Simplifying Psychometrics with SAS®
Charles A. DePascale, Center for Assessment (NCIEA)
Jennifer L. Dunn, Measured Progress

ABSTRACT

“The point of evidence displays is to assist the thinking of producer and consumer alike. Evidence presentations should be created in accord with the common analytical tasks at hand, which usually involve understanding causality, making multivariate comparisons, examining relevant evidence, and assessing the credibility of evidence and conclusions. … The metaphor for evidence presentation is analytical thinking” — Edward Tufte

There are countless statistics and procedures commonly used for describing, analyzing, and evaluating the quality of large-scale, K-12 educational achievement tests. However, many of these psychometric analyses are complex, difficult to interpret, and not easily understood by policy makers and education official responsible for administering the tests. Additionally, to a large extent, information provided by these statistics is static, making its use passive and limited.

In this application, relatively basic SAS/GRAPH® techniques were used to depict critical characteristics of tests in a manner that was understandable and easily accessible to the end users. Test and item characteristic curves, information functions, score distributions, and other statistics are presented individually and in combination. Overlays of the same graphs across years depict how an individual test has changed over time. Presentations of sets of graphs of various grade level tests within a year enables comparisons of components of tests across grade levels.

The primary purpose of the application was to enhance understanding of individual tests and changes in tests over time through the presentation of readily available, but little used, information about the tests. The graphic displays allowed an evaluation of individual statistics against clearly marked critical values. More importantly, however, the application focused discussion on questions such as a) How have the tests changed over time? b) Were the changes planned for and expected? and c) Is there internal consistency between the design of the tests and their intended uses?

INTRODUCTION

When preparing a paper such as this for NESUG, the goal is always to find the right balance between discussion of the application itself and discussion of the SAS procedures and techniques behind the application. In this case, the content of the application, psychometrics in support of the development and evaluation of educational assessments, is likely to be of limited interest to most participants in the conference. The overarching purpose of the application, however, to improve understanding through the effective communication of visual information, transcends the specific context of psychometrics. Consequently, the focus of this paper is on the purpose of and rationale for each of the graphs and charts developed for the application and on the SAS procedures and techniques used to produce those graphs and charts.

The graphs produced for this presentation are designed to visually depict how key characteristics of the tests vary across grade levels or have changed across years within a grade level. The presentation of any single curve depicted here provided key information about the quality and characteristics of the test. The organized presentation of sets of curves depicts additional information related to consistency and change in the program. In general, although change and variability may not be undesirable, it is critical that they are planned for and understood during the test development process.

The psychometrics are described briefly in the following section to provide a supporting context for the presentations, but the procedures used to compute the psychometric functions will not be discussed. Those interested in the details of generating the psychometric functions are encouraged to contact the authors directly. Subsequent sections of the paper are focused on the design of the graphs and a description of the specific SAS procedures and techniques used to development them.
PSYCHOMETRIC BACKGROUND AND CONTEXT

The field of psychometric is concerned with the measurement of psychological traits and constructs; and with the development of instruments to measure those constructs. In this case, the constructs being measured are students’ knowledge and skills in reading and mathematics at grades 3 through 8. The measurement of those constructs through statewide, standardized tests to measure those constructs with validity and reliability across multiple years is the context of the application described in this paper.

The development of standardized tests in reading and mathematics is a complex process that often involves a combination of classical measurement techniques and item response theory (IRT) techniques. In this application, we examine three basic psychometric functions: item characteristic curves, test characteristic curves, and test information functions. Other psychometric characteristics examined through charts and graphs are p-values and achievement level classifications based on raw scores. Each of these terms is described briefly in the following sections.

ITEM CHARACTERISTIC CURVE

The item characteristic curve (ICC) has been labeled the key concept or building block in item response theory (IRT). The ICC is a plot of the level of performance on some task or tasks against some independent measure such as ability, age, etc. (Hambleton and Swaminathan, 1985). The plot is usually a smooth, non-linear curve with level of performance increasing as ability increases. A common form of the ICC in educational testing depicts ability on the horizontal axis as a theta scale ranging from -4.0 to 4.0 and level of performance on the vertical axis as the probability (0.0 to 1.0) of providing a correct response to the item.

TEST CHARACTERISTIC CURVE

The test characteristic curve (TCC) is defined as the sum of ICC for a particular test. It is a monotonically ascending ogive showing depicting the relationship between ability and expected true score on the test. As with the ICC, ability is shown on the horizontal axis. On the vertical axis, true score may be represented as a total score ranging from 0 to the number of points on the test or as total percent correct on the test ranging from 0 to 100.

