of IRT and statistical equating it would not even be necessary to reuse the specific test items in the "market basket." Similar items could be used.

The market basket items would be clearly explained to the public and policy makers to enhance the meaning of such statements as "in 2000, the average American fourth grader obtained 37 out of 50 points on the assessment. This is 3 points higher than the results reported in 1999." Research using focus groups, such as that described earlier in this review, points out the difficulty some stakeholders may have with this statement. For example, there is no indication of confidence bands. Additionally, performance standards (e.g. basic, proficient, advanced) could be mapped onto the test score scale associated with the market basket assessment using test characteristic curves, and this would be more meaningful to many audience members.

However, according to Hambleton (2002) there are some problems to overcome in utilizing the market basket approach to score reporting. The first concern is whether the

items are reported or released to the public. These items could no longer be used in future tests, since the students might be taught the items and perform better. For the market basket to work, an equivalent set of items must be included in each assessment. However, this can be a difficult and expensive task for states. The second problem that might be encountered is that released items or tasks might cause the curriculum to be narrowed. States or teachers might try to focus in on items similar to those contained in the market basket, since they know these are the ones that will count and will be on the tests.

2.5.2 Diagnostic Scores

Another method for using collateral information to supplement score reporting is the use of so called diagnostic scores. These are using subscores of a larger test to infer information about a specific set of skills. For example, in a test of English language arts, we may want to know specifically about the examinee's reading comprehension ability. There may be a small subset of items that measure the examinees ability. However, when we use a smaller number of items, the reliability may be lower. Several methods involving prior distributions have been developed to try to overcome these limitations (Wainer, 2001; Luecht, 2003) and many have been promising (Haberman, 2008; Sinharay, 2009; Haberman & Sinharay, 2010; Roberts & Gierl, 2010).

2.5.3 Use of Collateral Information

Several methods have been developed to stabilize these subtest scores through the use of collateral information, and, in 1987, Yen developed a procedure to combine information from other portions of the test. Specifically, her procedure combined information from

the responses to subsets of items representing a specific educational objective with the score on the test as a whole, to produce estimates of the true score for the subset of items. That procedure formed the basis of the objective performance index (OPI) reported for some tests published by CTB/McGraw-Hill. Yen's procedure is based on assumed binomial distributions for the proportion correct scores. Additionally, her procedure treated the rest of the test as a unit in subscore estimation. Other methods are similar to Yen's but may use the normal distribution or may distinguish among any of the other subtest scores in the estimation of any one of these subtest scores (Wainer, 2001).

An example of this type of use of collateral information is the way the National Assessment of Educational Progress is carried out. The NAEP procedures calibrate the items on the exam and then make use of collateral information to estimate all subscores in one large iterative estimation system. Unlike Yen's procedure, this collateral information is not based on test or item performance. This method is more fully described in Mislevy et al. (1992).

2.5.4 Item Mapping

One final method of augmenting the test scores with additional information is usually referred to as item mapping (Zwick, Senturk, Wang, & Cooper-Loomis, 2001; Ryan, 2006; Wang, 2003). Using item response theory, the items are ordered in difficulty from easiest to hardest. This method has been used with the three-parameter model with NAEP, but can also be used with the one-parameter model with only the difficulty parameter. The scale can be described in terms of the content of the items that are "mapped" by ordering the items based on their difficulties. For example, the ability to

perform addition can be broken down into performing this with 1-digit or 2-digit numbers, and these can be mapped to more or less difficult items along the scale. The abilities of students are also mapped along this same scale, so we can determine what we might think a student with a particular score might know or be able to do. In the previous example, a student with a higher score would more likely be able to perform addition of 2-digit numbers, while a student with a lower score would not (Ryan, 2006; Wang, 2003).

Various individual studies that have looked at the reporting of large-scale assessment data, as well as the results of NAEP have been described up to this point. Ways of utilizing auxiliary information by different techniques have also been described. In the following section, several good summary studies will be discussed.

2.6 Summary Reviews and Guidelines for Score Reporting

Goodman & Hambleton (2004) produced a literature review that investigated current approaches for reporting student-level results on large-scale assessments. Student test score reports and interpretive guides from fourteen U.S. states, two Canadian provinces, and three commercial testing companies were examined. Based on past score reporting research, testing standards, and requirements of NCLB, a number of promising and potentially problematic features of these reports and guides were identified, and recommendations were offered to help enhance future score reporting designs and to inform future research in this important area.

This review study listed several important recommendations:

- Student score reports should be clear, concise, and visually attractive.

- They should also include easy-to-read text that supports and improves the interpretation of charts and tables.
- Care should be taken to not to try to do too much with a data display (i.e. displays would be designed to satisfy a small number of pre-established purposes).
- Devices such as boxes and graphics should be used to highlight main findings.
- Data should be grouped in meaningful ways.
- Small font, footnotes, and statistical jargon should be avoided.
- Key terms should be defined, preferably within a glossary (where they can be easily located by users).
- Reports should be piloted with members of the intended audience.
- Consideration should be given to the creation of specially-designed reports that cater to the needs of different users (i.e. a detailed score report may be appropriate for teachers, but a simpler report may be more appropriate for widespread distribution to parents).

Seven additional recommendations by Goodman and Hambleton (2004) are:

- Include all information essential to proper interpretation of assessment results in student score reports.
- Include detailed information about the assessment and score results in a separate interpretive guide, ideally one in which the student score report can be inserted.
- Personalize the student score reports and interpretive guides.
- 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 guides that illustrate the types of achievement represented by each performance level.
- Include a reproduction of student score reports in the interpretive guides to clearly explain the various elements of the reports.

An excellent resource written for both state educational agencies as well as local educational agencies to help them to be able to effectively report their accountability data is *A Guide to Effective Accountability Reporting* by Forte Fast (2002). It begins by helping the reader understand what the requirements of NCLB are and other relevant legislation as it pertains to reporting to comply with the legislation and gives very practical advice on the design of these reports. It deals with the composition of the team that will design these reports and who needs to be involved and tells exactly what specific information needs to be included in the reports.

According to Forte Fast (2002), the guide was not intended to provide an academic discussion of the nature of indicators and indicator systems, nor is it meant to cover the broad territory of accountability issues. It was meant to provide a resource for agencies and to spur the thought of practitioners, as accountability reporting systems are tooled to meet the requirements of NCLB. While each state or local agency will have different approaches and starting points for the development of accountability reports and different end products, it is a good idea for all agencies to *clearly document* whatever processes are used in report development. This will provide evidence that reports were developed in a

sound manner, help to ensure the process is comprehensive in addressing all relevant reporting requirements, and facilitate subsequent redevelopment plans.

Finally, the report described specifics on how the information needs to be displayed, and a list of additional resources is presented at the end of the report to pursue these points in further detail. A few example graphs are presented and then compared to an improved graph, while provided a critique of the displays. As will be described in Chapter 3, original and improved graphs were used in the current study.

Commissioned by the National Council on Measurement in Education (NCME), Deng and Yoo (2009) completed an annotated bibliography that has been a good source for additional examples of score reports used in different states and agencies as well as summaries of additional papers not able to be mentioned here. As will be mentioned next, looking at a variety of report examples is a beneficial aspect of creating score reports that score report builders should utilize.

Recently, Zenisky and Hambleton (2012a, 2012b) have proposed a model of score report development after integrating previous research. This model is defined by seven guiding principles:

1. Establish the purpose or purposes for the score report (sometimes this step may include a needs assessment)
2. Identify the intended audience
3. Review existing reports for ideas

4. Develop prototype reports
5. Field test reports
6. Revise and redesign, if necessary
7. Ongoing maintenance

Carrying out a needs assessment of what the report needs to accomplish is the first step in the score development process. This should include defining the purpose of the report(s) and consulting with key stakeholders. Second, score report builders need to carefully consider who will be viewing and/or using the score report, and, third, review existing report samples for ideas. Many examples of score reports are now available on the web. After the first three steps, prototypes of the score reports can be created and reviewed internally by the score report builders. Hambleton and Zenisky (2012b) created a 32 question check list that takes into account research findings and covers multiple report element areas. Fifth, field testing of these prototype reports and collecting data combining the stakeholders' opinions and understanding of potential repots. As Zenisky and Hambleton noted, this use of field testing is a common activity used throughout the test development process and can yield valuable information. Sixth, revising and redesigning these score reports should be carried out incorporating the field test data. More than one round of field testing and revision may be necessary. The final, seventh step is that a program of monitoring and maintaining score reports be carried out. Reports should be useful and function as intended, and this last step takes this into account.

2.7 Summary and Conclusions

This review of the literature on score reporting has attempted to present the key studies that pertain to this important and understudied area of testing and measurement. This review reveals the following: (1) It is clear that much of the focus of the research up to this point has been done on either NAEP or on large-scale or state assessments. (2) Although there has been a lot of research in this area, it is evident that there is not one simple summary of features and formats that make score reports informative and meaningful to all the intended audiences. Goodman and Hambleton (2004) summarized many features of score reports and is an excellent resource on the different strengths and weakness that should be considered when conceptualizing and designing score reports for a testing program. The more recent work by Zenisky and Hambleton (2012a; 2012b) has expanded much on the work up to now and offer a synthetic process of steps and principles for building score reports. Also, there exist several good sources dealing with the graphical display and reporting of statistical information (see for example Henry (1995), Nicol and Pexman (1999), or Morgan, Reichert, & Harrison, 2002). Remember that different audiences have slightly different purposes in reading the reports and therefore need slightly different pieces of information from these reports. (3) Many empirical studies have been carried out, suggestions and recommendations have been made, but problems still remain as revealed by most of these studies. (4) Improvements could be made in the design of these types of empirical studies to better reveal the complexity of the factors that have an impact on effective score reports.

This review of the literature pointed out that more research in this area is needed. Moreover, a more scientific approach using a quasi-experimental design should be used in a K-12 setting. This will carry on previous research done with national assessments (e.g. NAEP) into another important area of assessment (state) that has become more into the forefront.

CHAPTER 3

METHOD

3.1 Introduction

The primary goal of this study was to explore different methods for displaying test data (scores) that will be utilized by various stakeholders in large-scale testing, and in this chapter, the methodology for three studies is presented. Each study is presented separately, with the specific design, sampling method, and data analysis presented briefly for each. A more detailed section on the analysis of the data is next, followed by a summary section on the methodology presented in this chapter.

This research project consisted of three studies: Study 1 focused on one-on-one interviews with individuals from important stakeholder groups to obtain in-depth information about how the person is able to process the information he or she receives. In addition, participants were asked to give examples of some of the questions that they had before receiving reports. This study was used primarily as a pilot study to refine the questions and methods that were used in Studies 2 and 3 and helped to identify relevant issues or problems in a smaller more intimate setting before turning to the larger group settings. Study 2 used a focus group to provide feedback on various types of data displays. A focus group consisting of representative stakeholders was presented a series of graphs that had been modified either to comply or not to comply with several guidelines for data display that had been proposed by several previous researchers. Study 3 consisted of a survey administered to different sets of stakeholders that encounter data

reports in different degrees of frequency and who typically require different types of information from these reports.

3.2 Study 1: Think Aloud Study

Study 1 was a smaller scale study that involved one-on-one sessions with selected stakeholders from the respective groups. Two apiece from each of the following stakeholders were asked to answer a series of questions as they examined a series of score reports on a laptop for approximately half an hour: parents, teachers, students, and administrators. As many members per stakeholder group were interviewed as resources allowed. The goals of this study were (1) to preliminarily examine the general process that the participants go through as they attempt to retrieve the information needed from the various score reports and (2) to refine the questions and logistics of the methods to be used for the studies that will involve large groups.

The invited study participants were selected from an available pool by email, phone, and in person. The pool came come from employees of an assessment organization, customers and employees of a local business, educators and students from a local school district, and persons known by the researcher. Participants were briefed about the nature of the study and asked to participate for one time only. Ten to fifteen closed and open-ended questions were asked that ranged from asking the participant to find a particular piece of information from the set of displays, to their opinion on certain aspects of the data displays. Example questions are as follows: What types of questions do you have about how your student did on this type of test? Tell me out loud what is going on in your mind as you begin to look over these example reports to gather information and try

to answer these questions. These interview questions were based on previous research in this area of data displays and on previous literature on the use of interviews for research (NCES, 2002; Willis, 1999, 2005). This research technique has been used by used by national assessment entities and for-profit testing organizations to develop page layout, answer documents, and gain further insight into how students answer the test questions themselves (Zucker, et al., 2004).

These sessions were audio recorded using a small recorder with the permission of the participants to have a more accurate record of the sessions and notes were taken by the interviewer. Information gathered from these one-on-one sessions was analyzed by looking at the notes and listening to the recordings from the interviews and looking for common themes to be further expanded in Studies 2 and 3. For example, were there any particular pieces of information the participants typically seek and did they indicate any problems with the example reports. Since this study primarily served as a pilot study for the next two studies, aspects, such as particular questions or tasks, changed from interview to interview in order to further refine the process. There was an initial set of questions used that was augmented in the subsequent interviews based on more experience with the interviewing process. For this phase of the study, it was not necessary that all aspects of the interviews were completely standardized.

3.3 Study 2: Focus Group

This next phase of the overall research study involved gathering together a panel of stakeholders from different subgroups that may require different types of information from their respective areas. This study was carried out as in the spirit of a single-group

design, where the participants were shown two different types of displays that claim to impart the same information that is commonly reported in a large-scale testing situation. The focus groups were shown one version of the graphs that were to be the "original" graphs that do not comply with some of the current guidelines that have been developed for displaying visual information better. These guidelines came from several different sources, including previous studies on score reporting (e.g. Impara et al., 1991, Goodman & Hambleton, 2004) and literature on the graphical display of information in general (e.g. Henry, 1999). The focus groups also viewed other versions of the graphs that were "improved" by editing these original graphs to comply with the guidelines for the visual display of information. Numerous graphs were displayed to the participants, half being in the original format and half in the improved format, each group shown all of the displays. Some of the factors included were:

1. placement of text within the display
2. "density" of the graph
3. inclusion/exclusion of highlights of results
4. confidence bands
5. inclusion of narrative information at beginning of display

These factors were chosen because they seemed to be described in the literature and previous studies as most pertinent (e.g. Goodman & Hambleton, 2004; Zenisky et al, 2006a, 2006b). The participants answered ten closed-ended questions, as well as some open-ended questions to gather general opinions from the participants. Both questions about knowledge and understanding and questions to measure attitudes from the

participants were used. An example of a preset question and an open-ended question follow: Which classroom scored the highest in both Mathematics and Reading on the following score report? What do you like or think is useful in this score report? The participants were asked to respond both in a brief written format in a form provided to them and aloud so that the facilitator was able to ask follow-up questions for clarification. An emphasis was placed on inviting the participants to ask questions and discuss the material in an open and collegial manner. Transcripts of these sessions are included in Appendix B.

Eight to ten participants were included in each of the focus groups, and two focus groups (panels) were convened to try to determine the consistency of the responses. The groups convened for approximately one hour and were run according to established guidelines for focus groups used for marketing and product research, adopting them for this particular setting (Fern, 2001), and previous focus groups that analyzed results of a national assessment (Zenisky et al., 2006a, 2006b) were taken into account for what was asked and included in these groups. It was planned that the stakeholder groups represented would include policymakers, administrators, teachers, and students to try to get as much diversity as possible in order to bring to the discussion as wide a variety of views as possible. However, the final composition of the panels was determined by the availability of participants from the available pool of candidates, so representatives of the student stakeholder group could not be directly represented in the focus groups themselves. Participants' responses to questions and the overall discussion were examined to determine if any patterns were present. The main purpose of Study 2 was to

confirm the findings or extend the findings of previous research in this area on what elements are important to keep in mind when displaying graphical information (e.g. Goodman and Hambleton and Wainer et al.'s work) and the exploration of any additional guidelines. Basically, the goal was to determine what participants are paying and not paying attention to. Moreover, the results from Study 2 informed Study 3.

3.4 Study 3: Questionnaire

Study 3 was similar to Study 2, but involved giving surveys to a larger sample of different groups of stakeholders to react to a similar series of graphs (see Appendix A). The questionnaire and graphs were delivered to participants through a variety of methods with as much of a brief oral introduction to the survey as possible. Some were handed out to small groups, some were mailed, and some were handed out individually. A small pilot study involving a select sample of experts in research methodology and not reported here was done to refine the process on a small sample before the large sample was included.

Similar to Study 2, one version of the graphs was the original graphs that did not comply with some of the guidelines that have been developed for better visual display of information. These guidelines came from several different sources, including previous studies on score reporting and research on the display of graphical information in general. The other version of the graphs were improved by editing these original graphs to comply with the guidelines for the visual display of information and some changed to comply with additional aspects that might affect the ability of the participants to gather information. Four graphs were displayed to the participants in either the original format

or the improved format, and participants were exposed to both formats within the questionnaire. It had been planned to show more graphs to the participants, but after taking into account the amount of time the pilot sample was taking to go through the survey, one of the graphs had to be dropped and the number of questions decreased. As presented in Study 2, the factors manipulated included:

1. placement of text within the display
2. "density" of the graph
3. inclusion/exclusion of highlights of results
4. confidence bands
5. inclusion of narrative information at the beginning of the display

To make the study as strong and generalizable as possible, systematic assignment of the control versus experimental displays was included in the study where alternating Form A and B versions of the questionnaire were handed or sent out. The participants answered preset questions and responded to a written survey. The questions were written to either be scored dichotomously or the respondent indicated the degree to which they agreed or disagreed to a series of statements. The two sets of surveys and the corresponding questions are found in Appendix A, with an example introduction to be placed at the beginning of the questionnaire as well. Survey A was considered the control and survey B was considered the experimental version. To encourage participation and collection of the physical surveys, respondents were provided a postage-paid envelope to mail the completed survey back to the researcher. Respondents were also given the choice to send an email to the study email address to add their name to a lottery to win a small prize

(one of three gift cards) as an incentive. Because of a larger volume of participants than the first two studies, the survey utilized more selected response type items and fewer open ended type items.

The main emphasis of Study 3 was to determine what respective stakeholders wanted from score reports and to see if changing aspects of the displays affected the ability to gather needed information. What type of information do they need and are they able to get it from these exemplar score reports? Since the sample included teachers, parents, and even students who have taken the particular score reports, the results were expected to be more generalizable. The intent was to make the sample as representative as possible by balancing stakeholder and demographic variables. The composition of the sample is reported in the next chapter.

3.4.1 Development of the Survey

The purpose of this survey was to learn about the different ways stakeholders are able to effectively use and understand information contained in score reports. While one main goal was to keep the experiment version brief to encourage maximum return rate, it was necessary to try to get as much information as possible. In order to accomplish this, questions were developed to gather some general demographic information and then presented a series of displays that showed original and modified displays.

To help evaluate and improve the quality of the survey used for the comparison, a pilot test was carried out in the spring of 2009. Five individuals selected for their expertise in experimental design, psychometrics, and statistics served as reviewers to help maintain

the quality of the survey. Individuals from different testing organizations with extensive experience in the design and carrying out of research studies were recruited. Additionally, faculty members from a large research university were also asked to look at the displays and questions. All reviewers were asked to answer the survey under as realistic conditions as possible and then give any general as well as specific feedback. After this process, the survey was then able to be administered to the larger sample. Additionally, information gathered from studies 1 and 2 that was expected to improve the questionnaires was incorporated.

The questionnaire consisted of two parts. The first part contained several items designed to obtain demographic information about the respondents. As noted earlier, two versions of the questionnaire (A & B) were distributed to teachers, high school students, parents, and administrators. The information included the educational level of the respondents, the work environment of the respondent, the population of the local school district, the number of years of teaching experience, exposure and type of training with testing and measurement, the usefulness of the training, and ethnicity.

The second part of the questionnaire consisted of a mixture of Likert-type items and true/false items. For each Likert-type item, the respondents were asked to circle on a scale the number that most closely corresponded to their opinion on that item. A five-point scale was used for each item, where a value of "1" indicated the least favorable response and a value of "5" indicated the most favorable response. For example, an average response of less than "2" is indicative of a largely negative position on the item.

The true/false items indicated how the respondents felt about the series of statements dealing with the factual nature of the graphs. The choice "Neither" indicating neither true nor false was also provided. Additionally, there were items that were traditional multiple-choice that had a stem question and then four options for the respondent to choose from based on the graphical display that had been given to them. These were intended to check if the respondents were able to understand the displays and gather the necessary information intended to be projected from the graphics.

3.4.2 Sample Selection

In order to gather as large a sample size as possible, to be able to access the different stakeholders, and to obtain a sample as representative as possible, a comprehensive plan was initiated to accomplish these goals. Candidates were recruited from several different locations that individuals from the stakeholders' groups are known to frequent. Possible locations included, but were not limited to, testing organizations with a large sample of former teachers and parents of students taking large-scale assessments, local businesses that serve diverse populations, schools that would allow access to educators and students, and a network of individuals familiar with the researcher. Copies of the actual surveys were sent out to these locations and a recruitment flyer that includes the email address of the survey along with brief information about the study was also given. Follow-up activities were carried out to increase the response rate, and the inclusion of the opportunity to participate in a lottery for a small monetary compensation was added as an incentive.

3.4.3 Administration of the survey

As noted earlier, the survey was administered by distributing by mail and in person hard copies of the survey to participants after a brief oral introduction if possible. Packets of surveys were also sent out to helpers in local (Texas) and remote locations (Washington and New Jersey) for distribution of the surveys. Respondents mailed back the surveys to make collection as streamlined as possible, protect anonymity, and take into account geographic location.

3.5 Data Analysis

A variety of analytic strategies were employed to explore and interpret the data for this study. Descriptive statistics were computed to characterize the sample, including frequency distributions for every background question. Means and standard deviations were calculated for responses to the general opinion questions, the questions specific to the certain score reports, and the interpretation questions for score reports. Patterns of the responses to the general opinion questions were examined for any unusually strong opinions either in favor or in opposition to the proposed purposes for score reports. For the questions pertaining to specific score reports, response patterns were examined, as were comparisons between the respondents; opinions regarding one display of information versus another. For this analysis, a t-test was applied to the differences in means for each set of two corresponding purpose questions and significant differences ($p < .05$) were reported. Similar analyses were carried out too at the item level. These results are found in the next chapter.