TEST INFORMATION FUNCTION

The test information function (TIF) is measure of the precision of measurement along the ability scale. The larger the value of the information function, the more precise the measurement. The ability scale is depicted on the horizontal axis. The value of the test information function is shown on the vertical axis. The TIF for educational tests is commonly a bell-shaped curve with the maximum amount of information near the center of the ability scale or near a particular point of interest (e.g., an achievement level or pass/fail threshold).

P-VALUE

The p-value is a classical test theory statistic reflecting the proportion (0.0 to 1.0) of students responding to an item correctly. The p-value is also referred to as the difficulty of the item, although higher values indicate a less difficult item answered correctly by a large proportion of students. Scatterplots of item p-values are often examined as one step in evaluating change in performance from one test administration to the next.

ACHIEVEMENT LEVEL CLASSIFICATION

Since the dawn of the standards-based era in the 1990s and the ensuing passage of federal education legislation in 1994 (IDEA) and 2001 (NCLB), it has become commonplace to classify student test performance into one of three or more achievement level categories based on total test score. Total test score may be expressed as a raw score (i.e., number of points earned or questions answered correctly) or as a transformed scaled score. Often, these achievement levels are assigned names such as Basic, Proficient, and Advanced.

SAS GRAPHS AND CHARTS: EXAMPLES AND DESCRIPTIONS

In this section of the paper, examples of five SAS graphs and charts used in this application are presented and described. All of the graphs and charts were created using either the GPLOT or GCHART procedure in SAS/GRAPH. The following section provides details about the techniques used to produce each graphic.

Figure 1 depicts test characteristic curves (TCC) for four annual administrations of a grade 8 mathematics tests. The vertical lines on the graph represent the three achievement level thresholds (on the ability scale) separating...
performance into one of four achievement level categories. The TCC in Figure 1 show a test that was relatively stable in difficulty for three years and became significantly easier in the fourth year. Through the graph, this is shown by the separation between the TCC for years 1-3 and the TCC for year 4. That the test became less difficult in Year 4 is depicted by a higher percent correct on the Year 4 test than on the Year 1-3 tests being associated with performance at a given ability level. For example, at each of the three achievement level thresholds, a higher total score was needed in Year 4 than in Years 1-3. This type of change does not necessarily reflect a problem with the test, but should not occur unexpectedly.

Figure 2 depicts test information functions (TIF) for the same four grade 8 mathematics tests shown in Figure 1. The TIF show that although the difficulty of the tests remained constant across years 1-3, the test was undergoing some changes in where it was measuring most precisely. The TIF show that although the relative shape of the information function was fairly constant across years, the location of its peak shifted considerably. In Year 1, the TIF was centered near the threshold between achievement level 2 and achievement level 3. By Year 4, the TIF was centered much closer to the threshold between achievement level 1 and achievement level 2. That is, the grade 8 mathematics test in Year 4 was producing more precise measurement at a lower ability level than the corresponding test in Year 1. Again, this type of change should not occur accidentally. The horizontal reference line assists in the evaluation of the TIF – showing a benchmark for minimum level of information desirable at the achievement level thresholds. It can be seen in Figure 2, that although the value of the information function is well above this benchmark for the lower two achievement level thresholds, it is approaching the minimum benchmark at the highest achievement level threshold.
Figure 3 depicts the individual item characteristic curves (ICC) for each of the four grade 8 mathematics tests. Each curve on the graphs represents an ICC for an individual item. The relation of the item to the achievement level thresholds can be reviewed through these graphs. These graphs indicate that there are few items located near the highest achievement level threshold across the four years, and even fewer in Year 4 than in previous years.

In this figure, there is a separate graph for each test, but the four graphs are presented together. The presentation of multiple graphs allows for the direct comparison of the four tests in the same manner as Figure 1 and Figure 2. Obviously, there is too much information contained in each of the sets of ICC to attempt to combine the four years on a single graph. The presentation of a series of small, multiple graphs overcomes this obstacle. The intent of the presentation of the four graphs is not to elicit a detailed examination of any one graph, but simply to depict an overall representation of the tests across the four years. If the graphs reveal that the distribution of the ICC has changed across years (i.e., the pictures look significantly different) then it would be possible and advisable to examine one or more of the individual sets of ICC in more detail.