For the opinion questions, the degree of the respondents' agreement with each interpretation is reported as a mean ranging potentially from 1 (strong disagreement) to 5 (strong agreement), and the overall degree of respondents' agreement with these interpretations were compared between the different types of information displayed. For the interpretive questions, questions were scored dichotomously with a 1 for correct responses and a 0 for incorrect responses.

Since responses to the interpretive questions in comparing the difference ways of displaying the data, the background variables that related to training in testing and measurement issues will be of further interest in this study. Correlations between these background variables and interpretive questions were calculated and examined.

3.6 Summary

The three separate studies described above contributed to the overall research study by providing progressively more information on the use and interpretation of information from score reports by several different stakeholders. Moreover, these studies contributed to the overall goal of exploring different methods for displaying test data (scores) that is utilized by various stakeholders in large-scale testing.

CHAPTER 4

RESULTS

4.1 Introduction

This research project consisted of three studies. The first focused on one-on-one interviews with individuals from important stakeholder groups to obtain in-depth information about how the person is able to process the information he or she receives. Study 2 used two focus groups to provide feedback on various types of data display, and Study 3 consisted of a survey administered to different sets of stakeholders that encounter data reports in different degrees of frequency and who typically require different types of information from these reports. The results are present below.

4.2 Study 1: One-on-One Interviews

Parents, teachers, students, and administrators were asked a series of questions as they examined a series of score reports on a laptop for approximately half an hour. The goals of this pilot study were to preliminarily examine the general process participants go through as they attempted to retrieve the information needed from the various score reports and to refine the questions and logistics of the methods to be used for the studies that will involve large groups.

Selected by email, by phone, and in person from an available pool of employees of an assessment organization, customers and employees of a local business, educators and students from a local school district, and persons familiar with the researcher, participants

were briefed about the nature of the study and asked to participate. Interviewees were asked questions aimed to ask them to find a particular piece of information from the set of displays and to ascertain their opinion on certain aspects of the data displays. Some of the questions were as follows: What types of questions do you have about how your student did on this type of test? Tell me out loud what is going on in your mind as you begin to look over these example reports to gather information and try to answer these questions.

Audio recordings of these sessions were carried out with permission of the participants to have a more accurate record of the sessions and to supplement notes taken by the interviewer. Information gathered from these one-on-one sessions was analyzed by noting data from the interviews and looking for common themes to be further expanded in Studies 2 and 3. Since this study primarily served as a pilot study for the next two studies, aspects, such as particular questions or tasks, differed slightly from interview to interview in order to further refine the process.

4.3 Study 2: Focus Groups

Two panels of stakeholders, taken from different subgroups that may require different types of information from their respective areas, were convened for this second study. The participants were shown different types of displays aimed to impart the same information that is commonly reported in a large-scale testing situation. Each focus group was shown one version of the graphs that were original graphs that do not comply with some of the current guidelines that have been developed for better displaying visual information. They also viewed other versions of the graphs that were “improved” by

editing these original graphs to comply with the guidelines for the visual display of information. Numerous graphs were displayed to the participants—approximately half being in the original format and half in the improved format. These graphs reflected both individual student type reports as well as group type reports.

The participants were asked several open- and close-ended questions to gather general opinions from the participants. Both questions about knowledge and understanding and questions to measure attitudes from the participants were used. Some questions used during the focus groups are as follows: Which classroom scored the highest in both Mathematics and Reading on the following score report? What do you like or think is useful in this score report? The participants were asked to respond both in a brief written format, in a form provided to them, and aloud so that the facilitator was able to ask follow-up questions for clarification. An emphasis was placed on inviting the participants to ask questions and discuss the material in a collegial manner and elicit as much information and understanding how the panelists process the graphs as possible.

Six to eight participants were included in each of the two focus groups, and each convened for approximately one hour. Each was run according to established guidelines for focus groups used for marketing and product research, adopting them for this particular setting. Recruited from the performance scoring center of a testing organization, the two panels included administrators, teachers, and individuals who were recently students, and an attempt was made to achieve as much diversity as possible in order to bring to the discussion as wide a variety of views as possible. Unfortunately,

high school students were not able to participate in these focus groups but were included into the next phase of the research.

Transcripts of the focus group sessions were analyzed both in terms of whether the discussions with the participants agree with previous research on guidelines for what to include in score reports and how this study can inform the next study using questionnaires. Aligning with previous guidelines, the participants liked and understood some types of scores more than others. For example, one member didn't think that the scale scores were useful but admitted not knowing much about them. Percentile ranks were favored and somewhat more familiar, while the ability/achievement comparison (AAC) had not been seen before by participants. Grade equivalents were useful even if a little bit of explanation was needed. Some participants thought some reports were a little too complicated and might need to be simplified for some audiences. Including both graphical and textual explanations in reports was liked by several participants, along with comparisons on how an individual student did compared to a larger group such as school or nation. One participant made a point that echoes one of the rationale for this research: tell the different groups (e.g. parents, administrators) what they need to know—different audiences require different types of information. Additionally, including diagnostic type information so a parent or teacher knows what to tutor the student on was good to include in the reports. All of these points and more were taken into account for developing the questionnaires used in Study 3 reported next.

4.4 Study 3: Questionnaire

Study 3, and the most important one, used a questionnaire provided to the participants, with two versions (control and experimental) distributed to teachers, high school students, parents, and administrators. There were two parts to the questionnaire. The first part contained several items designed to obtain demographic information about the respondents such as the educational level of the respondents, the work environment of the respondent, the population of the local school district, the number of years of teaching experience, exposure and type of training with testing and measurement, the usefulness of the training, and ethnicity.

The second part of the questionnaire consisted of a mixture of Likert-type items and true/false items. For each Likert-type item, the respondents were asked to circle on a scale the number that most closely corresponded to their opinion on that item using a five-point scale. Additionally, there were items that were traditional multiple-choice that had a stem question and then four options for the respondent to choose from based on the graphical display that had been given to them. These were intended to check if the respondents were able to understand the displays and gather the necessary information intended to be projected from the graphics. They were also used as the basis for the summative score comparison between the control and experimental groups.

4.5 Description of the Sample for the Questionnaire

In order to be able to describe the participants, the survey included nine questions about the subject's local community, experiences, and training in educational testing. These demographic questions will be described in the following sections.

4.5.1 Numbers Returned

Approximately 250 questionnaires were distributed over the course of two months to gather as many respondents as possible. Out of this number, 110 were returned filled out and ready to be coded and entered into the database to be analyzed, giving a response return ratio of 44%.

4.5.2 Capacity Survey Filled Out

Respondents were asked in what capacity they completed the questionnaire. If the respondents had multiple roles, they were asked to indicate their primary one. The last category of Administrator includes Principals, Department Heads, Curriculum Directors, and Counselors.

Table 4.1 Capacity Survey Filled Out

Option	Count	Percentage of Sample
Parent	40	36
Teacher	35	31
Student	26	24
Administrator	9	8

In the table above it can be seen that most respondents were parents, followed by teachers, and then students. The category with the lowest number of respondents was Administrator. The percentages broken out by form are below:

Table 4.2 Capacity Survey Filled Out by Form

Option	Form A Percentage of Sample	Form B Percentage of Sample
Parent	31	41
Teacher	29	33
Student	31	17
Administrator	10	7

4.5.3 Community of Respondents

Respondents were asked to describe the community in which the school they were involved in is located. Out of the three choices of suburban, urban, and rural, suburban was chosen the most often at 72%.

Table 4.3 Community of Respondents

Option	Count	Percentage of Sample
Suburban	78	72
Urban	17	16
Rural	10	9
Other	3	3

The percentages of community broken out by form are below:

Table 4.4 Community of Respondents by Form

Option	Form A Percentage of Sample	Form B Percentage of Sample
Suburban	67	77
Urban	17	14
Rural	12	7
Other	4	2

4.5.4 Years of Experience or Involvement

The number of years of teaching or involvement had a slightly different interpretation depending on if the respondent was a teacher/administrator, parent, or student. Two separate questions were used to gather this information and are highlighted in the tables below. The first was for those filling out the survey as a teacher or administrator, and the second was for those filling out the survey as a parent or student.

Table 4.5 Years of Teaching and/or Administrative Experience

Option	Count	Percentage of Sample
Fewer than 3 years	4	4
4 to 9 years	11	12
10 to 20 years	10	11
More than 20 years	20	21
Does not apply to me	50	53

Table 4.6 Years of Involvement in the Local School System

Option	Count	Percentage of Sample
0 to 5 years	13	15
6 to 8 years	7	8
9 to 12 years	16	19
More than 12 years	21	24
Does not apply to me	29	34

Table 4.7 Years of Teaching and/or Administrative Experience by Form

Option	Form A Percentage of Sample	Form B Percentage of Sample
Fewer than 3 years	5	4
4 to 9 years	7	16
10 to 20 years	11	10
More than 20 years	23	20
Does not apply to me	54	51

Table 4.8 Years of Involvement in the Local School System by Form

Option	Form A Percentage of Sample	Form B Percentage of Sample
0 to 5 years	12	18
6 to 8 years	0	16
9 to 12 years	24	13
More than 12 years	22	27
Does not apply to me	41	27

For the overall sample, the largest percentage of teachers had more than 20 years of teaching experience, followed by 4 to 9 years. For parents or students, most had more than 12 years of involvement in the local system. The next highest percentages were 9 to 12 years and then 0 to 5 years.

4.5.5 Testing and Measurement Training

Most of the respondents indicated they had some exposure to the concepts that would be presented in a course on testing and measurement theory. These were the teachers or administrators. Parents and students were not asked if they had any exposure to the materials or courses. Exposure was typically either through a seminar or self-taught, for those respondents who indicated they had some exposure. Not many respondents had taken a formal course during teacher training in the subject area. Of those who did receive training, most had taken a workshop sponsored by the school or district. A course at college or university was the second most chosen option for type of training.

Table 4.9 Received Training in Educational Testing

Option	Count	Percentage of Sample
Yes	39	35
No	71	65

Table 4.10 Type of Training

Option	Count	Percentage of Sample
Workshop sponsored by school/district	22	38
Workshop while attending conference	7	12
Course at college or university	18	31
Book or online training	5	9
Other	6	10

**For types of training, there were multiple marks and these were adjusted for the counts for each type.*

Respondents that gave examples of "Other" included the following: Assessment Specialist training, class by teacher, instructor staff development course for military institution assignment, all of the above, research on my own online, and teacher in service.

4.5.6 Usefulness of Training

The respondents who filled out the survey as a teacher or administrator were also asked how well they thought their training or education in testing and measurement had prepared them to deal with, understand, and use the results of large-scale assessments that were used in their school and district. According to the counts and percentages in the table below, most respondents found the training "Somewhat Useful" or "Generally Useful". Fewer respondents were on the two extremes indicating the training was very useful or not useful at all.

Table 4.11 Usefulness of Training

Option	Count	Percentage of Sample
Very useful	4	12
Generally useful	9	27
Somewhat useful	12	36
Rarely useful	4	12
Not useful at all	4	12

4.5.7 Current Level of Education

Table 4.12 Current Level of Education

Option	Count	Percentage of Sample
Some high school	28	25
High School diploma/GED	7	6
Associates Degree	2	2
Bachelors degree	38	35
Post bachelors degree (Masters, Doctorate)	35	32

The current level of education of the respondents ranged from some high school to having a post bachelors degree. The question should be interpreted differently depending on if the respondent was a teacher/administrator, parent, or student. The two most frequently chosen options were “Bachelors degree” and “Post bachelors degree” (67% total). The levels of education broken out by form are in the table below:

Table 4.13 Current Level of Education by Form

Option	Percentage of Form A	Percentage of Form B
Some high school	33	19
High School diploma/GED	2	10
Associates Degree	2	2
Bachelors degree	31	38
Post bachelors degree (Masters, Doctorate)	33	31

4.5.8 Racial/Ethnic Background

The categories of the sample were 71% white and 29% combined for the other categories. The second highest category was “Asian/Pacific Islander”.

Table 4.14 Racial/Ethnic Background

Option	Count	Percentage of Sample
American Indian /Alaskan Native	1	1
Asian/Pacific Islander	20	18
Black, Non-Hispanic	1	1
Hispanic	10	9
White, Non-Hispanic	77	71

Broken out by form, the racial/ethnic percentages are seen below:

Table 4.15 Racial/Ethnic Background by Form

Option	Percentage of Form A	Percentage of Form B
American Indian /Alaskan Native	2	0
Asian/Pacific Islander	17	19
Black, Non-Hispanic	2	0
Hispanic	12	7
White, Non-Hispanic	67	74

4.5.9 Summary of the Sample

Overall, the sample reflected a wide distribution of demographic groups, sorted by education, school system experience or involvement, urbanicity, and ethnic background. In some areas such as ethnicity it was not as varied as was hoped for in the study.

4.6 Control Group Reports (Form A)

First, counts and percentages of each of the responses for the questions contained on the Control Form (Form A) will be given. The question, followed by a table for each question will be given. Within each table, the correct answer will be indicated by *italics* and have an asterisk. These will be presented for the three different score displays that were presented to the respondents. The original questionnaire and corresponding graphical displays can be found in Appendix A.

The first display of information was based on Report #1 for the fictional student named Sally taking a national standardized achievement test. Respondents were asked to look at the information and then answer the eight questions that follow.

10. According to the report, Sally is above average in Total Math compared to other 2nd graders in the nation.

Table 4.16 Question #10

Option	Count	Percentage of Form A
*True**	25	48
False	24	46
Neither	3	5

It can be seen that an almost equal number of survey takers gave a "True" or "False" response to this question. A small number of respondents felt the statement was neither true nor false.

11. According to the report, Sally got 82 percent of the questions correct on the Total Reading subtest.

Table 4.17 Question #11

Option	Count	Percentage of Form A
True	15	29
*False**	35	67
Neither	2	4

Most of the survey takers (67%) felt that this statement about Sally's performance on the test was false. Almost 30% felt that the statement was true, and a small percentage felt the statement was neither true nor false.

12. According to the report, Sally scored significantly better in Word Study Skills and Reading Vocabulary than Reading Comprehension.

Table 4.18 Question #12

Option	Count	Percentage of Form A
True	13	25
*False**	31	60
Neither	8	15

Only 25% of the survey takers thought the statement about Sally's score on Word Study Skills versus Reading Vocabulary was true. Sixty percent of the respondents thought it was false, while a slightly larger number than usual felt the statement was neither true nor false.

13. According to the report, Sally's test score in Spelling is above that for students with the same ability level.

Table 4.19 Question #13

Option	Count	Percentage of Form A
*True**	45	87
False	3	6
Neither	4	8

The vast majority of respondents thought the statement above was true (87%). An almost equal, much smaller number, either thought the statement was false or neither true nor false.

The next two questions had the same statements in order to help determine if respondents were carefully reading and going over the questions or were randomly marking their answers.

14. According to the report, Sally has the same proficiency of word study skills and knowledge as a third grader.

Table 4.20 Question #14

Option	Count	Percentage of Form A
True	40	80
*False**	6	12
Neither	4	8

Eighty percent of the survey takers felt the above statement was false, versus 12% who thought it was false, and 8% who thought it was neither true nor false.

15. According to the report, Sally has the same proficiency of word study skills and knowledge as a third grader.

Table 4.21 Question #15

Option	Count	Percentage of Form A
True	37	74
*False**	9	18
Neither	4	8

Almost one-fifth of the respondents thought the statement above was false, while almost three-fourths agreed with the statement above. A small number thought the statement was neither true nor false.

16. Assuming no learning occurs, if a thousand students with the same ability as Sally repeatedly took the Total Reading test, their PR scores will:

Table 4.22 Question #16

Option	Count	Percentage of Form A
Always be equal to 59	20	40
*Vary between 40 and 70**	19	38
Vary between 40 and 59	5	10
Vary between 59 and 70	6	12

An almost equal number of respondents chose that the percentile rank score would always be equal to 59 and that the score would vary between 40 and 70. Smaller numbers of respondents chose the final two options on this question.

17. This score report helps me understand Sally's strengths and weaknesses.

Table 4.23 Question #17

Option	Count	Percentage of Form A
Strongly Disagree	1	2
Disagree	10	19
Neutral	10	19
Agree	28	54
Strongly Agree	3	6

Slightly more than half of the survey takers thought this report helped them understand Sally's strengths and weaknesses. Slightly less than one half either thought the report did not help them understand or were neutral about it.

The next set of questions dealt with Report #2 that displayed information for a student by the name of Ferrus taking a national standardized achievement test. Respondents looked at the information and answered the eight questions that follow.

18. According to the report, Ferrus is above average in Total Math compared to other 11th graders in the nation.

Table 4.24 Question #18

Option	Count	Percentage of Form A
*True**	29	56
False	20	38
Neither	3	6

For this question, slightly more than half of the respondents chose "True" and 38% chose "False". A much smaller number chose "Neither", indicating the statement was neither true nor false.

19. According to the report, Ferrus got 82 percent of the questions correct on the Total Reading subtest.

Table 4.25 Question #19

Option	Count	Percentage of Form A
True	13	25
*False**	39	75
Neither	0	0

Respondents to this question either chose “True” or “False”, with none chose “Neither”.

Three-fourths chose “False” and one-fourth chose “True”.

20. According to the report, Ferrus scored significantly better in Math Procedures than Math Problem Solving.

Table 4.26 Question #20

Option	Count	Percentage of Form A
True	41	79
*False**	11	21
Neither	0	0

Seventy-nine percent of the survey takers agreed with the statement above concerning

Ferrus’ scores in mathematics. Twenty-one percent chose “False” and disagreed.

21. According to the report, Ferrus’ test score in Social Science is above that for students with the same ability level.

Table 4.27 Question #21

Option	Count	Percentage of Form A
True	8	15
*False**	42	81
Neither	2	4

The majority of respondents did not agree with the statement above and checked “False”

on the survey. About one-sixth chose “True” and agreed with the statement.

22. According to the report, Ferrus’ scale score in Math Problem Solving was about the same as the average score for a student in the first month of 12th grade had they taken this test.

Table 4.28 Question #22

Option	Count	Percentage of Form A
True	9	17
*False**	39	75
Neither	4	8

Seventy-five percent of the survey takers disagreed with the statement above and chose "False". Seventeen percent agreed and chose "True" as his or her answer.

23. If we want to see how Ferrus' school did compared to another school by average student scores in Math, Percentile Ranks would be the most appropriate score to use.

Table 4.29 Question #23

Option	Count	Percentage of Form A
True	37	71
*False**	10	19
Neither	5	10

The majority of survey takers agreed with the above statement and chose "True". Slightly less than one-fifth chose "False" and did not agree with the statement about Ferrus' classmates' performance.

24. You will notice on the report that Ferrus had a GE of 12.1 in Math Procedures on the report. What does this score of 12.1 mean?

Table 4.30 Question #24

Option	Count	Percentage of Form A
He should immediately be moved from the 11th grade and placed in a 12th grade Mathematics class.	2	4
*His scale score in Math Procedures is about the same as the average score for a student in the first month of 12th grade had they taken the test.**	40	77
He has mastered 11th grade Mathematics and can perform 12th grade work in Mathematics.	3	6
None of the above.	7	12

Seventy-seven percent of the respondents thought that Ferrus' GE score was about the same as the average score for a student in the first month of 12th grade. A combined 10% gave choices other than "None of the above".

25. More detailed explanations of terms with examples would have been useful in helping me understand the test results.

Table 4.31 Question #25

Option	Count	Percentage of Form A
Strongly Disagree	1	2
Disagree	5	10
Neutral	9	17
Agree	20	38
Strongly Agree	17	33

Seventy-one percent endorsed "Agree" or "Strongly Agree" when asked if they thought a more detailed explanation with more examples would have helped them understand the test results. Other than these two options, the next highest number of respondents chose "Neutral" for this item.

The next set of questions referred to Report #3 for a student by the name of James taking a national standardized achievement test. Survey takers looked at the information and then answered the six questions that follow.

26. According to the report, James might have more trouble with words having more than one meaning compared to two different words that mean the same thing.

Table 4.32 Question #26

Option	Count	Percentage of Form A
*True**	39	75
False	10	19
Neither	3	6

Three quarters of the respondents put "True" and agreed with the statement above. Approximately 20% disagreed with the statement and chose "False" on the survey.

27. According to the report, James scored higher than average in all clusters of the Spelling subtest.

Table 4.33 Question #27

Option	Count	Percentage of Form A
True	10	20
*False**	40	78
Neither	1	2

The majority of respondents (78%) chose "False" for this statement and disagreed that James had scored higher than average in all clusters of the Spelling subtest.

28. According to the report, it might be beneficial for James to work on organizing his essays and reviewing the different ways paragraphs can be organized.

Table 4.34 Question #28

Option	Count	Percentage of Form A
*True**	46	88
False	3	6
Neither	3	6

A clear majority of the survey takers agreed that James should work on organizing his essays. An equal but small number of respondents chose "False" or "Neither" on this question.

29. According to the report, James got fewer points correct in Language Expression than in Language Mechanics.

Table 4.35 Question #29

Option	Count	Percentage of Form A
*True**	45	87
False	4	8
Neither	3	6

Eighty-seven percent of the survey takers agreed with the statement above, while smaller numbers chose "False" or "Neither".

30. I found this graphical display to be useful in informing me about the test results for the student.

Table 4.36 Question #30

Option	Count	Percentage of Form A
Strongly Disagree	2	4
Disagree	4	8
Neutral	3	6
Agree	33	63
Strongly Agree	10	19

A little more than three-fifths of the respondents agreed that the graphical information used in this report was useful in informing about the student's test results.

31. How James performed on the different clusters was beneficial information to include in the report.

Table 4.37 Question #31

Option	Count	Percentage of Form A
Strongly Disagree	2	4
Disagree	0	0
Neutral	6	12
Agree	23	44
Strongly Agree	21	40

Eighty-four percent of the survey takers either agreed or strongly agreed that having the information for each of the different clusters was beneficial information to include in the report. Twelve percent indicated they were neutral about the inclusion of the information.

The following are general questions for all of the reports that you have looked at today. These questions are solely to help the researcher gather additional information and will not be associated with any particular individual.

32. Of the three reports you looked at, which one would be the most useful to you?

Table 4.38 Question #32

Option	Count	Percentage of Form A
Report #1 (Sally)	4	4
Report #2 (Ferrus)	1	5
Report #3 (James)	47	90

A clear majority of the survey takers (90%) thought that the third report was the most useful one. This report included some cluster score information for the student. Some of the explanations given by respondents included that the third report had more detail, was broken down into a more usable way, and specifically what areas he needed improvement.

33. Of the three reports you looked at, which one would be the least useful to you?

Table 4.39 Question #33

Option	Count	Percentage of Form A
Report #1 (Sally)	35	69
Report #2 (Ferrus)	10	20
Report #3 (James)	6	12

Almost 70% of the respondents thought that the first report was the least useful one. Some of the explanations given by respondents included it had too many "stats", was harder to understand, and needed more details.

34. Would you prefer to access score reports online or receive paper copies?

Table 4.40 Question #34

Option	Count	Percentage of Form A
Online	22	46
Paper	26	54

There was an approximately 50-50 split in preference for score reports online versus on paper by the respondents. Some of the explanations given included wanting both versions available, liking the fact that online is more environmental, and having a paper copy to put in the student's file is easier.

35. Which of the following scores did you find most useful in the reports? You may indicate more than one.

Table 4.41 Question #35

Option	Count	Percent of Form A
Grade Equivalent (GE)	38	37
Points Possible	19	18
Percentile Rank (PR)	18	17
Ability/Achievement Comparison (AAC)	16	15
Normal Curve Equivalent (NCE)	8	8
Scaled Score	5	5

**For score type usefulness, there were multiple marks and these were adjusted for the counts for each type.*

The three most useful scores according to the survey respondents were the grade equivalent, the points possible, and the percentile rank. The ability/achievement comparison was a close fourth in usefulness according to the survey takers.

36. Which of the following scores did you find least useful in the reports? You may indicate more than one.

Table 4.42 Question #36

Option	Count	Percent of Form A
Normal Curve Equivalent (NCE)	19	25
Percentile Rank (PR)	11	15
Ability/Achievement Comparison (AAC)	9	12
Points Possible	6	8
Scaled Score	5	33
Grade Equivalent (GE)	5	7

**For score type usefulness, there were multiple marks and these were adjusted for the counts for each type.*

Scaled scores and normal curve equivalents were found to be the least useful as far as the survey respondents were concerned. Some respondents also indicted that the percentile

rank and ability/achievement comparison were also not as useful for them in looking at the reports.

4.7 Experimental Group Reports (Form B)

In this section, counts and percentages of each of the responses for the questions contained on the Experimental Form (Form B) will be given. The actual survey question, followed by a table for each question will be given. As with the Control Form, within each table, the correct answer will be indicated by *italics*. Appendix A contains the original survey and data displays. Information that might be provided for a student by the name of Sally taking a national standardized achievement test was displayed in the first Report. Respondents looked at the information and then answered the eight questions that follow.

10. According to the report, Sally is above average in Total Math compared to other 2nd graders in the nation.

Table 4.43 Question #10

Option	Count	Percentage of Form B
*True**	32	55
False	20	34
Neither	6	10

Fifty-five percent of the survey takers agreed with the question and chose "True", while 34% did not and endorsed "False".

11. According to the report, Sally got 82 percent of the questions correct on the Total Reading subtest.

Table 4.44 Question #11

Option	Count	Percentage of Form B
True	16	28
*False**	42	72
Neither	0	0

Most of the respondents did not agree with the question and chose "False" (72%). All of the remaining respondents chose "True" and agreed with the statement.

12. According to the report, Sally scored significantly better in Word Study Skills than Reading Vocabulary.

Table 4.45 Question #12

Option	Count	Percentage of Form B
*True**	41	71
False	13	22
Neither	4	7

The majority of survey takers agreed with the above statement and chose "True". Slightly less than one-quarter chose "False" and did not agree with the statement about Sally's score in Word Study Skills and Reading Vocabulary.

13. According to the report, Sally scored significantly better in Word Study Skills than Reading Comprehension.

Table 4.46 Question #13

Option	Count	Percentage of Form B
True	43	74
*False**	13	22
Neither	2	3

Three quarters of the respondents chose “True” and agreed with the statement above.

Approximately one quarter disagreed with the statement and chose “False” on the survey.

14. According to the report, Sally’s test score in Spelling is above that for students with the same ability level.

Table 4.47 Question #14

Option	Count	Percentage of Form B
*True**	54	93
False	2	3
Neither	2	3

A clear majority of the survey takers (93%) agreed with the question and chose “True”.

Very small percentages chose “False” or “Neither” as responses.

15. According to the report, Sally has the same proficiency of word study skills and knowledge as a third grader.

Table 4.48 Question #15

Option	Count	Percentage of Form B
True	42	74
*False**	11	19
Neither	4	7

A majority (74%) of the survey takers agreed that Sally had the same proficiency of word study skills and knowledge as a third grader. Slightly less than 20% chose “False” and disagreed with the statement.

16. Assuming no learning occurs, if a thousand students with the same ability as Sally repeatedly took the Total Reading test, their PR scores will:

Table 4.49 Question #16

Option	Count	Percentage of Form B
Always be equal to 59	16	28
*Vary between 40 and 70**	30	53
Vary between 40 and 59	6	11
Vary between 59 and 70	5	9

Slightly more than half of the survey takers said that the percentile rank will "Vary between 40 and 70", and the next highest percentage said that the percentile rank will "Always be equal to 59". The smallest percentage indicated that the percentile rank will "Vary between 59 and 70".

17. I found the inclusion of the black confidence bands around the student's score (black diamond) in the graph on the right-side of the score report useful interpreting the scores.

Table 4.50 Question #17

Option	Count	Percentage of Form B
Strongly Disagree	2	3
Disagree	9	16
Neutral	18	31
Agree	23	40
Strongly Agree	6	10

Seventy-one percent endorsed "Agree" or "Neutral" when asked if they thought the black confidence bands around the student's score (black diamond) in the graph on the right-side of the score report was useful in interpreting the scores. Other than these two options, the next highest number of respondents chose "Disagree" for this item.

Report #2 displayed information for a student named of Ferrus taking a national standardized achievement test. Respondents looked at the information and then answered the eight questions that follow.

18. According to the report, Ferrus scored higher than 64% of the other students in his comparison group in Total Math.

Table 4.51 Question #18

Option	Count	Percentage of Form B
*True**	46	79
False	9	16
Neither	3	5

The majority of the respondents (79%) agreed with the question and chose "True". Sixteen percent disagreed with the statement and chose "False".

19. According to the report, Ferrus earned 82 out of a possible 114 points on the Total Reading subtest.

Table 4.52 Question #19

Option	Count	Percentage of Form B
*True**	54	93
False	3	5
Neither	1	2

Ninety-three percent of the survey takers agreed with the statement above, while smaller numbers chose "False" or "Neither".

20. A scaled score of 623 in Math Procedures and a score of 623 in Thinking Skills means that Ferrus did the same in both subtests.

Table 4.53 Question #20

Option	Count	Percentage of Form B
True	3	5
*False**	47	82
Neither	7	12

The majority of the survey takers chose "False" for this question, while the next highest percentage chose "Neither". The smallest percentage chose "True".

21. According to his stanine scores in the report, Ferrus was average compared to other students.

Table 4.54 Question #21

Option	Count	Percentage of Form B
*True**	45	78
False	5	9
Neither	8	14

The majority of survey takers agreed with the above statement and chose "True". One-seventh chose "Neither" and 9% chose "False" when referring to how Ferrus compared to other students.

22. According to the report, Ferrus' test score in Social Science is higher than that for students with the same ability level.

Table 4.55 Question #22

Option	Count	Percentage of Form B
True	4	7
*False**	50	86
Neither	4	7

Eighty-six percent of the respondents chose "False" for this question, while equal, smaller percentages chose "True" or "Neither".

23. If we want to see how Ferrus' school did compared to another school by average student scores in Math, Percentile Ranks would be the most appropriate score to use.

Table 4.56 Question #23

Option	Count	Percentage of Form B
True	27	47
*False**	24	42
Neither	6	11

Almost equal numbers put "True" or "False" for this question. Eleven percent put "Neither", indicating the statement is neither true nor false.

24. You will notice on the report that Ferrus had a GE of 12.1 in Math Procedures on the report. What does this score of 12.1 mean?

Table 4.57 Question #24

Option	Count	Percentage of Form B
He should immediately be moved from the 11th grade and placed in a 12th grade Mathematics class.	2	3
*His scale score in Math Procedures is about the same as the average score for a student in the first month of 12th grade had they taken the test.**	45	78
He has mastered 11th grade Mathematics and can perform 12th grade work in Mathematics.	4	7
None of the above.	7	11

Seventy-eight percent of the respondents thought that Ferrus' GE score was about the same as the average score for a student in the first month of 12th grade. Eleven percent indicated "None of the above" for their choice.

25. I found the detailed explanation of terms with examples to be useful in helping me understand the test results.

Table 4.58 Question #25

Option	Count	Percentage of Form B
Strongly Disagree	2	3
Disagree	2	3
Neutral	11	19
Agree	20	34
Strongly Agree	23	40

The majority of respondents put "Strongly Agree" and agreed that the detailed explanation of terms was useful in helping them understand the test results. A little more than half chose either "Agree" or "Neutral". Smaller numbers of respondents chose the final two options on this question.

Report #3 displayed information for a student by the name of James taking a national standardized achievement test. Respondents looked at the information and then answered the six questions that follow.

26. According to the report, James might have more trouble with words having more than one meaning compared to two different words that mean the same thing.

Table 4.59 Question #26

Option	Count	Percentage of Form B
*True**	42	75
False	10	18
Neither	4	7

Only 18% of the survey takers did not think James might have more trouble with words having more than one meaning compared to two different words than the same meaning. Three-quarters of the respondents thought it was "True", while only 7% felt the statement was neither true nor false.

27. According to the report, James scored higher than average in all clusters of the Spelling subtest.

Table 4.60 Question #27

Option	Count	Percentage of Form B
True	6	11
*False**	46	84
Neither	3	5

Eighty-four percent of the survey takers chose "False" and did not agree with the statement for the question. Eleven percent chose "True" and agreed that James scored higher than average in all clusters of the Spelling subtest.

28. According to the report, it might be beneficial for James to work on organizing his essays and reviewing the different ways paragraphs can be organized.

Table 4.61 Question #28

Option	Count	Percentage of Form B
*True**	42	75
False	4	7
Neither	10	18

Three-quarters of the survey takers chose “True” and agreed that it might be beneficial for James to work on organizing his essays and reviewing the different ways paragraphs can be organized.

29. According to the report, James got fewer points correct in Language Expression than in Language Mechanics.

Table 4.62 Question #29

Option	Count	Percentage of Form B
*True**	48	86
False	5	9
Neither	3	5

A clear majority of the survey takers agreed that James received fewer points correct in Language Expression than Language Mechanics. An almost equal but small number of respondents chose “False” or “Neither” on this question.

30. I found this graphical display to be useful in informing me about the test results for the student.

Table 4.63 Question #30

Option	Count	Percentage of Form B
Strongly Disagree	1	2
Disagree	3	5
Neutral	10	18
Agree	26	46
Strongly Agree	16	29

The majority of survey takers either agreed or strongly agreed that the graphical display was useful in informing them about the test results for the student. The next highest option was “Neutral” at eighteen percent.

31. How James performed on the different clusters was beneficial information to include in the report.

Table 4.64 Question #31

Option	Count	Percentage of Form B
Strongly Disagree	2	4
Disagree	1	2
Neutral	5	9
Agree	25	45
Strongly Agree	23	41

The vast majority of respondents (86%) endorsed "Agree" or "Strongly Agree" when asked if they thought including the information on how James performed on the different clusters was beneficial information to include in the report. Only a combined six percent chose "Disagree" or "Strongly Disagree".

The following are general questions for all of the reports that you have looked at today. These questions are solely to help the researcher gather additional information and will not be associated with any particular individual.

32. Of the three reports you looked at, which one would be the most useful to you?

Table 4.65 Question #32

Option	Count	Percentage of Form B
Report #1 (Sally)	4	7
Report #2 (Ferrus)	7	13
Report #3 (James)	45	80

A clear majority of the survey takers (80%) thought that the third report was the most useful one. This report included some cluster score information for the student. Some of

the explanations given by respondents included that the performance clusters would be very useful for teachers by allowing them to pinpoint areas for improvement and might be easier for a parent to comprehend. On the other hand, a respondent that preferred the second report thought it has just the right amount of information—not too little as the first report or not too much as in the third report.

33. Of the three reports you looked at, which one would be the least useful to you?

Table 4.66 Question #33

Option	Count	Percentage of Form B
Report #1 (Sally)	38	68
Report #2 (Ferrus)	9	16
Report #3 (James)	9	16

Sixty-eight percent of the survey takers thought that the first report was the least useful one. Some of the explanations given by respondents included it needed more detail or that it needed to provide more explanations of the terms. Here, there were also some respondents that felt that too much information was given in the final report and it might be a little confusing.

34. Would you prefer to access score reports online or receive paper copies?

Table 4.67 Question #34

Option	Count	Percentage of Form B
Online	30	53
Paper	27	47

There was an approximately 50-50 split in preference for score reports online versus on paper by the respondents. Many explanations given included being better able to keep a record—this referred to the electronic or the paper depending on the respondent. Other points made were having a tangible paper copy can be better for explaining to parents, less clutter with online, and liking the ability to annotate a hard copy.

35. Which of the following scores did you find most useful in the reports? You may indict more than one.

Table 4.68 Question #35

Option	Count	Percentage of Form B
Percentile Rank (PR)	33	28
Grade Equivalent (GE)	23	19
Ability/Achievement Comparison (AAC)	22	18
Points Possible	18	15
Normal Curve Equivalent (NCE)	13	11
Scaled Score	11	9

**For score type usefulness, there were multiple marks and these were adjusted for the counts for each type.*

The most useful scores according to the survey respondents was the percentile rank. Grade equivalent, ability/achievement comparison, and points possible were next and had percentages that were close to each other, but still several percentage points less than the percentile rank.

36. Which of the following scores did you find least useful in the reports? You may indict more than one.

Table 4.69 Question #36

Option	Count	Percentage of Form B
Scaled Score	27	34
Normal Curve Equivalent (NCE)	15	19
Ability/Achievement Comparison (AAC)	14	18
Grade Equivalent (GE)	10	13
Points Possible	9	11
Percentile Rank (PR)	4	5

**For score type usefulness, there were multiple marks and these were adjusted for the counts for each type.*

Scaled scores, normal curve equivalents, and ability/achievement comparisons were found to be the least useful, as far as the survey respondents were concerned. Scaled score received 34% of all responses, normal curve equivalent received 19%, and ability/achievement comparisons received18%. The smallest percentage was for the percentile rank.

4.8 Item and Total Statistics

The last part of Study 3 involved scoring the responses of the participants to the set of eighteen questions that could be marked as “True”, “False”, or “Neither”. For these questions, there were correct answers based on the available reports, and these were given a value of 0 for incorrect and 1 for correct and then added to get a sum score with a total of eighteen possible. Within this set of eighteen questions, twelve were common to the two forms and summed to get a subscore. Item means and correlations to the scored items for respondents are given in the tables below by form. The first table has the items common across the two forms aligned so that comparisons may be made, while the second table has items that were not common across forms and therefore cannot directly be compared. For example, item 10 is the same across forms and comparisons can be

made, while item 12 on Form A is different than item 12 on Form B and comparisons cannot be made. The item means give an idea on how difficult the individual items were for the respondents.

Table 4.70 Item Means for Common Items

	Control (Form A)		Experimental (Form B)
Question	**Mean**	**Question**	**Mean**
10	0.48	10	0.54
11	0.67	11	0.74
13	0.87	14*	0.95
15	0.71	15	0.74
16	0.37	16	0.53
21	0.81	22*	0.88
23	0.19	23	0.42
24	0.77	24	0.79
26	0.75	26	0.74
27	0.77	27	0.81
28	0.89	28	0.74
29	0.87	29	0.84
N=52		N=57	

**Questions 13 & 21 in Form A are the same as 14 & 22 in Form B, respectively.*

Table 4.71 Item Means for Non-Common Items

Control (Form A)	
Question	**Mean**
12	0.60
14	0.12
18	0.56
19	0.75
20	0.21
22	0.75
N=52	

Experimental (Form B)	
Question	**Mean**
12	0.72
13	0.21
18	0.81
19	0.95
20	0.83
21	0.79

N=57

To check on the internal consistency of the questionnaire, Coefficient Alpha was calculated both for all eighteen scored items and for the twelve common items. The alpha value for the eighteen items was low and since this study is interested in how the groups in the control and experimental groups compare, the alphas using the twelve items and the corresponding subscore are presented below:

Table 4.72 Coefficient Alpha of Questionnaire

Coefficient Alpha	Control	Experimental
Raw	0.65	.71

These two values are within commonly acceptable values.

4.9 Sum Scores Across Groups

To gather an overall picture on the results of the scored items, comparisons of mean scores, standard deviations (SD), and counts for key demographic variables are presented in the tables below, with all respondents combined.

Table 4.73 Capacity Filled Out Survey

Option	Mean	SD	Count
Student	7.96	2.63	26
Parent	7.95	2.35	40
Teacher	9.24	1.65	34
Administrator	8.89	1.54	9

Table 4.74 Years of Teaching and/or Administrative Experience

Option	Mean	SD	Count
Fewer than 3 years	9.75	1.26	4
4 to 9 years	9.55	1.63	11
10 to 20 years	9.00	2.12	9
More than 20 years	9.00	1.49	20

Table 4.75 Years of Involvement in the Local School System

Option	Mean	SD	Count
0 to 5 years	8.58	2.61	12
6 to 8 years	8.14	2.67	7
9 to 12 years	7.75	3.02	16
More than 12 years	8.10	1.97	21

Table 4.76 Received Training in Educational Testing

Option	Mean	SD	Count
Yes	8.79	1.95	38
No	8.23	2.36	71

Table 4.77 Level of Education

Option	Mean	SD	Count
Some high school	7.75	2.69	28
High School diploma/GED	6.33	2.73	6
Associates Degree	7.5	2.12	2
Bachelors degree	9.08	1.75	38
Post bachelors degree (Masters, Doctorate)	8.69	1.94	35

In the tables above it can be seen that teachers and administrators had higher mean scores than students or parents. An increase in the number of years of involvement in the local school system or years teaching didn't appear to have an impact on the mean scores. Those participants with and without training in educational testing had mean scores that were similar. Finally, participants holding bachelors degrees and above had higher mean scores than those without.

4.10 Control vs. Experimental Group Sum Score Comparison

Within the set of eighteen questions that could be marked as "True", "False", or "Neither", twelve were common to the two forms and summed to get a subscore. These subscores of the control and experimental groups were then statistically compared using an independent-samples t-test. An independent-samples t-test was conducted to compare whether the addition of the supplemental information recommended by previous research aided the participants and made a difference in the sum subscores. There was not a significant difference between the control (M=8.1, SD=1.9) and experimental (M=8.7, SD=2.5) groups; $t(107) = -1.33$, $p = 0.19$. These results suggest that including the various pieces of information did not have an effect on the correct understanding of the score report readers.

4.11 Control vs. Experimental Group Common Item Comparison

To attempt to tease out any specific questions where the control and experimental groups statistically differ, t-tests were performed for the twelve items that were common across the two forms. As can be seen below, only item 23 had a statistically significant difference. This question dealt with the most appropriate score to use for school to

school comparisons. Additional interpretive material helped the experimental group score higher on this particular item.

Table 4.78 Common Item Statistical Comparisons

Question*	Form A Mean	Form B Mean	Form A SD	Form B SD	t	p-value	Statistically Significant
10	0.48	0.54	0.50	0.50	-0.65	0.51	
11	0.67	0.74	0.47	0.44	-0.07	0.47	
13 (14)	0.87	0.95	0.34	0.23	-1.48	0.14	
15	0.17	0.19	0.38	0.40	-0.27	0.79	
16	0.37	0.53	0.49	0.50	-1.69	0.09	
21 (22)	0.81	0.88	0.40	0.33	-0.99	0.32	
23	0.19	0.42	0.40	0.50	-2.63	0.01	**
24	0.77	0.79	0.43	0.41	-0.25	0.80	
26	0.75	0.74	0.44	0.44	0.16	0.91	
27	0.77	0.81	0.43	0.40	-0.48	0.63	
28	0.88	0.74	0.32	0.44	1.97	0.05	
29	0.87	0.84	0.34	0.37	0.30	0.73	

**Number in parentheses is for Form B item number.*

CHAPTER 5

DISCUSSION AND RECOMMENDATIONS

5.1 Introduction

Two versions of a questionnaire were distributed to teachers, high school students, parents, and administrators. One was the control version and had information presented in the conventional way, while the experimental version tried to improve upon these displays by incorporating elements discussed in the literature. The first part of the questionnaire was identical in the two forms and contained several items designed to obtain demographic information about the respondents. The information included grade levels taught or experienced, the work environment of the respondent, the population of the local school district, the number of years of teaching experience, exposure and type of training with testing and measurement, the usefulness of the training, gender, and ethnicity. Based on the descriptions of the total sample and of the two subsets that took the two different versions are found in the Results section, it can be seen the participants that took the experimental and control versions were similar in make-up. There are some slight differences in percentages, but these can be attributed the modest sized samples that took each form.

The second part of the questionnaire consisted of a mixture of Likert-type items and true/false/neither items. For each Likert-type item, the respondents were asked to circle on a scale the number that most closely corresponded to their opinion on that item. A five-point scale was used for each item, where a value of “1” indicated the least favorable

response and a value of “5” indicated the most favorable response. For example, an average response of less than “2” is indicative of a largely negative position on the item. The choice “NA” (the abbreviation of “not applicable”) was also provided. The true/false/neither items indicated how the respondents felt about the series of statements dealing with the factual nature of the graphs. Additionally, there were items that were traditional multiple-choice that had a stem question and then four options for the respondent to choose from based on the graphical display that had been given to them. These were intended to check if the respondents were able to understand the displays and gather the necessary information intended to be projected from the graphics.