The sets of ICC in Figure 3 also contain an additional piece of information. The various colors of the ICC represent the content standard measured by the item. This information can be used to evaluate the relative difficulty of the items within and across content standards.
Figure 4 is a simple scatterplot depicting the p-values of a set of items administered on two occasions. Each point on the (x,y) plot represents the p-values of the item on the two administrations. In this case, the plot depicts p-values obtained from administrations of the items in consecutive years. The plot is used as a preliminary indicator of change in overall performance from one year to the next as well as an indicator of any unexpected changes in item behavior from one year to the next. The inclusion of the y=x reference line facilitates comparisons of performance across years. Items located above the line have higher p-values in Year 2 than in Year 1, and items located below the line have lower p-values in Year 2 than in Year 1.

The use of colors and labels to represent content strands provides additional information in the scatterplot. As with the item characteristic curves presented in Figure 3, the use of colors and/or labels allows the examination of item performance by content strand. In the example depicted in Figure 4, it appears that there is a difference in the pattern of performance across years based on content strand. Performance improved from Year 1 to Year 2 on Strand 1, remained the same on Strand 2, and declined on Strand 3. The use of color and labels makes it easy to see that, in general, students’ performance was strongest on the items in Strand 3 and weakest on the items in Strand 3.
Figure 5 contains two charts presenting two depictions of student score distributions on the same grade 8 mathematics test. The chart on the left contains the distribution of transformed scaled scores. The chart on the right contains the distribution of raw scores on which those scaled scores are based. Both charts show a distribution that is somewhat negatively skewed.

Viewed side by side, there are clear differences in the distributions of raw scores and scaled scores. The raw score distribution is smooth across the score scale, as would be expected in a distribution of the scores of tens of thousands of students. In the scaled score distribution, there is a collapsing of scores to form a peak at the low end of the distribution, and a second lone peak near the middle of the scaled score distribution. The introduction of colors into the charts to indicate performance levels associated with each raw and scaled score clearly shows that these peaks occur at the thresholds between performance levels. The peaks are an artifact of the scaling procedures to transform raw scores to scaled scores rather than a feature of student performance.
SAS GRAPHS AND CHARTS: TECHNIQUES AND PROCEDURES

In this section of the paper, the basic SAS techniques and procedures used to produce the graphs shown in Figures 1-5 are presented. As previously mentioned, all of the graphs and charts were created using either the GPLOT or GCHART procedure in SAS/GRAPH. Within those procedures, basic options were applied to overlay multiple curves on the same plot, assign specific colors to individual curves or portions of the chart, and to produce multiple plots on the same page. The basic SAS code used to produce each of these effects is presented and discussed.

OVERLAY PLOTS IN A SINGLE GRAPH

Figures 1, 2, and 3 depict examples of plots that contain multiple curves plotted on the same graph. In SAS, this is accomplished through the OVERLAY option in the PLOT statement of PROC GPLOT. In simplest form, graphs containing four curves with the same axes, similar to those shown in Figure 1 and 2, can be produced with the following SAS code. The OVERLAY option in the PLOT statements places all four curves on the same graph. In this example, the array variables year1 through year4 contain the level of performance for a given test administration (e.g., proportion correct) associated with each value of theta.

```sas
proc gplot;
  plot year1*theta year2*theta year3*theta year4*theta /overlay;
run;
quit;
```

Additional options can be used to refine the look and feel of the graphs. The code shown above, for example, would produce a graph containing scatterplots of unconnected symbols representing each plotted pair of data points. The four graphs would be distinguished by the color of the symbols. Use of the SYMBOL statement prior to the PROC GPLOT code can assign a specific color to each graph and also produce connected lines rather than unconnected symbols.

```sas
symbol1 color = red interpol=join;
symbol2 color = blue interpol = join;
proc gplot;
  plot year1*theta year2*theta year3*theta year4*theta /overlay;
run;
```
ASSIGN COLORS TO VALUES THROUGH SYMBOL STATEMENTS

In the same manner that the SYMBOL statement was used to specify the color of the curves in the example above, symbol statements can also be used to assign colors based on specific values of variables as shown in the ICC in Figure 3. A minor modification to the code shown above will produce curves colored by content standard.

```plaintext
symbol1 color = red interpol=join;
symbol2 color = blue interpol = join;
proc gplot;
  plot item1*theta=1 item2*theta=2 item3*theta=2 item4*theta=1 /overlay;
run;
quit;
```

In this example, the SYMBOL states are used to assign a color to each content standard: red for standard 1, blue for standard 2. The PLOT statement will generate ICC for four items. The ‘=1’ or ‘=2’ following each plot assigns the appropriate content standard (i.e., SYMBOL statement) to each item. Item 1 and Item 4 are assigned to content standard 1 and will be represented by a red line. Item 2 and Item 3 are assigned to content standard 2 and will be represented by a blue line. Additional standards can be assigned through additional SYMBOL statements and additional items can be included by specifying additional plots in the PLOT statement.