5.2 General Questions

The general questions were in the second part of the questionnaire after the specific questions on individual reports, but will be discussed here first. The first of these questions asked the participants which data display they found useful and then which they found the least useful. The results for both the experimental and control groups are mixed in that the addition of more information was not always found to be useful for the participants. There were many that welcomed and wanted more detailed information (e.g. diagnostic in Report #3) but some participants expressed that more information was a little too much and may have distracted them. These two points of view were reflected in the answers to which was most and which was least useful to the respondents to interpret the score reports.

Regarding the preference for online versus paper reports, there was about an even split in the sample. Some of the respondents preferred to have both in fact. It is hard to say if

there is any general pattern by age, since this piece of information was not collected, but it is also hard to find another general pattern. Some teachers and parents preferred paper, while some teachers and parents preferred online. The above seemed slightly surprising, in that when this question was written, it was expected that most participants would choose online. This is something to keep in mind as states and testing organizations move forward disseminating the results of assessments.

Percentile ranks appeared to be most useful to the participants, while scaled scores, normal curve equivalents, and ability/achievement comparisons were found to be less so. A percentile rank is probably found to be more easily understood by the general population and by educators than the latter three types of scores. Even with the addition of more interpretive information in the reports, these last three types of scores may still continue to give the end users of these reports more trouble. More technically minded psychometricians and statisticians like these other types of scores, but they may never be as useful to others unless ways can be found to educate these groups on these other important score scales. Though, in hindsight, probably NCEs will never have much value to parents, teachers, and administrators. On the other hand, they were designed for and valuable to persons involved in program evaluation. Impara et al. (1991) in their report for NAEP had a similar sentiment.

5.3 Interpretive Questions and Group Comparisons

These were the questions that were designed to test the experimental condition that using the types of information recommended in previous research would truly make a difference for report readers. The questions were designed to have a correct answer that

was based on the information presented and then the means and standard deviations of the summative scores for all the items on each form were compared. The twelve common items between the two versions were also compared using statistical analysis, and in this case it was found that there was not a statistical difference in the summative scores between the control and experimental groups. This may seem surprising since it does not follow the generally accepted guidelines for what items to include in designing a good set of score reports. However, all research can have limitations and these will be discussed more in the following section.

5.4 Limitations and Recommendations

The results of this experiment reported in Study 3 did not show that using conventional wisdom for what to include when designing score reports for assessment results should stop. Several of these were purposely included in the design of the score reports used in this research in the hope of confirming that inclusion would make the experimental group have a better understanding and therefore have higher summative scores. Again, no statistically significant differences were seen between the experimental and control groups. However, when these two different sets of score reports were being developed, it was found that the complete absence of certain information was not possible and that the complete absence or addition of the experimental manipulators was more of a gradient These gradients of the factors may have affected the results. For example, some basic information on the definitions of the score types needed to be included in the control form, while more extensive definitions and examples were given in the experimental condition. Initially, there were only to be explanations of the abbreviations in the graphs

and no additional information so that there would be a true absence or presence of the experimental manipulation.

There are two recommendations that can be drawn from this study that may be used as guidelines for reporting results of large-scale assessments. First, designing and constructing score reports and interpretive materials based on previous recommendations discussed in the literature should continue. While this particular research does not necessarily support them, neither does it automatically lead to a recommendation to dismiss them. The next section reiterates some of the previous recommendations. Secondly, further experimental research on this topic is needed to draw definitive conclusions on the use of particular scores and designs for score reports. This study was broader in nature than additional research that could focus only on score types, for example, or other designs. The inclusion of individual interviews, focus groups, and surveys attempted to gather a broad range of information. Use of only focus groups for a particular stake holder group with a narrower focus may be more useful for a particular state or assessment organization.

In light of the movement toward multi-state assessments by consortia such as the Common Core State Standards (CCSS) and Partnership for Assessment of Readiness for College and Careers (PARCC), there will be further opportunities to study reporting assessment scores that have more of an emphasis on electronic distribution of results. Both CCSS and PARCC have requirements for online testing of students rather than the traditional paper and pencil that have been predominantly used until now. A follow-up

study that focuses on electric presentation of reports that can incorporate factors not available to paper results can be done. For example, the use of a drill down that has results for a school, where the recipient can then click on these higher level results to go to a particular classroom or even down to a particular child. Videos or links to other resources such as explanations of the assessment in general or what particular types of scores mean can be added. The use of audio and visual forms of communication as another medium of instruction in addition to just written explanations of terms may add to the reader's comprehension of the results. The effectiveness of these additional pieces of information can be studied. Factors that are considered during a typical software usability study such as amount of time spent on a particular web page or the sequence links are clicked can help determine what are the more useful pieces of information to present to the report recipients. These areas can be explored in a focus group type setting or can be expanded to participants in an online survey. These types of studies can be more complicated, but can potentially yield troves of useful information.

5.5 Recommendations for Presenting Assessment Results

During the design of the sample score reports used during this study, recommendations and guidelines from previous studies and research were kept in mind. As noted above, the current study does not automatically dismiss these recommendations, so it would be good to again reiterate some of these guidelines and add when appropriate how they were used in the current study:

- Student score reports should be clear, concise, and visually attractive.
- They should also include easy-to-read text that supports and improves the interpretation of charts and tables.

- Care should be taken to not to try to do too much with a data display (i.e. displays would be designed to satisfy a small number of pre-established purposes).
    - Having a crowded versus not crowded display due to whitespace was a factor manipulated in the study.
- Devices such as boxes and graphics should be used to highlight main findings.
- Data should be grouped in meaningful ways.
- Small font, footnotes, and statistical jargon should be avoided.
- Key terms should be defined, preferably within a glossary (where they can be easily located by users).
    - This was one of the manipulated factors in this study.
- Reports should be piloted with members of the intended audience.
    - Individual interviews and focus groups composed of intended stake holders were key steps in this study.
- Consideration should be given to the creation of specially-designed reports that cater to the needs of different users (i.e. a detailed score report may be appropriate for teachers, but a simpler report may be more appropriate for widespread distribution to parents).
- Include all information essential to proper interpretation of assessment results in student score reports.
- Include detailed information about the assessment and score results in a separate interpretive guide, ideally one in which the student score report can be inserted.
- Personalize the student score reports and interpretive guides.

- 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.
    - A more diagnostic type report was included in the study.
- Include sample questions in the interpretive guides that illustrate the types of achievement represented by each performance level.
- Include a reproduction of student score reports in the interpretive guides to clearly explain the various elements of the reports.

5.6 Conclusion

The studies in this research contributed to the overall research study by progressively providing information on the use and interpretation of information from score reports by several different stakeholders, and they hoped to contribute to the overall goal of exploring different methods for displaying test data (scores) that will be utilized by various stakeholders in large-scale testing. While the results of comparing the experimental and control group in this study statistically have not been able to provide definitive conclusions, they have provided additional information that can add to the corpus of best practices for dissemination information from assessments to various stake holders.

APPENDIX A

CONTROL AND EXPERIMENTAL GROUP SURVEYS

FORM A

CONTROL GROUP VERSION

QUESTIONNAIRE ON REPORTS OF EDUCATIONAL TEST SCORES

In this questionnaire you will be asked for your opinion on several educational test scores. The results that follow may easily represent results of a state testing program such as the Texas Assessment of Knowledge and Skills (TAKS) or another large-scale assessment such as the Stanford Achievement Test, 10th Edition (SAT10). The reports are for students from different grade levels who took an assessment consisting of several subtests such as Reading, Mathematics, or Science. Although the scores and results are fictitious and have been created for this study, they are similar to the results that are routinely reported in practice.

You will also be asked a few background questions to help describe the persons participating in the study. Please answer them as honestly and conscientiously as possible so that the results can best represent the opinions of you and other users of score reports. You will be shown a series of score reports and asked to answer a small number of questions about each one. Please try to answer the questions in the order they are given and do not go back to previous questions to change your answers.

After you have completed the questionnaire and, if you wish, you can send an email with your name to the email address below to have your name entered into a drawing for one of four $50 gift cards. Your name and email address will not be associated with a specific questionnaire but will only be used for the drawing.

Sign the enclosed consent form and return one copy with the questionnaire. Participants who are students will also need to have a parent or guardian sign the form. Please return the questionnaire by December 9th or within two weeks of receipt.

Thank you again for your participation.

Stephen J. Jirka
Senior Research Associate
University of Massachusetts Amherst
scorereportsurvey@gmail.com
(210) 555-8596

Background Questions

The first nine questions are about your local community, experiences, and training in educational testing. Your answers will help us describe the persons who are participating in the study and will not be associated with any particular individual.

1. In what capacity will you complete this questionnaire? If you have multiple roles, please indicate your primary one.

   - o Student
   - o Parent
   - o Teacher
   - o Administrator (e.g. Principal, Dept. Head, Curriculum Director, Counselor)

2. How would you describe the community in which the school you are involved in is located?

   - o Rural
   - o Suburban
   - o Urban
   - o Other

3. If you are filling out this survey as a teacher or administrator, how many years of teaching and/or administrative experience do you have?

   - o Fewer than 3 years
   - o 4 to 9 years
   - o 10 to 20 years
   - o More than 20 years
   - o Does not apply to me

4. If you are filling out this survey as a parent or student, how many years of involvement do you have in the local school system?

   - o 0 to 5 years
   - o 6 to 8 years
   - o 9 to 12 years
   - o More than 12 years
   - o Does not apply to me

5. Have you ever received training in educational testing?

   - o Yes (go to question 6)
   - o No (go to question 8)

6. What type of training was it?

   - o Workshop sponsored by school/district
   - o Workshop while attending conference
   - o Course at college or university
   - o Book or online training
   - o Other. Please specify _____.

7. If you are filling out this survey as a teacher or administrator, how well do you think this training or education in testing and measurement has prepared you to deal with understand and use the results of large-scale assessments that might be used in your school and district?

   - o Not useful at all
   - o Rarely useful
   - o Somewhat useful
   - o Generally useful
   - o Very useful

8. What is your current level of education?

   - o Some high school
   - o High school diploma/GED
   - o Associates degree
   - o Bachelors degree
   - o Post bachelors degree (Masters, Doctorate)

9. What is your racial/ethnic background?

   - o American Indian/Alaskan Native
   - o Asian/Pacific Islander
   - o Black, Non-Hispanic
   - o Hispanic
   - o White, Non-Hispanic

Please go on to the next page where you will see a student report, followed by several questions that refer to this report. There are three reports followed by questions in this survey. Please try to go through the reports and answer the questions in the order they are given and do not go back to previous questions to change your answers.

Research Report #1

Student Report for Sally Wang

Teacher: Ramirez
School: Oak
District: Westin

Grade: 2
Test Date: April 2009

Age: 9
Student Number: 1233456

Subtests and Totals	Points Possible	Points Correct	Scaled Scores	National PR-S	National NCE	GE	AAC Range	National Grade Percentiles 1 10 30 50 70 90 99
Total Reading	114	82	639	59-5	54.8	2.7	MIDDLE	
Word Study Skills	30	25	664	76-6	64.9	3.1	HIGH	
Reading Vocabulalry	30	22	627	46-5	47.9	2.5	MIDDLE	
Reading Comprehension	54	35	634	53-5	51.6	2.6	MIDDLE	
Total Math	80	56	633	64-6	57.5	2.7	MIDDLE	
Math Problem Solving	48	30	623	54-5	52.1	2.5	MIDDLE	
Math Procedures	32	26	650	74-6	63.5	3.1	HIGH	
Language	48	28	610	39-4	44.1	1.2	MIDDLE	
Language Mechanics	24	15	617	46-5	47.9	1.8	MIDDLE	
Language Expression	24	13	603	36-4	42.5	1.4	MIDDLE	
Spelling	40	30	647	73-6	62.9	2.8	HIGH	
Science	40	30	643	69-6	60.4	2.7	MIDDLE	
Social Science	40	22	607	40-5	44.7	2.3	MIDDLE	
Listening	40	22	608	35-4	41.9	2.1	MIDDLE	
Thinking Skills	190	122	623	56-5	53.2	2.5	MIDDLE	
Basic Battery	322	218	NA	57-5	53.6	2.5	MIDDLE	
Complete Battery	402	270	NA	56-5	53.4	2.5	MIDDLE	

PR-S = Percentile Rank and Stanine **NCE = Normal Curve Equivalent** **AAC = Achievement/Ability Comparison** **GE = Grade Equivalent**

Scaled Score: the student's reported score
PR-S: percentage of a student's peer group with scores less than or equal to that particular score
NCE: ranging from 1-99, indicating how many students out of a hundred had a lower score
AAC: relationship between an individual's score on a subtest of an achievement test and the scores of other students of similar ability as measured by an ability test.
GE: estimate of the performance that an average student at a grade level is assumed to demonstrate on the test at a particular time in the school year

The following questions refer to Report #1.

In Report #1 appears a display of information that might be provided for a student by the name of **Sally** taking a national standardized achievement test. Look at the information and then answer the eight questions that follow.

10. According to the report, Sally is above average in Total Math compared to other 2nd graders in the nation.

    - o True
    - o False
    - o Neither

11. According to the report, Sally got 82 percent of the questions correct on the Total Reading subtest.

    - o True
    - o False
    - o Neither

12. According to the report, Sally scored significantly better in Word Study Skills and Reading Vocabulary than Reading Comprehension.

    - o True
    - o False
    - o Neither

13. According to the report, Sally's test score in Spelling is above that for students with the same ability level.

    - o True
    - o False
    - o Neither

14. According to the report, Sally has the same proficiency of word study skills and knowledge as a third grader

    - o True
    - o False
    - o Neither

15. According to the report, Sally has the same proficiency of word study skills and knowledge as a third grader.

    - o True
    - o False
    - o Neither

16. Assuming no learning occurs, if a thousand students with the same ability as Sally repeatedly took the Total Reading test, their PR scores will:

    - o Always be equal to 59
    - o Vary between 40 and 70
    - o Vary between 40 and 59
    - o Vary between 59 and 70

17. This score report helps me understand the Sally's strengths and weaknesses.

    - o Strongly Disagree
    - o Disagree
    - o Neutral
    - o Agree
    - o Strongly Agree

Research Report #2

PR-S = Percentile Rank and Stanine
NCE = Normal Curve Equivalent
AAC = Achievement/Ability Comparison
GE = Grade Equivalent

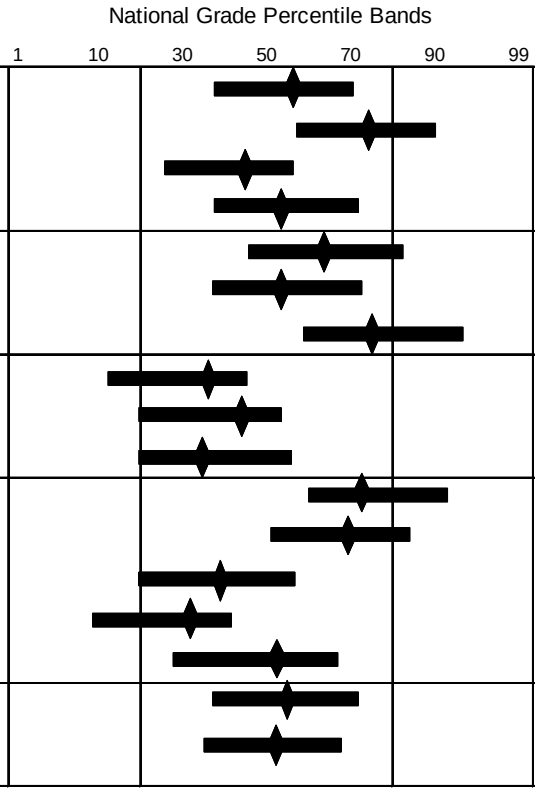
Student Report for Ferrus Abbda

Teacher: Lee
School: Palm
District: Comfort

Grade: 11
Test Date: April 2009

Age: 16
Student Number: 6445522

Subtests and Totals	Points Possible	Points Correct	Scaled Scores	National PR-S	National NCE	GE	AAC Range
Total Reading	114	82	639	59-5	54.8	11.7	MIDDLE
Word Study Skills	30	25	664	76-6	64.9	12.1	HIGH
Reading Vocabulalry	30	22	627	46-5	47.9	11.5	MIDDLE
Reading Comprehension	54	35	634	53-5	51.6	11.6	MIDDLE
Total Math	80	56	633	64-6	57.5	11.7	MIDDLE
Math Problem Solving	48	30	623	54-5	52.1	11.5	MIDDLE
Math Procedures	32	26	650	74-6	63.5	12.1	HIGH
Language	48	28	610	39-4	44.1	10.2	MIDDLE
Language Mechanics	24	15	617	46-5	47.9	10.8	MIDDLE
Language Expression	24	13	603	36-4	42.5	10.4	MIDDLE
Spelling	40	30	647	73-6	62.9	11.8	HIGH
Science	40	30	643	69-6	60.4	11.7	MIDDLE
Social Science	40	22	607	40-5	44.7	11.3	MIDDLE
Listening	40	22	608	35-4	41.9	11.1	MIDDLE
Thinking Skills	190	122	623	56-5	53.2	11.5	MIDDLE
Basic Battery	322	218	NA	57-5	53.6	11.5	MIDDLE
Complete Battery	402	270	NA	56-5	53.4	11.5	MIDDLE



Scaled Score: the student's reported score
PR-S: percentage of a student's peer group with scores less than or equal to that particular score
NCE: ranging from 1-99, indicating how many students out of a hundred had a lower score
AAC: relationship between an individual's score on a subtest of an achievement test and the scores of other students of similar ability as measured by an ability test.
GE: estimate of the performance that an average student at a grade level is assumed to demonstrate on the test at a particular time in the school year

The following questions refer to Report #2.

In Report #3 appears a display of information that might be provided for a student by the name of **Ferrus** taking a national standardized achievement test. Look at the information and then answer the eight questions that follow.

18. According to the report, Ferrus is above average in Total Math compared to other 11th graders in the nation.

    - o True
    - o False
    - o Neither

19. According to the report, Ferrus got 82 percent of the questions correct on the Total Reading subtest.

    - o True
    - o False
    - o Neither

20. According to the report, Ferrus scored significantly better in Math Procedures than Math Problem Solving.

    - o True
    - o False
    - o Neither

21. According to the report, Ferrus' test score in Social Science is above that for students with the same ability level.

    - o True
    - o False
    - o Neither

22. According to the report, Ferrus' scale score in Math Problem Solving was about the same as the average score for a student in the first month of 12th grade had they taken his test.

    - o True
    - o False
    - o Neither

23. If we want to see how Ferrus' school did compared to another school by average student scores in Math, Percentile Ranks would be the most appropriate score to use.

    - o True
    - o False
    - o Neither

24. You will notice on the report that Ferrus had a GE of 12.1 in Math Procedures on the report. What does this score of 12.1 mean?

    - o He should immediately be moved from the 11th grade and placed in a 12th grade Mathematics class.
    - o He is scale score in Math Procedures about the same as the average score for a student in the first month of 12th grade had they taken the test.
    - o He has mastered 11th grade Mathematics and can perform 12th grade work in Mathematics.
    - o None of the above.

25. More detailed explanations of terms with examples would have been useful in helping me understand the test results.

    - o Strongly Disagree
    - o Disagree
    - o Neutral
    - o Agree
    - o Strongly Agree

Research Report #3

Student Report for James Garcia

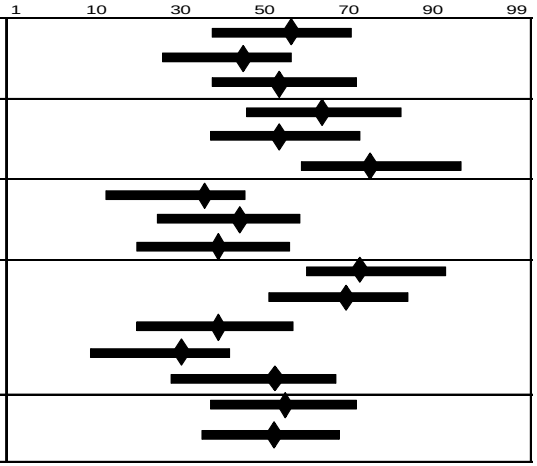
Teacher: Moto
School: Birch
District: Marriot

Grade: 8
Test Date: April 2009

Age: 14
Student Number: 223218

Subtests and Totals	Points Possible	Points Correct	Scaled Scores	National PR-S	National NCE	GE	AAC Range
Total Reading	84	57	639	59-5	54.8	8.7	MIDDLE
Reading Vocabulalry	30	22	627	46-5	47.9	8.5	MIDDLE
Reading Comprehension	54	35	634	53-5	51.6	8.6	MIDDLE
Total Math	80	56	633	64-6	57.5	8.7	MIDDLE
Math Problem Solving	48	30	623	54-5	52.1	8.5	MIDDLE
Math Procedures	32	26	650	74-6	63.5	9.1	HIGH
Language	48	28	610	39-4	44.1	7.2	MIDDLE
Language Mechanics	24	15	617	46-5	47.9	7.8	MIDDLE
Language Expression	24	13	603	36-4	42.5	7.4	MIDDLE
Spelling	40	30	647	73-6	62.9	8.8	HIGH
Science	40	30	643	69-6	60.4	8.7	MIDDLE
Social Science	40	22	607	40-5	44.7	8.3	MIDDLE
Listening	40	22	608	35-4	41.9	8.1	MIDDLE
Thinking Skills	190	122	623	56-5	53.2	8.5	MIDDLE
Basic Battery	322	218	NA	57-5	53.6	8.5	MIDDLE
Complete Battery	402	270	NA	56-5	53.4	8.5	MIDDLE

National Grade Percentile Bands



PERFORMANCE ON CLUSTERS

	Bel Avg	Avg	Abv Avg
Reading Vocabulary		*	
Synonyms		*	
Multiple Meaning Words	*		
Content Clues		*	
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Reading Comprehension		*	
Literary		*	
Informational		*	
Functional		*	
Initial Understanding		*	
Interpretation		*	
Critical Analysis		*	
Strategies		*	
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Mathematics Problems Solving		*	
Number Sense & Operations		*	
Patterns/Relationships/Algebra		*	
Data, Statistics, & Probability			*
Geometry and Measurement		*	
Comunication & representation		*	
Estimation	*		
Mathematical Connections		*	
Reasoning & Problem Solving			*
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Mathematical Proceduree		*	
Computation w ith w hole Numbers			*
Computation w ith fractions		*	
Computation w ith decimals		*	
Symbolic Notation Computation		*	
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Language Mechanics		*	
Capitalization			*
Usage		*	
Punctuation		*	

	Bel Avg	Avg	Abv Avg
Language Expression		*	
Sentence Structure		*	
Prew riting		*	
Content & Organization	*		
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Spelling		*	
Phonetics principles		*	
Structural principles		*	
Homophones		*	

	Bel Avg	Avg	Abv Avg
Science		*	
Life			*
Physical		*	
Nature of Science		*	
Earth		*	
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Social Science		*	
History		*	
Geography		*	
Political Science		*	
Economics		*	
Appl of Know ledge		*	
Org, sum, interp information		*	
Determination of cause/effect		*	
Thinking Skills		*	

Scaled Score: the student's reported score
PR-S: percentage of a student's peer group with scores less than or equal to that particular score
NCE: ranging from 1-99, indicating how many students out of a hundred had a lower score

AAC: relationship between an individual's score on a subtest of an achievement test and the scores of other students of similar ability as measured by an ability test.
GE: estimate of the performance that an average student at a grade level is assumed to demonstrate on the test at a particular time in the school year

The following questions refer to Report #3.

In Report #3 appears a display of information that might be provided for a student by the name of **James** taking a national standardized achievement test. Look at the information and then answer the six questions that follow.

26. According to the report, James might have more trouble with words having more than one meaning compared to two different words that mean the same thing.

    - o True
    - o False
    - o Neither

27. According to the report, James scored higher than average in all clusters of the Spelling subtest.

    - o True
    - o False
    - o Neither

28. According to the report, it might be beneficial for James to work on organizing his essays and reviewing the different ways paragraphs can be organized.

    - o True
    - o False
    - o Neither

29. According to the report, James got fewer points correct in Language Expression than in Language Mechanics.

    - o True
    - o False
    - o Neither

30. I found this graphical display to be useful in informing me about the test results for the student.

    - o Strongly Disagree
    - o Disagree
    - o Neutral
    - o Agree
    - o Strongly Agree

31. How James performed on the different clusters was beneficial information to include in the report.

    - o Strongly Disagree
    - o Disagree
    - o Neutral
    - o Agree
    - o Strongly Agree

The following are general questions for all of the reports that you have looked at today. These questions are solely to help the researcher gather additional information and will not be associated with any particular individual.

32. Of the three reports you looked at, which one would be the **most useful to you**?

    - o Report #1 (Sally)
    - o Report #2 (Ferrus)
    - o Report #3 (James)

    Please explain your choice briefly.

33. Of the three reports you looked at, which one would be the **least useful to you**?

    - o Report #1 (Sally)
    - o Report #2 (Ferrus)
    - o Report #3 (James)

    Please explain your choice briefly.

34. Would you prefer to access score reports **online** or receive **paper** copies?

    - o Online
    - o Paper

    Please explain your choice briefly.

35. Which of the following scores did you find **most useful** in the reports? You may indicate more than one.

    - o Points Possible
    - o Scaled Score
    - o Percentile Rank (PR)
    - o Normal Curve Equivalent (NCE)
    - o Grade Equivalent (GE)
    - o Ability/Achievement Comparison (AAC)

36. Which of the following scores did you find **least useful** in the reports? You may indicate more than one.

    - o Points Possible
    - o Scaled Score
    - o Percentile Rank (PR)
    - o Normal Curve Equivalent (NCE)
    - o Grade Equivalent (GE)
    - o Ability/Achievement Comparison (AAC)

Thank you for taking time to respond to this questionnaire. Please feel free to use the space below or the back of this page to make any additional comments or questions you might have. If you would like a summary of the results of this study or have any further questions, please send an email to: **scorereportsurvey@gmail.com**

FORM B

EXPERIMENTAL GROUP VERSION

QUESTIONNAIRE ON REPORTS OF EDUCATIONAL TEST SCORES

In this questionnaire you will be asked for your opinion on several educational test scores. The results that follow may easily represent results of a state testing program such as the Texas Assessment of Knowledge and Skills (TAKS) or another large-scale assessment such as the Stanford Achievement Test, 10th Edition (SAT10). The reports are for students from different grade levels who took an assessment consisting of several subtests such as Reading, Mathematics, or Science. Although the scores and results are fictitious and have been created for this study, they are similar to the results that are routinely reported in practice.

You will also be asked a few background questions to help describe the persons participating in the study. Please answer them as honestly and conscientiously as possible so that the results can best represent the opinions of you and other users of score reports. You will be shown a series of score reports and asked to answer a small number of questions about each one. Please try to answer the questions in the order they are given and do not go back to previous questions to change your answers.

After you have completed the questionnaire and, if you wish, you can send an email with your name to the email address below to have your name entered into a drawing for one of four $50 gift cards. Your name and email address will not be associated with a specific questionnaire but will only be used for the drawing.

Sign the enclosed consent corm and return one copy with the questionnaire. Participants who are students will also need to have a parent or guardian sign the form. Please return the questionnaire by December 9th or within two weeks of receipt.

Thank you again for your participation.

Stephen J. Jirka
Senior Research Associate
University of Massachusetts Amherst
scorereportsurvey@gmail.com
(210) 555-8596

Background Questions

The first nine questions are about your local community, experiences, and training in educational testing. Your answers will help us describe the persons who are participating in the study and will not be associated with any particular individual.

1. In what capacity will you complete this questionnaire? If you have multiple roles, please indicate your primary one.

   - o Student
   - o Parent
   - o Teacher
   - o Administrator (e.g. Principal, Dept. Head, Curriculum Director, Counselor)

2. How would you describe the community in which the school you are involved in is located?

   - o Rural
   - o Suburban
   - o Urban
   - o Other

3. If you are filling out this survey as a teacher or administrator, how many years of teaching and/or administrative experience do you have?

   - o Fewer than 3 years
   - o 4 to 9 years
   - o 10 to 20 years
   - o More than 20 years
   - o Does not apply to me

4. If you are filling out this survey as a parent or student, how many years of involvement do you have in the local school system?

   - o 0 to 5 years
   - o 6 to 8 years
   - o 9 to 12 years
   - o More than 12 years
   - o Does not apply to me

5. Have you ever received training in educational testing?

   - o Yes (go to question 6)
   - o No (go to question 8)

6. What type of training was it?

   - o Workshop sponsored by school/district
   - o Workshop while attending conference
   - o Course at college or university
   - o Book or online training
   - o Other. Please specify _____.

7. If you are filling out this survey as a teacher or administrator, how well do you think this training or education in testing and measurement has prepared you to deal with understand and use the results of large-scale assessments that might be used in your school and district?

   - o Not useful at all
   - o Rarely useful
   - o Somewhat useful
   - o Generally useful
   - o Very useful

8. What is your current level of education?

   - o Some high school
   - o High school diploma/GED
   - o Associates degree
   - o Bachelors degree
   - o Post bachelors degree (Masters, Doctorate)

9. What is your racial/ethnic background?

   - o American Indian/Alaskan Native
   - o Asian/Pacific Islander
   - o Black, Non-Hispanic
   - o Hispanic
   - o White, Non-Hispanic

Please go on to the next page where you will see a student report, followed by several questions that refer to this report. There are three reports followed by questions in this survey. Please try to go through the reports and answer the questions in the order they are given and do not go back to previous questions to change your answers.

Research Report #1

Student Report for Sally Wang

Teacher: Ramirez
School: Oak
District: Westin

Grade: 2
Test Date: April 2009

Age: 9
Student Number: 1233456

Subtests and Totals	Points Possible	Points Correct	Scaled Scores	National PR-S	National NCE	GE	AAC Range	National Grade Percentile Band * 1 10 30 50 70 90 99
Total Reading	114	82	639	59-5	54.8	2.7	MIDDLE	
Word Study Skills	30	25	664	76-6	64.9	3.1	HIGH	
Reading Vocabulalry	30	22	627	46-5	47.9	2.5	MIDDLE	
Reading Comprehension	54	35	634	53-5	51.6	2.6	MIDDLE	
Total Math	80	56	633	64-6	57.5	2.7	MIDDLE	
Math Problem Solving	48	30	623	54-5	52.1	2.5	MIDDLE	
Math Procedures	32	26	650	74-6	63.5	3.1	HIGH	
Language	48	28	610	39-4	44.1	1.2	MIDDLE	
Language Mechanics	24	15	617	46-5	47.9	1.8	MIDDLE	
Language Expression	24	13	603	36-4	42.5	1.4	MIDDLE	
Spelling	40	30	647	73-6	62.9	2.8	HIGH	
Science	40	30	643	69-6	60.4	2.7	MIDDLE	
Social Science	40	22	607	40-5	44.7	2.3	MIDDLE	
Listening	40	22	608	35-4	41.9	2.1	MIDDLE	
Thinking Skills	190	122	623	56-5	53.2	2.5	MIDDLE	
Basic Battery	322	218	NA	57-5	53.6	2.5	MIDDLE	
Complete Battery	402	270	NA	56-5	53.4	2.5	MIDDLE	

PR-S = Percentile Rank and Stanine **NCE = Normal Curve Equivalent** **AAC = Achievement/Ability Comparison** **GE = Grade Equivalent**

***** The black bars around the student's scores (black diamond) indicate the area that one can be 95% sure contains the true test score, given that any measurement contains some error.

Scaled Score: the student's reported score.
PR: percentage of a student's peer group with scores less than or equal to that particular score.
Stanine: nine-point scale for test scores with the mean of 5, and1 being the highest and 9 the lowest.
NCE: a measure of where a student falls on a normal curve, indicating a student's rank compared to other students. It ranges from 1-99.
AAC: where an individual's score on a subtest of an achievement test compares within the range of scores of other students of similar ability as measured by an ability test.
GE: describes the grade level and month where the student's test score would have been average. e.g. 10.5 means the 5th month of 10th grade

The following questions refer to Report #1.

In Report #1 appears a display of information that might be provided for a student by the name of **Sally** taking a national standardized achievement test. Look at the information and then answer the eight questions that follow.

10. According to the report, Sally is above average in Total Math compared to other 2nd graders in the nation.

    - o True
    - o False
    - o Neither

11. According to the report, Sally got 82 percent of the questions correct on the Total Reading subtest.

    - o True
    - o False
    - o Neither

12. According to the report, Sally scored significantly better in Word Study Skills than Reading Vocabulary.

    - o True
    - o False
    - o Neither

13. According to the report, Sally scored significantly better in Word Study Skills than Reading Comprehension.

    - o True
    - o False
    - o Neither

14. According to the report, Sally's test score in Spelling is above that for students with the same ability level.

    - o True
    - o False
    - o Neither

15. According to the report, Sally has the same proficiency of word study skills and knowledge as a third grader.

    - o True
    - o False
    - o Neither

16. Assuming no learning occurs, if a thousand students with the same ability as Sally repeatedly took the Total Reading test, their PR scores will:

    - o Always be equal to 59
    - o Vary between 40 and 70
    - o Vary between 40 and 59
    - o Vary between 59 and 70

17. I found the inclusion of the black confidence bands around the student's score (black diamond) in the graph on the right-side of the score report useful interpreting the scores.

    - o Strongly Disagree
    - o Disagree
    - o Neutral
    - o Agree
    - o Strongly Agree

Research Report #2

Student Report for Ferrus Abbda

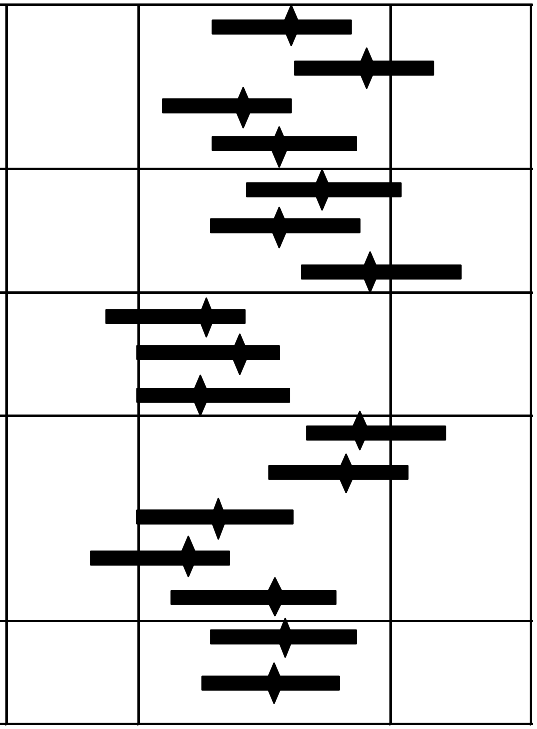
Teacher: Lee
School: Palm
District: Comfort

Grade: 11
Test Date: April 2009

Age: 16
Student Number: 6445522

Subtests and Totals	Points Possible	Points Correct	Scaled Scores	National PR-S	National NCE	GE	AAC Range
Total Reading	114	82	639	59-5	54.8	11.7	MIDDLE
Word Study Skills	30	25	664	76-6	64.9	12.1	HIGH
Reading Vocabulalry	30	22	627	46-5	47.9	11.5	MIDDLE
Reading Comprehension	54	35	634	53-5	51.6	11.6	MIDDLE
Total Math	80	56	633	64-6	57.5	11.7	MIDDLE
Math Problem Solving	48	30	623	54-5	52.1	11.5	MIDDLE
Math Procedures	32	26	650	74-6	63.5	12.1	HIGH
Language	48	28	610	39-4	44.1	10.2	MIDDLE
Language Mechanics	24	15	617	46-5	47.9	10.8	MIDDLE
Language Expression	24	13	603	36-4	42.5	10.4	MIDDLE
Spelling	40	30	647	73-6	62.9	11.8	HIGH
Science	40	30	643	69-6	60.4	11.7	MIDDLE
Social Science	40	22	607	40-5	44.7	11.3	MIDDLE
Listening	40	22	608	35-4	41.9	11.1	MIDDLE
Thinking Skills	190	122	623	56-5	53.2	11.5	MIDDLE
Basic Battery	322	218	NA	57-5	53.6	11.5	MIDDLE
Complete Battery	402	270	NA	56-5	53.4	11.5	MIDDLE

National Grade Percentile Bands *



PR-S = Percentile Rank and Stanine
NCE = Normal Curve Equivalent
AAC = Achievement/Ability Comparison
GE = Grade Equivalent

***** The black bars around the student's scores (black diamond) indicate the area that one can be 95% sure contains the true test score, given that any measurement contains some error.

Explanation of Terms

Points Possible: The number of raw score points available per subtest or total. Individual test items may be worth one point or several points.

Points Correct: The number of points the student obtained on the subtest or total by indicating correct responses. Blank, missing, and omitted responses are not counted correct.

Scaled Scores: A conversion of a student's raw score on a test or a version of the test to a common scale that allows for a numerical comparison between students. They are particularly useful for comparing test scores over time, such as measuring semester-to-semester and year-to-year growth of individual students or groups of students in a content area. However, within the same test, different content areas are typically on different scales, so direct comparisons cannot be made. For example, a score of 244 in Mathematics may not mean the same as a scaled score of 244 in Reading. The scaled score range on this assessment is 200-800.

National PR-S: The percentile rank (PR) is the percentage of a student's peer group (e.g., grade level) that a student's score surpassed. It is useful in comparing an individual student's performance with those of other students within the nation. For example, a student receives a test score of 66 and a percentile rank of 83. This means that a score of 66 is higher than 83% of the comparison group. The stanine (S) is another useful comparison. There are nine stanine units, ranging from 1 to 9. Typically, stanine scores are interpreted as below average (3, 2, 1), average (6, 5, 4), and above average (9, 8, 7). Because it uses only nine numbers, stanine scoring is usually easier to understand than other scoring models, and are useful in comparing a student's performance across different content areas. For example, a 6 in Mathematics and an 8 in Reading generally indicate a meaningful difference in a student's learning for the two respective content areas.

National NCE: A way of measuring where a student falls along the normal curve. The numbers on the NCE line run from 1 to 99, similar to percentile ranks, which indicate an individual student's rank, or how many students out of a hundred had a lower score. NCE scores have a major advantage over percentiles in that they can be averaged. That is an important characteristic when studying overall school performance, and in particular, in measuring school-wide gains and losses in student achievement.

AAC Range: This is the Ability/Achievement Comparison, and is the relationship between an individual's score on a subtest of an achievement test and the scores of other students of similar ability as measured by an ability test. If a student's achievement test score is higher than those of students of similar ability, the AAC is HIGH. If the achievement score is about the same as the scores of similar-ability students, the AAC is MIDDLE; if the score is lower, the AAC is LOW.

Grade Equivalent: This allows one to compare students based on the performance of other students relative to the school year. Based on a 9-month school year (typically September through May), the score represents a period during the school year, displayed as a number to show a grade and a month. The score is an estimate of the performance that an average student at a grade level is assumed to demonstrate on the test at a particular time in the school year. For example, a score of 6.8 represents a performance level typical of sixth-grade students in the eighth month (April) of the school year*. It is important to note that grade equivalent scores outside the current grade are common and should be interpreted with caution*. For example, a fifth-grade student could receive a grade equivalent score of 7.4. This does not mean the student can perform seventh-grade work – the student would not have been exposed to seventh-grade content, nor would a fifth-grade test contain seventh-grade content. It suggests that a typical seventh grader in the fourth month would have received the same score if seventh graders had taken the fifth-grade test.

The following questions refer to Report #2.

In Report #2 appears a display of information that might be provided for a student by the name of **Ferrus** taking a national standardized achievement test. Look at the information and then answer the eight questions that follow.

18. According to the report, Ferrus scored higher than 64% of the other students in his comparison group in Total Math.

    - o True
    - o False
    - o Neither

19. According to the report, Ferrus earned 82 out of a possible 114 points on the Total Reading subtest.

    - o True
    - o False
    - o Neither

20. A scaled score of 623 in Math Procedures and a score of 623 in Thinking Skills means that Ferrus did the same in both subtests.

    - o True
    - o False
    - o Neither

21. According to his stanine scores in the report, Ferrus was average compared to other students.

    - o True
    - o False
    - o Neither

22. According to the report, Ferrus' test score in Social Science is higher than that for students with the same ability level.

    - o True
    - o False
    - o Neither

23. If we want to see how Ferrus' school did compared to another school by average student scores in Math, Percentile Ranks would be the most appropriate score to use.

    - o True
    - o False
    - o Neither

24. You will notice on the report that Ferrus had a GE of 12.1 in Math Procedures on the report. What does this score of 12.1 mean?

    - o He should immediately be moved from the 11^{th} grade and placed in a 12^{th} grade Mathematics class.
    - o He is scale score in Math Procedures about the same as the average score for a student in the first month of 12^{th} grade had they taken the test.
    - o He has mastered 11^{th} grade Mathematics and can perform 12^{th} grade work in Mathematics.
    - o None of the above.

25. I found the detailed explanation of terms with examples to be useful in helping me understand the test results.

    - o Strongly Disagree
    - o Disagree
    - o Neutral
    - o Agree
    - o Strongly Agree

Research Report #3

Student Report for James Garcia

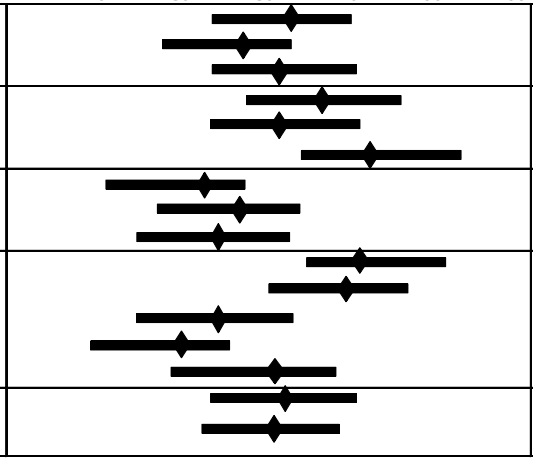
Teacher: Moto
School: Birch
District: Marriot

Grade: 8
Test Date: April 2009

Age: 14
Student Number: 223218

Subtests and Totals	Points Possible	Points Correct	Scaled Scores	National PR-S	National NCE	GE	AAC Range
Total Reading	84	57	639	59-5	54.8	8.7	MIDDLE
Reading Vocabulalry	30	22	627	46-5	47.9	8.5	MIDDLE
Reading Comprehension	54	35	634	53-5	51.6	8.6	MIDDLE
Total Math	80	56	633	64-6	57.5	8.7	MIDDLE
Math Problem Solving	48	30	623	54-5	52.1	8.5	MIDDLE
Math Procedures	32	26	650	74-6	63.5	9.1	HIGH
Language	48	28	610	39-4	44.1	7.2	MIDDLE
Language Mechanics	24	15	617	46-5	47.9	7.8	MIDDLE
Language Expression	24	13	603	36-4	42.5	7.4	MIDDLE
Spelling	40	30	647	73-6	62.9	8.8	HIGH
Science	40	30	643	69-6	60.4	8.7	MIDDLE
Social Science	40	22	607	40-5	44.7	8.3	MIDDLE
Listening	40	22	608	35-4	41.9	8.1	MIDDLE
Thinking Skills	190	122	623	56-5	53.2	8.5	MIDDLE
Basic Battery	322	218	NA	57-5	53.6	8.5	MIDDLE
Complete Battery	402	270	NA	56-5	53.4	8.5	MIDDLE

National Grade Percentile Bands *



PERFORMANCE ON CLUSTERS

	Bel Avg	Avg	Abv Avg
Reading Vocabulary		*	
Synonyms		*	
Multiple Meaning Words	*		
Content Clues		*	
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Reading Comprehension		*	
Literary		*	
Informational		*	
Functional		*	
Initial Understanding		*	
Interpretation		*	
Critical Analysis		*	
Strategies		*	
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Mathematics Problems Solving		*	
Number Sense & Operations		*	
Patterns/Relationships/Algebra		*	
Data, Statistics, & Probability			*
Geometry and Measurement		*	
Comunication & representation		*	
Estimation	*		
Mathematical Connections		*	
Reasoning & Problem Solving			*
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Mathematical Proceduree		*	
Computation w ith w hole Numbers			*
Computation w ith fractions		*	
Computation w ith decimals		*	
Symbolic Notation Computation		*	
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Language Mechanics		*	
Capitalization			*
Usage		*	
Punctuation		*	

	Bel Avg	Avg	Abv Avg
Language Expression		*	
Sentence Structure		*	
Prew riting		*	
Content & Organization	*		
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Spelling		*	
Phonetics principles		*	
Structural principles		*	
Homophones		*	

	Bel Avg	Avg	Abv Avg
Science		*	
Life			*
Physical		*	
Nature of Science		*	
Earth		*	
Thinking Skills		*	

	Bel Avg	Avg	Abv Avg
Social Science		*	
History		*	
Geography		*	
Political Science		*	
Economics		*	
Appl of Know ledge		*	
Org, sum, interp information		*	
Determination of cause/effect		*	
Thinking Skills		*	

***** The black bars around the student's scores (black diamond) indicate the area that one can be 95% sure contains the true test score, given that any measurement contains some error.

Scaled Score: the student's reported score
PR-S: percentage of a student's peer group with scores less than or equal to that particular score
NCE: a measure of where student falls on normal curve, indicating a student's rank compared to other students. It ranges from 1-99.

Stanine: nine-point scale for test scores with the mean of 5, and1 being the highest and 9 the lowest.
AAC: where an individual's score on a subtest of an achievement test compares within the range of scores of other students of similar ability as measured by an ability test.
GE: describes the grade level and month where the student's test score would have been average. e.g. 10.5 means the 5th month of 10th grade

The following questions refer to Report #3.

In Report #3 appears a display of information that might be provided for a student by the name of **James** taking a national standardized achievement test. Look at the information and then answer the six questions that follow.

26. According to the report, James might have more trouble with words having more than one meaning compared to two different words that mean the same thing.

    - o True
    - o False
    - o Neither

27. According to the report, James scored higher than average in all clusters of the Spelling subtest.

    - o True
    - o False
    - o Neither

28. According to the report, it might be beneficial for James to work on organizing his essays and reviewing the different ways paragraphs can be organized.

    - o True
    - o False
    - o Neither

29. According to the report, James got fewer points correct in Language Expression than in Language Mechanics.

    - o True
    - o False
    - o Neither

30. I found this graphical display to be useful in informing me about the test results for the student.

    - o Strongly Disagree
    - o Disagree
    - o Neutral
    - o Agree
    - o Strongly Agree

31. How James performed on the different clusters was beneficial information to include in the report.

    - o Strongly Disagree
    - o Disagree
    - o Neutral
    - o Agree
    - o Strongly Agree

The following are general questions for all of the reports that you have looked at today. These questions are solely to help the researcher gather additional information and will not be associated with any particular individual.

32. Of the three reports you looked at, which one would be the **most useful to you**?

    - o Report #1 (Sally)
    - o Report #2 (Ferrus)
    - o Report #3 (James)

    Please explain your choice briefly.

33. Of the three reports you looked at, which one would be the **least useful to you**?

    - o Report #1 (Sally)
    - o Report #2 (Ferrus)
    - o Report #3 (James)

    Please explain your choice briefly.

34. Would you prefer to access score reports **online** or receive **paper** copies?

    - o Online
    - o Paper

    Please explain your choice briefly.

35. Which of the following scores did you find **most useful** in the reports? You may indicate more than one.

    - o Points Possible
    - o Scaled Score
    - o Percentile Rank (PR)
    - o Normal Curve Equivalent (NCE)
    - o Grade Equivalent (GE)
    - o Ability/Achievement Comparison (AAC)

36. Which of the following scores did you find **least useful** in the reports? You may indicate more than one.

    - o Points Possible
    - o Scaled Score
    - o Percentile Rank (PR)
    - o Normal Curve Equivalent (NCE)
    - o Grade Equivalent (GE)
    - o Ability/Achievement Comparison (AAC)

Thank you for taking time to respond to this questionnaire. Please feel free to use the space below or the back of this page to make any additional comments or questions you might have. If you would like a summary of the results of this study or have any further questions, please send an email to: **scorereportsurvey@gmail.com**

APPENDIX B

FOCUS GROUP TRANSCRIPTS

Focus Group #1

Leader: So did you guys have a minute to take a look at it? I would like to just ask a couple of general questions. How do you guys look at reports like this before, sort of like student reports, kind of some of the results of tests and when you are either at home or a kid or a student. You can see from this that it is for Sally Wong, we just made up the name. It has some of the information. She is a second grader. She took it this year and these are the sub-tests. These are some of the different scores it has. Some of those are briefly defined down here. And now, look at these scores. You can see that there is "Number Possible" and "Number Correct" and "PRS". Are there any of these types of scores that you are familiar with from before, like these scales or maybe a PR (percentile rate)?

Group 1: I think that we need to review it just a little bit.

Leader: We have grade equivalent.

Group 3: Grade equivalent I have. So she is best qualified for third grade?

Leader: There are some short descriptions down here.

Group 2: Go over PR first or even go over them all.

Leader: We will go over those in a little bit. It is kind of the stages that we are on.

Group 2: Okay.

Leader: Again, down here is a brief description of what some of these are. PR is a percentage of the student's peer group that have a score less than or equal to the one that they have. And grade equivalent is basically, you know, it has to be basically what the average student has at the grade level. Because I got these for example score reports that came from both standardized tests and tests that we used in some of the programs.

Group 2: So this is your own form then.

Leader: Well, it is a mock-up that I did and it was based on the reports that are distributed out into the field. And there are two purposes here today. One is to sort of, you know, go over and have you guys answer the questions and sort of an interactive sort of way to sort of see how the understanding is and the wants and wishes of what folks are getting reports. The other is to sort of refine this, because this is really a survey which I am going to, you know, basically finalize. I am going to be giving it out. Doing a focus group is a good way of modifying it.

Group 2: This one that you have prepared is for all levels, administrators, teacher and students, or is it just for parents?

Leader: Well, here is what is an individual student report, so it can be used by the parents and by all levels, which typically look at this one.

Group 2: Okay.

Leader: So, are you saying that really parents aren't as familiar with an NCE or a test score or as you can see there is a total of number correct here as well. Over here there is a graph that sort of gives you an idea of how Sally is doing. Let's go ahead and you guys can take a look at the questions on the next page, starting with question number nine. There are five questions with the report and there are some opinion questions obviously.

Group 2: Well it says the above average math, that is the total math right? Not the individual one.

Leader: Yes, that is total math.

Group 3: The nine along this page, are they questions for these all right here? You had changed the font, so I was hoping that was what was going on with the change in the font.

Leader: Yes. How are we doing? What do you guys in general think of this kind of graph report? Do you think that it is sort of in a useful format? Do you think that there is too much data in here? Do you like the graphs?

Group 3: I like the graphs. The only thing that I might do would be to put a line at the fifty, or have something at the bottom so that I don't, you know, when I write I slant. The other ones are a matter of knowledge and familiarity and all of them are very helpful except, for me, the scale scores. That one is not. I don't have any experience with that to say that is helping me.

Leader: I think that parent's and a lot of folks, typically, it sort of depends and also often they, like here, I try to include some additional information and sort of like in a graph the co-representation and you know also the numbers in here. But I guess sometimes it can be a lot of stuff.

Group 2: Well, the question is, are you planning to give exactly the same form for each of the three groups, or are you planning to abbreviate this form?

Leader: Well, I mean, the plan was that I was going to give, (interruption) there was actually multiple versions of this kind of graph that we are going to go through here, but it is sort of looking at what….(interrupted by Group) This is what is typically sent out to parents, or something that is like this.

Group 1: I really see it a little bit complicated. Too complicated. I've have had one of my friends would show me a test report and I'd have to go through and explain it, and it was much, much simpler. Then this one, this one seems very complicated to me.

Leader: Okay

Group 3: I personally like the fact that there are different ways of comparison, how many students are less than, how does it fit, such is the NCE average, how many are less and the percentile rankings. The only thing that I did notice was the AAC and the GE on the bottom part there are not in the linear order of the graph.

Leader: Okay.

Group 3: And that will trip up people. Not that they won't be tripped up some other way.

Group 1: Basically, I like the concept that you are trying to do the comparison of how their grade level, their range and everything. That part is good. But I am still seeing some, you know, it being a little bit complicated.

Group 2: And in the other way, I don't think that you need to dumb down to people.

Group 1: No. Well I don't either.

Leader: I have seen all kinds of reports. This actually looks very similar to a product that I know very well and has one that looks kind of like this. I have seen others that are kind of more of a narrative and you will see some examples here and also the graphs. I don't have any example of it, but I have seen one that looks like a dashboard of a car and it has arrows and stuff like that. In fact it was my sister who is a special ed. teacher who said that different people like to look at stuff in different ways.

Group 1&3: Exactly!

Group 3: For me, the only thing that was useless was the scale scores because I don't know enough to use it.

Group 1: You see, that is the way it is. Like, yeah, the scale, the national the natural. Now it is important for me to know the grade equivalent and you know the range maybe, but the others …..

Group 2: I want to know how much of it is being missed. I also think that we may have to go into what is it the NAEP, one of these things. I want to know also how my child also is shaping up against other children, it is other children of the same age?

Leader: Yes. Let's go ahead and move on to report two. Pretty similar to the other one. It just has the traditional type of wordage again. Take a look at the questions that follow it and see what you think about those questions.

Group 3: The questions are on page 7.

Leader: This is a tenth grader here, Maria.

Group 2: Now as opposed to the one in second grade, was Maria not tested on some of these skills?

Leader: She took the shorter battery. It is marked here.

Group 2: This is what they would see in the upper grades. See, all I know is elementary.

Leader: This is just an example.

Group 1: Now what stands out to me is this explanation. I think this is helpful to parents because this also would tell parents how their child is doing. That explanation is good. I like that.

Group 2: Are you trying to come up with one test for all levels? Or are you trying to do different tests for different levels? What you are trying to do for your work.

Leader: You know, it is basically looking at score reports and what is useful about it, what do people find informative, not informative, basically. And as you can see, they are looking at different ones……

Group 2: Because even though it has an explanation for these, I don't think for administrators, it doesn't go into some of these listening and thinking skills that the elementary one does. But maybe they don't test them on that in high school. I don't know. Reading, math and language seems a little brief if they are also testing them in high school on social studies and science.

Leader: It could be that if it is the state testing program, maybe those are the three that they chose to only include.

Group 2: And if your child needs tutoring in Science, wouldn't you want to know? Or would you need to develop programs for teachers in social studies. Depending on what tests they give in each state. Don't they give science and social studies in Texas? I think that they do.

Leader: Yes, I believe that they do. This could be a different state. It is a subject sort of like Texas, but whatever the program has.

Group 2: In other words, the research report should match what the state tests, I guess is my question.

Leader: Right, we are just going to say that this is what this state is doing.

Group 1: But basically what you are doing is trying to see which kind of a report looks good, is a good type of report.

Leader: I am trying to find out what is informative or not. That is why you guys are here.

Group 1: That is why I jumped at this, because I said, this is what parents would want to know. About the student scores. I think that would be a good explanation to the parents.

Leader: Have you guys had a chance to look at the questions.

Group 3: No, not yet.

Leader: Have you guys…

Group 3: I am ready.

Leader: Maybe we should move on. Is there anything else you would like to say about report 2?

Group 3: Yes, I just liked the fact that it doesn't have on the graph the vertical bars.

Group 1: The black bars there. That falls in that range where the…

Group 3: The bars on the individual, like these have individual, those things, the triangles… where the actual child is in relation to everyone else. That is not communicating that information. And I downgraded the survey. Before I said it was strongly rated, now I am down to neutral on the first one, because I don't know where the student is. I know where the world is, but not where the student is.

Group 1: All in this range or in that range.

Group 3: I want to see my kid in the top of the range.

Group 2: I like the way the first subset things are a little bit thicker on the top. The reading is thicker, the total math is thicker. The language is thicker.

Group 3: I definitely agree with the comment about how to evaluate the scores, the synopsis and wording which is in fact another way to say the same thing.

Group 1: I think that that is a lot more meaningful to the parents too.

Group 3: There is less interpretation. It is a statement of what is happening as opposed to seeing the data that you have been given.

Group 1: Because some parents do not know how to interpret the data. So if they have this all, then …

Group 2: Then they can schedule conferences with the counselor.

Leader: So number three. You have the reports, then you have a lot more explanations of the terms. Okay, it sort of tries to go beyond the little stuff that is on there.

Group 3: I keep having the question of number possible of number correct. Are they graded for correct answers or for guessing one?

Group 1: To me this is overload.

Leader: Overload? Okay. Yeah, I think that it is trying to strike a balance between not being too little, too much and ….

Group 3: And it is all going to change day to day. When we get through, I will tell you a great story about maps that I read. I have to agree about overload. If there is a counselor, they have to elaborate about that. Of course, if there isn't a counselor, they may want to know that. They you get the question of how much paper do you want to use. Now do we go ahead and do the evaluation.

Leader: Yes, go ahead and put the questions twenty-five.

Group 2: I think it is my problem with my sight that I have to put a line under each line.

Group 3: That is what I do too. I put some lines.

Group 1: On our stat sheets, they are so tiny, I have to….

Group 2: Now this one here is better than the others because it has the notch where the student is.

Group 3: Yes, this one here did not have the notch, and I downgraded it because it did not provide the data on my kid.

Group 2: So even though it looks jumbled, it really has the information that you need.

Group 1: Oh yeah.

Leader: Are you guys ready to talk about this?

Group 1: I am just still saying that that particular page, parents would probably not understand it. I mean, it's too much for the parents.

Leader: Too much information?

Group1: Yeah. They need more simplistic.

Leader: What did you think about the, there are a couple of examples within each one of those that… are examples helpful?

Group 1: Uh huh.

Leader: Like we were saying grade 6 and math….

Group 1: Well, if I was handed the paper, I would not want to read through all of that. I mean, to be honest……

Group 2: But not every parent feels that way.

Leader: But is it useful to have just in case? Do you think that someone might need it?

Group1: Yes…..

Group 2: Go ahead and finish.

Group 1: I mean, it would be useful, but, I mean if they wanted to have it, but that, it would discourage, I mean, I would look at it like…..

Group 2: Is that supposed to go along with this report.

Leader: Yes. So maybe a balance between…. Some of you guys are saying that you thought that this was too much on the explanation of terms, but then, what do you think about on the bottom of each one that I sort of had a very brief, brief explanation. Is that enough, too little?

Group 1: Yes, yes.

Group 3: I can finally read and understand what these grade equivalent .7 and .2 are by reading this thing, which I did not understand through the first two. I read what the scale scores mean and I again will say that it is useless to me.

Group 2: But it might not be useless to somebody else.

Group 3: Right. But I am able to understand details, so. Having the page, and probably having this simple thing if one wants to know more, then they could. I am still trying to figure out the PR-S and the second number means.

Leader: Maybe a sort of a variety of resources and then also there could be the school counselor or a testing person or another resource.

Group 3: I would say that this definitely needs to be included, but for most people, I would say that they would also want just a simple thing. And if I want to know more, then I look at this. To read this to try to figure out that would be, I have to agree, overwhelming. In most cases, I am not interested. And again, now that it has the notches on where the child is, it describes much more useful. Where as before it did not tell me anything that I wanted to know.

Group 2: I just don't want the idea of dumbing down. You know, I think that we have done an awful lot of it in you know our insurance policies, in getting our directions for getting our new drivers license and everything. You get to such a lowest common denominator that it really isn't ….(interrupted) this is for a range of people, it is not just for one.

Group 3: I know.

Group 1: But I am looking for an average, working person and ….

Group 2: But why, that is only one of the groups of people…

Group 1: That is true. I mean you have all kinds of….

Leader: What I have been trying to do is get a variety of groups together, so.

Group 2: You would get very educated people that would want all of that, that would know and vary.

Group 3: My first question was why is it 2.1. What is the ".1"? I did not know what the 2.1 was until I read, oh second grade first month.

Leader: Speaking of the grade equivalent, I think that is a part of it, because I know it can be an often misunderstood kind of a score sometimes. Is that something that you guys think is a good kind of a score to include?

Group 1: No, I think that it is very good. The grade equivalent is very useful.

Leader: I just know, I think that someone who is close, near and dear to me heart, my mother. Because I remember that I had to take a test and the grade equivalent was like a 12.9 and you know I was like in the eighth grade and she was like, oh, you are doing college work! But, you know, there is more to it. So with part of these I am trying to get some of the questions alluding to common misunderstandings that folks have about certain scores and trying to get to all of them.

Group 2: So you cherry pick from this chart which ones you want.

Leader: Yes, for the questions.

Group 3: And then of course, there is also the factor of what we talked about the second grader or the fifth grader.

Leader: Well is there anything else that you guys want to say about this report number three and the explanation of terms?

Group 3: It's not necessary, and I have already said that. Simplified, the dumbed down version.

Group 1: The dumbed down version, this has to be something that goes to the whole school district, doesn't it?

Leader: Yes Ma'am.

Group 3: It's like I was telling a guy who was a foreign exchange bank trader making millions per year for himself and for the bank. And I said, put your columns in and if this the date that I cannot work with this information that you are giving me, I just don't want to see any more. He was putting it in a different order. And I said, then put the time. And if I have a dentist's appointment then, I can't use it. So I don't want to see the rest.

Group 2: But that is you as opposed to all of the people, bank officers in the bank as opposed to all of the branches of the bank and ….

Group 3: They did not know how to do it, and I knew how to do it.

Leader: Let's go ahead and look at the last report.

Group 1: Now this one to me, what jumps out at me is the bottom part of it that is showing the performance in clusters. I like that. If I remember right, the PRAXIS is a lot like that. It may be similar to that, I am not sure. I like this too. That is what jumps out so that the parent at a glance can look at it.

Group 3: Key word, at a glance.

Group 1: The name is James, eighth grade.

Group 2: This is on the state test?

Leader: It could be an example of a national test as well.

Group 3: So do we go ahead to page twelve or page thirteen. Yeah, I think that I have shared everything that would be for the report types, so we can move on.

Group 3: Well, I have to say that I am like those people that can talk about the science fiction novels and the good old hard copy.

Group 1: Yeah I do too. I want the copy. I do not want to go on-line.

Group 3: Part of that may be because of the quality of screen.
Group 1: Well, you can always print it. But I want it handed to me.

Group 3: Well, I am just saying that an average cathode ray tube is not very eye friendly.

Leader: As far as like an on-line report, they have some that I have seen, that folks have produced, that are like, first you get a graph, then you can click on it. You can scroll down and sort of get the level of detail that you want. If you want to look at maybe a, you know, a certain district, or a, you know, a group of kids, or something like that. Is that?

Group 1: No. I don't like to mess with stuff like that. I am not a, I am not technically challenged. I do not want to deal with computers that much.

Group 2: Well, we are old. We are not part of the generation that is on-line.

Group 1: But personally, I like the last one with the …

Leader: This one right here?

Group 1: Yes. That one to me was the most friendly.

Leader: So having the clusters and then just sort of looking at the case, if they are below average, or ….

Group 1: Uh huh. But you are still maintaining the stats. So you still have those, so. In the graph there beside, so.

Group 3: Yeah, that is why I just wanted the whole copy, the hard copy. Because the screens were never good enough for what I wanted in my eyes. That has been changing now. So….

Leader: So it is definitely useful to have this sort of more specific information so that like, you know like what your kid needs to look at or you know….

Group 3: It gives a focused target group.

Leader: Yeah. Like average, above average and then there are some items….

Group 1: Right. So I like that information.

Group 3: You are trying to trick me on question number 46.

Leader: I was in another state and there are more categories that they put.

Group 1: So really, that makes a big difference to that?

Leader: There was the building that I was in for Grad School. There was the men's room the female and the gender neutral. Anybody could go in that one.

Group 3: You could have a unisex bathroom like in my church. Mainly for baby changing.

Group 2: Yeah, they do that in movie theaters too. It says diaper changing though. It does not say mother's only.

Group 3: Not that I did not take every opportunity I could to not change diapers.

Group 2: Yeah, but you are old. The younger generation just does it.

Leader: I know that we are wrapping up. But if there are any particular questions that you thought were problematic, or were not clear, just circle them.

Group 1: I think that I go all of my questions wrong with this SE.

Group 3: I would like to throw in one thing. On some of these evaluations, in that, when I took a test that said what careers are you going to be in when I was in night school, (interruption) they gave me two things, a computer programmer and an Army officer, where I was "outstanding". I found out later that the fact that they said Army officer was a big change in my life because I detested the Army. I would have been a Naval officer. So that if they had said a military officer…they have to watch the wording on what you are presenting. I did not want to be in the Army. The Navy fine, Army no. I did not see that apply to anything in here yet. That was the first time I ever said how pissed I was when I figured out what they had done to me. I will reiterate what I said about on-line versus hard copy. With our plasma screens now, you can count the hairs out of place. It is not going to become an issue anymore for me. What I see was a consolidation of reports one with the explanation. One of them, number three did not have any explanation. Neither did this one, so you need two explanations. That one presented the data better. The explanation needs to have simplified and the enhanced complete. It did not have any explanations.

Group 1: It would be nice to fix up sort of a hybrid to have the clusters and a little bit about the scores.

Group 3: It could all be done on one page. Simple explanations on top and here.

Leader: Okay. Do you think having those parts, typically on the bottom helped answer the questions? Some of the types of graphs….. as the graphs had more of that information, it helped answer the questions. Like either the narrative at the bottom or the clusters.

Group 1: Yes. But it would be nice to have a hybrid with the clusters and the narrative.

Leader: Anything else that you guys would like to say.

Group 3: Thank you.

Leader: There is some contact info that you should, I think that you already have that on the consent form.

Group 1: So approximately when do you think that you might have it all together?

Leader: It is not going to be for a little while.

Group 1: Okay, I mean, like January? I just mean a kind of time frame.

Leader: I am thinking January.

Group 3: I am glad that someone is making a report on how to make reports.

Group 1: Well, you know, a lot of people just do not understand them like I have a niece that I had to sit down and I had to tell her, okay, your sons scores are this. And this is what it means. So that is the perspective that I am coming from, is, a lot of parents don't.

Leader: Well, I guess that is it.

Focus Group #2

Group 1: What do you mean by training and educational testing?

Leader: Well, let's see, the options are maybe a workshop that the school gave them, or is conducting it or giving it or some info on that. It could be that some of the districts have a research person, I don't know if that person comes and does that. They usually put like a teacher or maybe like a principal that would often take the test measurement class.

Group 2: You should have warned us that we were going to have a flashback.

Group 4: We did enough quality testing back in the seventies. I don't think that was MAT8.

Group 3: I put MAT8 and I put Stanford.

Leader: Go ahead and put those down.

Group 4: Okay

Group 2: By the reading subtests, you are talking about the total reading?

Leader: Right. We will get started and we will look at that guy.

Group 2: Oh, I thought that we were supposed to be filling it out. I'm sorry.

Leader: That's fine.

Group 2: I misunderstood what you said a while ago. I didn't think so at first and then I did it.

Leader: Has everybody had a chance to fill out the consent form? Kept a copy and gave one to me? And you have a copy of the survey? Let's go ahead and get started. Okay, I am S---, a psychometrician here at Pearson and this is my colleague, K--. Well I guess the question is, why are we here? And it is really to do this focus group and its going to be on score reports. And really, what we are going to do is we are going to basically look at four mock score reports that I created and we are going to, you know, they are kind of representative of different states, but not really any particular state. So, you think I could be a typical state you might have. And really, we are going to take a look at each one of those and sort of, I am going to ask your opinions on them, what you like, what you don't like, what you think is useful, not useful. Then you are going to answer some questions about them afterwards. Some are informational, where you are sort of interpreting results. The others are really for your opinion about the score reports. We are going to sort of walk through each one and give you some time to look at the questions. We should probably spend about fifteen minutes for each score report, or roughly, and leave some

time for discussion. So we will try to keep that pace and see how it goes, depending on the time. If we don't have more discussion then we will stop. This is going to be for my dissertation. That is what the results are going to be for. And it will also probably be for a research report and then also a publication. And any questions?

Group 1: What are you doing your dissertation on, why are you doing a dissertation, for what?

Leader: It is psychometrics. Testing, test and measurement. I think that my diploma would actually say research and evaluation methods. They are all, depending on the school, the same thing.

Group 4: Do you also involve students in this, in your focus groups too? Like maybe high school age. I don't know if any of these pertain to that.

Leader: The reports themselves or the folks participating?

Group 4: I know that they are made up, but...

Leader: Well, these right here, there is going to be like elementary level, and I think a junior high. Most of these reports have basically the same information. We are going to be adding additional stuff. When I am doing my survey I will give it to high school kids too. So today we are going to have two focuses. We are going to give our opinions of these reports and then also it is also going to help with the survey, which is going to be another part of the study. Okay, why don't we just briefly say your first name and maybe something brief about yourself.

Group 3: My name is G--- and I checked here as an administrator, but I am a father of four adult children on the north side and seven grandchildren on the north side of B---.

Group 2: I was an administrator for the --- school district and have two children that went through north side schools. They both graduated from J--- High School. And I have a granddaughter that is in kindergarten in a --- school this fall. So she is the first of the next generation on north side.

Group 1: My name is S---, I have one son who is ten and in the fourth grade. And he has had speech problems and has been in school since he was three for speech. That is one thing that I like about San Antonio, is that they have speech programs for kids that start at three years old. At no cost to parents, which is always nice.

Group 4: I am E--- and I taught for thirty-three years in the S--- school district. I have one child who went through A--- school district and graduated. And I now have a grandson that is also in A--- school district in a Spanish emersion program in fifth grade.

Leader: It sounds good. We have a lot of experience and that is what we want here is to take advantage of that. Okay, let's start with the first one. I am going to give you guys a

copy of all of these, because I know that it is pretty small to read up here. So you have a copy of all of these reports. But basically this one is sort of a prototypical one that I kind of created and it is for Sally Wong. There is a little bit of info up here. It looks like she took the test this year and she is a ninth grader. And these are the sub-tests that she took.

Group 1, 2 and 3: I thought that she was second grade.

Leader: Okay, sorry, she is in second grade, and she is nine years old. These are some of the tests that she took and some of the different scores that she has and there is also some of the graphical representations of that and down here on the bottom, sort of some brief definitions of some of these scores right here. And taking a look at this example report, are there any of these scores that maybe you guys are familiar with or have seen before? Like the scale score, or maybe a percentile range or grade equivalent or any of these?

Group 3: Yeah, basically all of them except this looks more like a norm reference, this especially over here.

Group 4: I had not seen this AAC range before. This is kind of new to me and I think that helps that, and I thought that was kind of interesting.

Leader: Yeah, that one is more of a, it could be more typical like a Stanford 10 or a Metropolitan 8 type. Maybe some of the scores are familiar types, but not necessarily all of them. Some of them you may have seen before and others maybe not. Did you say that AAC is not one that folks might not be as familiar with, but any of the other scores that you guys have maybe seen before or maybe used. The scale score?

Group All: Yes

Leader: Okay, what about percentile rank?

Group 4: That has been on before.

Leader: Okay. Grade equivalents?

Group All: Oh yes.

Group 2: The grade equivalent in my experience was the one that parents seemed to misunderstand most often because they would say, well if she is in fifth grade and this means that she is working on a seventh grade level…and it doesn't. It means that she did as well on this test as a seventh grader would. Sometimes they wanted to put pressure to get you to elevate them up to a higher group or something. And sometimes it was appropriate, but it wasn't always.

Group 3: In context, I think that we more than at the class administrative level more than the classroom level. Of course we would usually have someone who was responsible for administering these tests, like the school counselor. But the thing that we would get most

often would be a report of the whole class and it was used an awful lot in terms of, we would get like a summary of all of the kids in that room and it would more or less show us, you know, basically we could see if they needed to bump a kid up from a year before or better and stuff like that. Unfortunately it was used a lot to see in how the teacher effectiveness.

Group 4: And also we would use them, you know, like you said, checking on student progress, but if they were weak in a certain area too then that's, you were trying to pull them through together for a group for the following year to really begin working on that one weak area, or whatever area was weak.

Group 3: And when we would do the decision making and got all of the academic themes and stuff, we would actually use these scores sometimes to group students for, quite like modified job programs and stuff like that, or to just put them in a reading group to try to meet needs, you know even sometimes across grade levels.

Leader: S---, is this, does this look like a typical report maybe that you have seen?

Group 1: No. Maybe how many problems they have and how many they have gotten correct. But like a scaled score are not ones that I am used to dealing with on things that my son brings home. Um, and the number of what the PRS I think it is percentile. Usually they don't usually say that you have that many people less or above your child, not really equal to, less or above. Because I mean, you could have so many that are equal and you really don't know if they are right on mark or if there is more or less below them.

Leader: Okay.

Group 4: But I know I noticed one thing on my grandson's TAKS last year. It has just a line you know that if he continues on this track, then next year he should have no difficulty passing the fifth grade one. Which I thought, hmm, that's an interesting prediction.

Leader: Okay.

Group 1, 3: I hadn't seen that.

Leader: We are going to go through other reports, but is there anything, kind of, that you like or maybe don't like about this? We are going to put through several others, but sort of…

Group 1: I don't like the scale score, because it doesn't tell me scaled to what.

Leader: Okay.

Group 3: From the perspective, I think that the varies that you see over here on these national grade percentile bands, I think that the variants that you see in those in terms of

the discipline, like you know determining the math and things like that, that they are a little atypical . I don't know if they designed them that way or not but they seem…

Leader: What do you mean by atypical?

Group 3: Well, for example in the reading vocabulary, and I guess in the total reading overall, you know going from the vocabulary and the reading comprehension and the word study skills, going from the thirty to a ninety on this would be different indicators is not real typical of what you might see.

Group 2: It seems like the range is really odd. That is what I thought when I first glanced at it.

Group 3: Yeah, the range are kind of off.

Leader: The range right here or just the scores?

Group 2: No the way that it is represented there on the, I mean like, 35 is pretty low and 70 is pretty high. And I am not sure whether that, on total reading, it starts at below the 35 and goes up to about 71 or something. And I don't know whether that would be very meaningful to a person. It just seems…you might expect this child to perform anywhere from pretty low to pretty high according to that. I don't know if that is realistic or not.

Leader: Do you think that this confidence band is useful then or …

Group 2: I do but I am wondering if it ought to be that broad.

Leader: I was trying to make it realistic as possible.

Group 3: The other thing that I thought kind of stood out in my mind is global. Is, I thought that the correlation because usually if they do that well in reading you would expect a higher percentile rank in let's say for example language. Because those two are usually language and reading are more closely associated.

Group 1: My problem with the national grade percentile bands is all of the states have different exams for the students at different grades. You know each state has their own exam and who's to say that one exam is better than the other and easier for students to do. So I think that kind of skews the report.

Leader: Well this is, like I say, trying to be an example student, you know. And this is mock data that I made up.

Group 2: What kind of test is this, this looks like just a regular achievement test. Norm referenced.

Leader: Right. So anything else on this report. Let's go ahead and have you guys take a look at these questions that follow that report. I guess nine through sixteen. They are all found one page back. They all refer to this report number one. I just put them so that you can have them both and take a look at them. We will take a couple of minutes for you guys to go through those questions.

Group 2: This says on this spelling question above average for students at the same ability level. Where is the ability level indicated?

Leader: Well, that is part of the question.

Group 2: In other words, that is part of the question. Okay I get you.

Group 3: I had an interesting experience being that I was a school administrator and had a friend through the church that I thought was as well or better educated than me that came to me with all of these concerns about his daughter's achievement tests. This was back in the days before the state testing programs. It was like a Stanford student test. And really, she was way above average, but this guy didn't think so from what she brought home. It was revealing to me.

Leader: Are you done looking at those questions? Just quickly with the questions. Are there any particular questions that maybe stood out as, you know…

Group 2: You mean the questions on your survey form?

Leader: Yeah, questions nine through sixteen so far….

Group 3: Yeah what really kind of stands out… what was hard for me to answer, particularly number eleven. Sally scored significantly better in word study skills. You know because she is better in one and not in the other. You know that was kind of….

Group 2: I think that is the appropriate answer needed.

Leader: Okay.

Group 3: And then there was another one with two components to it, let's see, something about being better….according to the report, Sally has the same proficiency of word study skills and knowledge as a third grader. That one, it was like, the way I read it was like she had a 3.1 is like when the scores are 2.9, 3.0 and 3.1, you know if it said… I went ahead and said that was True, but you know what I was thinking was, you know if it would have been 3.8 I would have had no hesitation.

Group 1: Yeah.

Group 3: The fact that it was only 3.1, that wasn't very simple.

Group 2: The test was given in April. So she wasn't much of a "grade level".

Group 4: No, I didn't think that she was a grade level at that time. A third grader.

Group 3: Let's see, when is her birthday? Well you said age nine, so she is an older second grader.

Group 2: Oh that is right, she would normally turn eight here in second grade, not nine. So by age, if in some of them they do report by age, she would not be very high at all.

Leader: Okay. So with that particular question, I was looking more for like what the grade equivalent and maybe like we had talked about earlier. It reminds me of when I took a test and I was in ninth grade. I got like a 12.9 grade equivalent. And my Mom was like, oh that means that you have the same, you are doing the work of a twelfth grader.

Group 3: Well you are! (laughing)

Leader: That is what I was kind of talking about with that particular question. So maybe I need to tweak it a little bit so that…

Group 3: Yeah, this thing about the age too. I mean, I am sitting here and I would be in fourth grade when I was nine years old.

Group 2: Oh yeah, most fourth graders are nine. My granddaughter was nine in fourth grade.

Leader: So that should be eight years old then.

Group 1: Well my son was ten and in the fourth grade, but he has also repeated the first grade, so technically he would have been nine in the fourth grade.

Group 2: Well and by April most of them, well more than 50% of them are 8 by April.

Leader: So let's go ahead and move on to the next one.

Group 3: So what are we doing now?

Group 4: Looking at the second report.

Leader: Looking at the second report and taking a minute to look at it and then look at the questions that… questions seventeen through twenty-four.

Group 3: Now this thing that you are about to see the scores. You say that this is part of the report that would maybe go home with the parents when they come back to the school. Correct?

Leader: Correct. You can see that it is similar to the other report but it has some additional information.

Group 3: Oh, it is better.

Leader: I will ask you guys about that.

Group 3: Are we making any distinctions here as to who this is going to be read by or that it going to be read by us the educators or whether it is read by parents when it gets home, or both or what?

Leader: I am trying to get diverse views. That is why I try to get parents and administrators and teachers. And I know that a lot of you guys have multiple perspectives. So if you, if there is something that you know would be useful for the conversation...

Group 3: Just based on this right here, principals would love that.

Group 1: Parents too.

Group 3: Principals and counselors and parents and teacher. We spend so much time giving this kind of, trying to do this kind of data and explaining this kind of data to parents. I mean, something like that where there was a computer program that could get this kind of data out would be wonderful. It would make our jobs so much easier.

Leader: So about the student scores on the bottom of the form too.

Group 3: Well, the explanation of it.

Group 1: And your question does say just the graphical display. Are you considering the graphical display also with what they write?

Leader: I am saying the whole score report.

Group 1: That is a lot easier than going through the numbers. Having a summary like that. Are we allowed to write notes next to our answers on the questions?

Leader: Yeah, I am going to go ahead and collect these afterwards.

Group 1: Well, I mean, write notes to y'all. Why we think that it is easier to read.

Leader: Sure. So are we ready to talk about this report number two? So I think from what you guys were talking about just now, you like this about the student scores on the bottom. And……

Group 2: I have become accustomed to sending that kind of thing home before I left public education. And it is a big help for parents to have that because we didn't always do

that. We didn't always make those kind of explanations and when we were reporting nothing but grade equivalents if I recall at one time, that is all that we were reporting. I thought that it could be pretty misleading or it could make the parent feel a little over secure about how well their child was doing because they would say, oh my goodness, he reads as well as a twelfth grader. Well I don't think that is the way to interpret that. He did that well on this test but that does not necessarily mean that you are going to be able to have a hand in a novel or something that is written for graduating seniors and have him do as good a job on it as they would.

Leader: Yeah, we will talk about the grade equivalent a little bit more. There will be more information about that. Do you like having both the numbers and the graphical? I think that they are different ways. I have seen sort of both extremes. I have seen some older reports where I call it just data dump. It was just tons of numbers. And then I have even seen, I guess one of the newer reports that was really almost completely graphical. I think that for the scores it looked like a speedometer and it just had a rotating needle. You know whether the kid was proficient or not. I have seen reports that were like that. So.

Group 1: I think that just with all of the numbers too, a lot of parents aren't used to statistical information and trying to evaluate their child compared to that you know and a lot of them don't want to ask questions because they don't understand it. I have been lucky enough. I found out a lot about it because my son had to have IAP's every year. I go in for you know two to three meetings a year that when I had reports and different testing, I could sit there and pick all of the peoples brains who gave him the test and who evaluated what does this mean and how does that affect my son and things like that. That makes a difference. So, having the summaries really helps.

Group 3: One of my grandchildren who is multiple handicapped and is hearing impaired and language impaired and a number of things. And his mother grew up on the north side and is the senior vice president of USAA with six thousand employees. And she takes me in and she still takes me in to help her on those ARD's to make sure that every time that she gets reports, she gives them to me to try to help her understand what is going on. So I am saying that these kind of narratives and stuff like that break it all down.

Group 2: I found that if we are not careful, I would sit in ARD meetings with parents and the teachers would start talking about IEP's and XYZ's and PDQ's and I would say, and that means…. and the teacher would say, oh yeah, I should have told you that. And not that I did a great job with that. I mean just that we sometimes talk over people's heads without even realizing it. Just like the computer people I have been meeting with the last few weeks because I have a new computer and I don't understand what they are telling me. Anyway, the situation that I was referring to a while ago; I had a good friend who was an engineer at southwest research; I think he had a masters degree. He had a daughter in like fifth grade or something like that. He came to me and said, she just did terrible on these achievement tests. I said, well, I am surprised. What are you talking about? Well, she only scored 80% percentile in reading. And I said, well, what does that mean to you? Well, 80, that is 80%, that's a B. I said, no, that means that 80 kids out of

100 scored lower than she did. Oh, he said, that sounds pretty good. So you'd think that a person with that kind of educational background would already know some of the things that he did not and that was kind of a wake-up call for me in making sure that we explain things to people.

Group 3: I know the word co-morbidity and that was like well, I said, you know like co-existence.

Leader: Yeah, we will talk about those points. They are good points and sort of leads into what we are doing here. All right, what about, I put these brief definitions on the bottom of each page and actually on the next report they are going to be more in depth. But are they, is it useful to have those? Actually do you really you need to have those there. Do you think that most parents or students might look at those, or administrators?

Group 1: Even if they look at them they wouldn't know necessarily how to interpret them.

Group 2: You said the scale score is one that I don't think, I doubt that 10% of the people would understand. And I am in those 90% that don't. Scale scores, I never have had them explained to me but I have a hard time. I am sitting here blushing because I am trying to remember.

Group 3: I am sitting here blushing because I am trying to remember what…..

Group 4: I think that it has something to do with the weight of each question.

Group 3: Okay. That is what I thought.

Leader: So we can go back afterwards, when I try to touch on each report if we can.

Group 2: The student report is on a converted scale. Converted from what to what and why? (Laughing)

Group 1: Yeah, that is mine. Like what is the point if they don't know why or what it is, or.

Group 2: I really think that all of these explanation of terms that we are …

Group 4: I think that it is good down here because we had a positive person that was followed by you know it was weak in these areas. So that was good.

Group 3: Yeah, that was well, pretty well designed report.

Group 2: Oh I see what you are saying. Actually, I think that those things are almost more useful to teachers and administrators and I mean you know….

Group1: And more useful to parents.

Leader: It is really hard to come up with a, you know, trying to make it as simple as possible a short definition for these.

Group 2: They are kind of complex.

Group 4: Yes.

Leader: Any other thoughts in mind at this point for report number two? Okay. Let's go ahead and take a look at number three and take a look at it and start the questions. Number twenty-five through thirty-two. So we have that research report on page 8 and then there is the explanation of terms.

Group 2: Now this sheet, the explanation of terms, will be sent with this if it was going home or be put with the record or whatever.

Leader: Yes.

Group 2: Okay.

Group 3: How is it that none of these seem to have birthdates on them. Is that not usually a part of your report anymore? We usually always had date of birth.

Leader: You can put it on there.

Group 3: And sometimes when we have analyzed tests, especially for a child that has been retained a year and maybe was almost seven when he started first grade in the beginning, and so he is now nine years old and in the first grade. You compare him with his own grade equivalency and he is okay. You compare him with his peers in terms of age and he is way down.

Leader: Can we go ahead and talk about this one, number three?

Group 2: You know the thing that I have to say, this is just incidental, sort of a random, abstract side point, is in terms of principals used and or so on of all of this kind of data. Depending on the, you know we have often said that elementary teachers were generalists, were as junior high schools you've got one that specializes in this and that and the other. Your elementary teachers are really, have to be content generalists in all of these areas of social science and math and so on. And likewise, many times, the principal in his strength may be in something other than interpreting the data. You know what I am saying. He may be really good on discipline and teacher motivation. But he has delegated the whole interpretation of this stuff to a counselor or to a lead teacher or something like that.

Group 1: Actually, the explanation of terms for the grade equivalent, they did a really good explanation I would say. But um, between the number possible and the number correct, by just looking at the graphs without those explanations, I would have looked at it as those are the possible number of questions and how many of number questions they

got right, instead of for number possible it says the number of raw score points. So some questions could be worth more points than others. We don't which, which kind, how they are weighted. So that totally screwed up my whole understanding of those.

Leader: Okay. So what did you guys think about these explanations right here on this explanation of terms page? Did you, was it useful? Because most of them tried to have an example, a specific example. Was that useful in helping to explain it?

Group 1: For grade equivalents, I think so.

Leader: All right.

Group 2: I didn't read them word for word but, they did a pretty good job I would say.

Group 4: I like it better on one page. Right now, you know, maybe if you are sending this home for the first time, like with first grade. But year after year they are not going to read that over and over. And normally, I would hang on to everything, you know, as a parent. Now I am talking as a parent.

Group 2: I agree with that, but this is quite a bit more detailed.

Group 1: With all of the things that kids bring home, I can't hold onto everything.

Group 4: Well test scores, you know, they are usually all stapled together. That's another thing. They are all stapled together and you are going to flip them back and forth like this.

Group 2: I am thinking in like an urban district like S--- ISD or C--- in B--- or something like that. If these women are Hispanic, you are going to have to do an awful lot of sitting and translating and trying to explain all of this to them. This is going to be totally lost on them.

Leader: Well, something like this would probably have, I know I worked in C--- it was in, I don't know, twenty different languages, everything that went home from. Anyway, it was in Russian, in whatever. What about the level of detail of this? Like these explanations on the explanation of terms page, versus at the bottom of most of the other graphs that we were talking about, like the scale score and those shorter definitions.

Group 4: As a parent, I would definitely prefer this.

Leader: So the longer explanation.

Group 4: Yes, because, you know the short definitions, if you are used to dealing with information like that, you can understand it. These, for the lay person, much easier to understand.

Group 2: I would think something like this from the standpoint of cost benefits from how much more is going to be involved. It could almost be like if I were going to have a choice of if I was going to waste that paper, I would say that there is a detailed explanation available from your teacher or counselor.

Group 4: Okay, but what if a parent looks at just the graph and they are like, oh, they think that their student is doing really well because they misinterpreted it. And they have something like this and they interpret it correctly they see that their student is not doing well and they can help them improve.

Group 2: Let me explain what the problem is in an urban district. You're not, you know, all parents love their kids. They want the best for their kids and all of that. You are an intelligent, concerned parent. Many of them are so overwhelmed by stuff like this that they just feel absolutely lost. They are embarrassed or even afraid to ask questions, because it is just so totally….

Group 4: Well that is why it would be good to send this home then.

Group 1: Exactly. Because this would help them understand it better.

Group 2: But I am saying, if they ask for it or something like that.

Group 1: But if they had this along with a summary, like we had on the second report that would be really nice. Because then I would understand more of what was going on with my child instead of just throwing a graph out at me and saying, here interpret it. You know?

Leader: Okay.

Group 2: Well, like I said, I was just thinking it might be a little, we have to be really careful in inner city urban districts.

Group 1: With all of the junk that they send home and all of the cost they waste by printing things that mean nothing. I would rather have something that is going to help me understand tests you'll are making my child take.

Group 2: What I am saying is that what you feel. But something like that to a lot of parents would be so overwhelming.

Group 4: We had a lot of young parents when I was teaching and they had a very difficult time just understanding that their child was not making fifty percent. That was not a grade, you know? That was another thing. Yeah.

Group 2: I mentioned that with the guy who had the engineering degree. He didn't understand either.

Group 3: Yeah, see that is what I am saying. And what we are saying is that this kind of gap data is just overwhelming and we don't want to get it to the point where the parents feel unenpowered, I guess is the word I am looking for. We have to be careful that we don't alienate the parents by overkill on data. Because there is a certain point where they will just go….

Group 4: Well that is what happens a lot of times when you just throw a graph at them that just don't understand. And if you have an explanation, they can at least try.

Group 1: Yeah.

Group 3: Oh, I am not knocking this, I don't mean that.

Group 4: They do need a legend on the graph though that shows the tic marks or the student within the band. I just had to ask. I didn't know that.

Group 2: Well l am trying to remember. Most of the ones that we did in the last several years I was there was more like the second one which didn't have that in it. I tend to prefer this. However, when you really think about it, then that may be all that they look at. They may not look at the possibility that there is a range. I don't know. I know that when I was in the school where we did a pile of reporting to parents every six weeks. And we used a check-list. And it was just, it took hours for the teachers to do the check-list. And our intent was that we were going to give the parents more information than we did with just an A, B, or C. They would say, but all of this is not telling me what he is doing. Does he have an A or does he have a B? Well it's pretty hard for us to use our grading system to equate to an A or a B and it was very frustrating. So in the end, low and behold, my son was in one of the schools that was in the project, in that same reporting project. He was there from kindergarten through third grade. And he always brought home all of these great reports. We transferred him to a private school, not because of being dissatisfied with the public school, but we put him in a private school in fourth grade until he was a senior. He brought home the first six weeks, he had a couple of B's and he was absolutely just you know, he was down on the floor. How could I possibly make a B. And I explained to him that, well with this system, they keep all of the grades and if you have a couple of papers that weren't turned in, they are going to hurt your grade. It might not hurt what you know unfortunately, but that is going to hurt your grade. Oh you mean, OK, and after that he made A's.

Group 2: Some parents, quite frankly, are only concerned if they are passing or if they are failing. They don't want any….

Group 3: They will just say, is he passing or is he failing? And then, when my kids were in high school, in honors classes and all, there were parents in there that the only thing that they were interested in was is it all A's or not. If it is not all A's. If my daughter doesn't get all A's that is a reflection on me. So she is going to produce all A's or else… and my daughter would come home and say, I am so thankful that you don't put me under all of that pressure. If I get an A- this reporting period, well I got an A+, she said,

you wouldn't be mad about it. No, I said and I never will. You would be surprised that most of the Dads, especially in our neighborhood; if they went from a A+ to an A-, they went on restriction.

Group 1: Oh, that is how I grew up. My father, I got money for A's. B's I wouldn't get anything. Anything lower than a B, which I never got, you got grounded.

Group 3: Well, they would ground them for anything below an A+ sometimes in those honors classes.

Group 1: One thing that I was also just looking at, under the explanation of terms, number correct, length missing of minute responses are not counted correct. But it doesn't show anywhere on here like on the graph, that it would show me, hey, my child didn't even attempt or respond to do certain kinds of questions. So they could have a really low score just because they didn't attempt it or they didn't know how and I wouldn't know that. I would have just thought that they got it wrong.

Leader: Okay.

Group 1: so that would be something that I would want to know as a parent. You know if there are certain things that my child does not know how to do this, that is something they need to learn. As compared to their just getting it wrong.

Group 3: I have told my kids many times, be sure you put some kind of an answer on every question. And they still won't do it because they are afraid that they are going to get it wrong.

Group 1: Or they just might not know how to do it. And to me a blank is more so telling me that you don't know how to do it instead of just guessing and getting it wrong, or attempting it and getting it wrong.

Leader: So let's go ahead and start taking a look at number four. You can see that is basically the same info as the others and some more on the bottom. So go ahead and take a look at that and the next set of questions which are thirty-three through forty. And we will start talking about it in maybe five minutes or something like that. If you finish through number forty and you don't want to wait for everyone else to finish, you can do the last set of questions.

Group 4: Besides answering the questions, can we write you notes on that paper too?

Leader: Sure. Go ahead. If there is any particular question, something that stood out in your mind or you liked it or something like that, you can circle it.

Group 4: Actually, there are pros and cons to all of them. There is like different aspects that I take from each one and put them all together to make a report.

Leader: Yeah. I think that we only have about ten or fifteen minutes left, so I want to have a little more discussion. We will start talking in about another minute or two if that is fine. Are we ready to go ahead and say a last few, a little bit of discussion?

Group 1: Okay.

Leader: So, on this last report, number four. What did you guys think about this one and the inclusion of the performance on the clusters down here at the bottom?

Group 4: I liked that because it broke the subsets up into each component that makes it up because you can be strong in one area and not another, which I think is good. It gives you specific things that you need to work on for that individual student.

Group 3: I think that one that breaks it down this far would be, where it would be most useful would be if you actually have a conference with the parents and or the student if they are older and explain and go over the things in this. I think that it would be extremely valuable for them. I am not sure otherwise that it would be that much more beneficial because I am just thinking on peoples busy lives and I am not sure they would do it if it was not in the context of a teacher conference of something like that. I think that it would be great for a teacher conference.

Leader: Okay.

Group 2: For all of them, I made note of this that this was actually my favorite report because it was so comprehensive that I felt like beyond that, you know, I am coming at this from two angles. I am coming at this from the standpoint of the parent and I am coming at it from the standpoint as a former teacher and administrator. I think that a graphic like that with a below average, average, above average and put that that is the dot where you are, you know, naturally. One of the things that we run into, especially in elementary school. We get parents sending those kids into pre-kinder and kinder that have never had any norm group to compare their children to. So we will get parents coming in to tell us how gifted their child is and how brilliant their child is and they are not even fully toilet trained. Am I correct?

Group 3: I didn't have that kind of experience that you did, but um, as I was at one school much longer than I was at any of the others. And it was a neighborhood that was, had been highly military, and military people come in and use the VA loan and buy a new house and so by the time the subdivision has been there fifteen years, you have got fourth and fifth owners. And we had like a whole block that was a pretty nice looking subdivision, but we would have whole blocks that were owned by real estate companies and it was all rental. So we had a totally different clientele after even ten years than we did at the beginning. And so I began to experience some of what you were talking about. We had people coming in; kids who came that did not know a red crayon was red.

Group 2: Well see, in the area where my teaching experience was, as compared to where my administrative for twenty years in administration, because I was always were kids

were in housing projects pretty much, federally owned housing projects. Whereas when I was principal, it was in the s--- area of San Antonio, you know. Over by W---. But by the same token, compared to what you had at C--- High School or A---, my goodness…

Group 4: I was South of town. That is where I was teaching. South of downtown. You know.

Leader: What do you think about this particular report?

Group 4: I think that it is pretty good. I would definitely hope that there was some type of a cover or packet or something that goes, not a packet, something or kind of folder that goes with it that explains each one of those skills. Because in the same instance they may not know exactly what do synonyms have to do with reading…context clues and breaking all of that down. But yet you don't want to break that down too far because your paper can only hold so much. So I am hoping that something goes home with it because I am going, well…..

Group 3: That is why I said that it would be so valuable in a parent teacher conference or something, because I wonder how many people with master's degree know what a homophone is. I bet you that a lot of people with a masters degree have no idea with a homophone is. So I think that it would be tremendously useful if it is all explained. You could send a cover letter which would be better than not having anything. The most valuable would be if you were actually talking to an educated.

Group 1: Well, I don't think that is necessarily true. Because looking at the parent tell me reading vocabulary comprehension, well what includes that? I mean, at least breaking it down, I have some semblance of… if I don't know what one of these words means, I can go and look it up. But at least I have a category or a breakdown of the categories so that I know what it involves. And even without an additional explanation of each one of those things, I like it.

Leader: Okay.

Group 4: And again, this is middle school too, so a lot of this is probably relative to middle school terms too.

Group 1: If a parent doesn't know what a homophone is, they can probably ask their eighth grader?

Group 3: An eighth grader would come here and know it and his parent probably wouldn't. Because when we started teaching genre in third grade, that kind of….most of the parents said, what are you talking about?

Leader: I guess we are running out of time. One last question is, I asked you if you prefer to access score reports online or paper hard copies. Now a lot of it is that we have vendors here that have a lot of that information that can be accessed online. Either for a

parent or for an administrator or even teachers. Some of it you know is that they can drill down. Like it will have a certain level and then they will click on a graph and it will have like reading, and then it goes even further down and like that. Do you think that it….

Group 2: I think that it is rather futuristic, but, the online component. But I have actually scored both of those. But you know I have moments where my computer is down or I can't receive e-mail or, and I am thinking so a paper copy is pretty old school but still it is….I think both. You still want the paper but you would like to be able to access it online if you had accessibility. Even here, we are starting to go now, where some people are concerned because they are not going to telephone people to offer them work anymore. It is all going to be done online. So, if they don't have a computer at home, they are going to have to go to a public library or I mean….don't get so lost in this technological world over here that you forget that you know, there is 50% of the people in this world that don't have computers.

Group 3: I found out some kind of interesting things. My son carries a blackberry all of the time. And he works for a law firm downtown. There are times that I can get an answer from him more quickly by sending him an e-mail then by calling him on the phone because his secretary will tell me; he is in a conference right now. But even if he is in conference, he looks at that blackberry every two minutes and I will get an answer from him while he is in the meeting. So we are just, I feel like, I am a totally different generation from my children in terms of where we go for information.

Group 4: If you are able to click on a certain area and see that information. I just think if you need a hard copy, you can print it out. And if the parent didn't have a computer access, they can go up to the school. Usually, there is some kind of ….

Group 3: I think that you would be running much too much risk of being demeaning to the parents to say, well if you don't have a computer, you can come up to the school. I think that you should just sent it home.

Group 4: A lot of times they will ask, do you want a hard copy or do you want to view it online.

Group 2: I see where this is going though. The person that sends out these reports can send them to the school or district and it is the districts job to print all of that up.

Group 1: I have never gotten a report that is just on one page from any of my son's school stuff. My thing is that if there is information that I need from both pages, I can put them side by side with hard copies and look at it and be able to go back and forth without having to flip pages or scroll down screens. I like being able to do that.

Group 3: When I was not yet accustomed to scoring these test here on machine, we were trying to qualify them all on hard copies. And I feel like I know this stuff pretty well. I would go into the scoring room and turn on the computer and ….I had never seen that before, it was just a difference that you had to get accustomed to. Now I do fine when I

turn on the computer but I didn't for about a year or more. I had to get totally re-accustomed. My mind had to be re-programmed from going from here to the computer screen.

Group 1: Well I can do it both ways, but I prefer to be able to have it in front of my face. And I think that a lot of people are that way.

Group 3: A lot of people your age would come here …well I think that you are unusual. I think that most people your age would come here and it wouldn't make them a bit of difference.

Group 2: I think that you are atypical. I mean that in a complimentary way. You are educated and you have a degree and that makes a big difference. Parents that have none of what you have in the way of academic competencies love their child with the same intensity that you do. What we had to work at as administrators and parents like that, is to develop competencies little by little to teach them how to help their children. Parent training to just show parents, to give them skills to help their children at home and stuff like that. Because, I mean the parents were only third grade if that or fourth grade level of competency if that themselves.

Leader: Okay, I think that we are just about out of time, but is there anything else that you guys want to say as far as one of these reports that was useful. I need the surveys back.

Group 1: If I had to pick one it would be that one. If I would design one on my own, I would use a combination of the different reports and different aspects from them. I would use the explanation of terms page, the bottom graph…..

Leader: I am sort of doing the whole process and the study is to have these different reports and see what is useful and what is not.

Group 3: I suspect the superintendent would start saying okay, that one is really great, but on a per student basis, how much more is that going to cost me than the other one. And if it is very much more, I don't think that too many districts could afford it because their budgets right now are ….

Leader: I think that is it, because we are out of time.

APPENDIX C

CONSENT FORM

University of Massachusetts Amherst
Informed Consent Form for Participation in Score Reporting Research Study

Study Title: *Empirical Study of Stakeholders' Understanding of the Information Contained in Score Reports for Large-Scale Assessment*

Principal Investigator: Stephen Jirka, Center for Educational Assessment, University of Massachusetts, Hills South, Room 152, Amherst MA 01003
Phone: 413-555-1066 E-mail: sjirka@educ.umass.edu

Introduction to the study: We are inviting you to take part in a survey on student score reports for state department of education tests. This research is being conducted through the Center for Educational Assessment at the University of Massachusetts Amherst. The purpose of this study is to find out how useful and meaningful student-level score reports from state department of education tests are to parents and educators, and how these reports might be improved. The results of this study will be published in a research report and possibly a professional journal, and we expect to share these publications with many departments of education, including the department of education in Massachusetts.

What will happen during the study: During the survey, you and other students, parents, or teachers will be asked to review several sample student score reports and to answer some questions about the reports and how useful and meaningful they are to you. The intent of the study is to evaluate various reporting methods, not to evaluate any participant. The survey should take approximately 20 minutes to complete and you will be asked to complete a brief questionnaire that will allow you to provide individual feedback on the reports and allow us to describe the demographic characteristics of the survey participants as a whole.

Who to go to with questions: If you have any questions or concerns about being in this study you may contact Stephen Jirka of the Center for Educational Assessment, whose contact information is given at the top of this page. He is available to answer your questions. You will also have opportunities to ask questions about any aspect of the study before or after completing the survey.

How participants' privacy is protected: We will make every effort to protect your privacy. We will not use your name in any of the information we obtain from this study or in any of our research reports. We will not link any comments to specific individuals, schools, or districts.

Risks and discomforts: We do not foresee participants experiencing any risk or discomfort from being in this study. The intent of this study is to evaluate the effectiveness of different reporting methods, not to evaluate the individuals who take part in it. The study, in no way that we can think, would be judged as controversial. It is no more or less than a study aimed at improving the ways we communicate test results to educators, parents, and students.

Benefits: Your personal benefit from being in the study will be limited to an increased knowledge of how student results on large-scale assessments can be reported. We hope that the results of this study will have important societal benefit by providing necessary direction

for reporting these results to educators, parents, and students in a useful and meaningful manner.

Your rights: You should decide on your own whether or not you want to be in this study. You will not be treated any differently if you choose to not participate. If you choose to participate, you have the right to withdraw from part or all of this study at any time.

Human Subjects Review Approval: The Human Subjects Review Committee in the Department of Psychology at the University of Massachusetts Amherst has approved this study. If you have any concerns about your rights as a participant in this study you may contact the Human Research Protection Office via email (humansubjects@ora.umass.edu), telephone (413-545-3428), or mail (Office of Research Affairs, 108 Research Administration Building, University of Massachusetts, 70 Butterfield Terrace, Amherst, MA 01003-9242).

PLEASE READ THE FOLLOWING STATEMENT
AND SIGN IN THE SPACE PROVIDED IF YOU AGREE

Study title: *Empirical Study of Stakeholders' Understanding of the Information Contained in Score Reports for Large-Scale Assessment*

I have had the opportunity to ask any questions I have about this study and my questions have been answered. I have read the information in this consent form and I agree to be in the study. There are two copies of this form. I, the participant, will keep one copy and return the other with the survey, which will be kept on file with Ronald K. Hambleton, Executive Director of the Center for Educational Assessment, University of Massachusetts Amherst.

_____ _____

Participant Signature Date

_____ _____

Parent or Guardian Signature Date

BIBLIOGRAPHY

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice, 26(1),* 36–46.

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

Deng, N., & Yoo, H. (2009). *Resources for reporting test scores: A bibliography for the assessment community*. Bibliography prepared for the National Council on Measurement in Education. Available at www.ncme.org/resources/biblio1.cfm

Fern, E. F. (2001). *Advanced focus group research.* Thousand Oaks, CA: Sage.

Forte Fast, E. (2002) A g*uide to effective accountability reporting*. Washington DC: Council of Chief State School Officers and US Department of Education

Forte Fast, E., & Tucker, C. (2001, April). *Redesign of the student assessment reporting system in Connecticut*. Paper presented at the meeting of the American Educational Research Association, Seattle, WA.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17,* 145-220.

Haberman, S. J. (2008). *Subscores and validity* (ETS Research Report No. RR-08–64). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75(2),*209–227.

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: Allyn & Bacon.

Hambleton, R. K., & Slater, S. (1995). Using performance standards to report national and state assessment data: Are the reports understandable and how can they be improved? *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments,* Volume II (pp. 325-343). Washington, DC: National Assessment Governing Board, National Center for Educational Statistics.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Harris, R. L. (1999). *Information graphics: A comprehensive illustrated reference.* New York: Oxford University Press.

Henry, G. T. (1993). Using graphical displays for evaluation data. *Evaluation Review,17*, 60–78.

Henry, G. T. (1995). *Graphing data: Techniques for display and analysis.* Thousand Oaks, CA: Sage.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers?. *Educational Measurement: Issues and Practice*, *10*, 16-18.

Jorgensen, M. A. (2005). *Systematic feedback for more effective teaching and learning*. San Antonio, TX: Harcourt Assessment.

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research (*3rd edition.). Thousand Oaks, CA: Sage.

Luecht, R. R. (2003, April). *Applications of multidimensional diagnostic scoring for certification and licensure tests*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Mislevy, (1998). Implications of market-basket reporting for achievement-level setting. *Applied Measurement in Education*, *11*, 49-60.

Morgan, S. E., Reichert, T., & Harrison, T. R. (2002). *From numbers to words: reporting statistical results for the social sciences*. Boston: Allyn and Bacon.

National Center for Education Statistics. (2002). *The measurement of instructional background indicators: Cognitive laboratory investigations of the responses of fourth and eighth grade students and teachers to questionnaire items.* Washington, DC: U.S. Department of Education.

National Education Goals Panel (1998). *Talking about tests: An idea book for state leaders*. Washington, DC: U. S. Government Printing Office.

Nicol, A. A., & Pexman, P. M. (1999). *Presenting your findings: A practical guide for creating tables.* Washington, DC: APA.

No Child Left Behind Act of 2001. Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Roberts, M. R., & Gierl, M. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice,29 (3),* 25–38.

Ryan, J. M. (2006). Reporting scores and subscores for large-scale assessment programs: Strategies, issues, and concerns. In S. Downing and T. Haladyana (Eds.), *The handbook of test development*. Mahwah, NJ: Lawrence Earlbaum Associates.

Sinharay, S. (2009).*When can subscores be expected to have added value? Results from operational and simulated data* (ETS Research Memorandum). Princeton, NJ: Educational Testing Service.

SurveyMonkey (n.d.) Retrieved October 1, 2007, from http://www.surveymonkey.com *(private organization, no author, no date, Each page has a different URL)*

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press

Tufte, E. R. (2001). *The visual display of quantitative information, second edition.* Cheshire, CT: Graphics Press.

Tufte, E. R. (2003). PowerPoint is evil. *Wired, 11* (9).

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science,* 5 3, 327-339.

Wainer, H. (1990). Graphical visions from William Playfield to John Tukey. *Statistical Science*, 5, 340-346.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*, 14-23.

Wainer, H. (1996). Using trilinear plots for NAEP data. *Journal of Educational Measurement, 33*, 41-55.

Wainer, H. (1997a). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*, 1-30.

Wainer, H. (1997b). *Visual revelations: Graphical tales of fate and deception from Napolean Bonaparte to Ross Perot.* New York: Copernicus Books.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*, 301-335.

Wang, N. (2003). Use of Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement, 40*, 231-253.

Willis, G. B. (1999). *Cognitive interviewing: A "how to" guide.* Research Triangle Park, NC: Research Triangle Institute. Retrieved January 21, 2004, from http://appliedresearch.cancer.gov/areas/cognitive/interview.pdf

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage.

Zenisky, A. L., Delton, J., & Hambleton, R. K. (2006a). *State reading content specialists and NAEP data displays* (Center for Educational Assessment Report No. 598). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Zenisky, A. L., & Hambleton, R. K. (2012a). Developing test score reports that work: The process and best practices for effective communication. . *Educational Measurement: Issues and Practice, 31,* 21-26.

Zenisky, A. L., & Hambleton, R. K. (2012b). From "Here's the Story" to "You're in Charge": Development and maintaining large-scale online test and score reporting resources. (pp. 175-185). In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment: Theory, issues, and practice*. London: Taylor & Francis.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education,22 (4),* 359–375.

Zenisky, A. L., Hambleton, R. K., & Smith, Z.R. (2006b). *Do math educators understand NAEP score reports? Evaluating the utility of selected NAEP data displays* (Center for Educational Assessment Report No. 587). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Zucker, S., Sassman, C., & Case, B. (2004). *Cognitive labs*. San Antonio, TX: Harcourt Assessment.

Zwick, R., Senturk, D., Wang, J., & Cooper-Loomis, S. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20,* 15-25.