AN ALTERNATE APPROACH TO ASSIGNING COLORS AND OVERLAYING PLOTS

The technique used to produce the scatterplot in Figure 4 applies a slight variation to the use of the SYMBOL statement and an alternate method to overlay two plots. As in Figure 3, the SYMBOL statement is used to assign a specific color to a content strand. Additionally, the VALUE option is used to assign the strand number as the symbol used to represent the point on the plot. Unlike the approach used in Figure 3, however, in this example the variable name STRAND is used in the PLOT statement and the value of STRAND is determined for every point plotted. The appropriate SYMBOL statement corresponding to the value of STRAND is applied to each point.

```plaintext
symbol1 value='1' color=red;
symbol2 value='2' color=blue;
symbol3 value='3' color=green;
symbol4 i=join color=black;
axis1 label=none major=none value=none minor=none order = (0 TO 100 by 10);
axis2 order = (0 TO 100 by 10) label=('Year 1');
axis3 order = (0 TO 100 by 10) label=(angle=90 'Year 2');
proc gplot;
  plot Y*X = strand /vaxis = axis3 haxis = axis2 nolegend;
  plot2 X*X = 4 /vaxis = axis1;
run;
quit;
```

In Figure 3, the OVERLAY option is used to place to plots on a single graph. In Figure 4, the PLOT2 statement is used rather than the OVERLAY statement to plot the y=x identity line as a reference line on the graph. The PLOT2 statement is often used to place two plots with different vertical axes, but a shared horizontal axis, on the same graph. In this example, the vertical axis is also the same in both graphs as shown in the AXIS1 and AXIS3 statements defining the vertical axes. However, the OVERLAY option does not work with the Y*X=Z type plot, where in this example Z is the variable STRAND.

ASSIGNING COLORS TO VALUES IN A BAR CHART
In Figure 5, the SUBGROUP option in the VBAR statement of PROC GCHART is used to assign colors to the vertical bars based on the value of a specific variable. In this example, each bar reflects the number of students earning a particular scaled score or raw score. The colors of the bars reflect the performance level that corresponds to each raw score or scaled score.

```sas
label mrawsc = 'Raw Score';
label mscaleds = 'Scaled Score';
axis1 label= (angle=90 'Number of Students') order = (0 to 8000 by 1000);
proc gchart;
  vbar mscaleds mrawsc /discrete
  subgroup=mperflev
  raxis = axis1
  nolegend;
run;
quit;
```

The code above used to produce the charts in Figure 5 does not make use of the SYMBOL statement to assign specific colors based on the value of the performance level variable – MPERFLEV. Instead, the program is allowed to cycle through the default list of colors as different values of the variable are encountered. To be certain that the same color will be applied to a particular performance level across charts, the SYMBOL statement should be used.

**PRODUCING MULTIPLE GRAPHS ON A SINGLE PAGE**

Figure 3 contains an example of multiple graphs being produced on a single page. As discussed previously, this allows the reader to make direct comparisons across the graphs without the need to shuffle or lay out multiple pieces of paper. The number of graphs included on a single page can be defined through the program, and the size of the graphs displayed is adjusted accordingly.

In BASE SAS, multiple plots per page were produced through PROC PLOT using the HPERCENT and VPERCENT options to the PROC PLOT statement. For example, including HPERCENT = 50 and VPERCENT 50 would divide the page into four equal plot areas (2 X 2). Similarly, the statements HPERCENT = 33 and VPERCENT = 33, cutting the horizontal and vertical space into thirds, would divide the page into nine equal plot areas.

A separate plot statement for each desired plot within the PROC PLOT procedure would produce multiple plots on the same page. The code shown below would produce four ICC plots on a single page rather than overlaying them in a single graph.

```sas
proc plot hpercent=50 vpercent=50;
  plot item1*theta;
  plot item2*theta;
  plot item3*theta;
  plot item4*theta;
run;
```

In SAS/GRAPH, multiple plots per page are easily produced through the use of PROC GREPLAY, previously produced plots, and templates for creating multiple panels on a single page. The specific syntax of PROC GREPLAY is not critical to this presentation. Good sources for information on producing multiple plot displays is the 2002 publication by Perry Watts, Multiple-Plot Displays: Simplified with Macros; and the 1995 publication by Carpenter and Shipp, Quick Results with SAS/GRAPH Software.

**CONCLUSIONS**

The examples and code provided in this paper provide the basics for producing a series of relatively simple and straightforward graphs. These graphs form the basis of an application, but are not themselves the application. The final step in the development of the application is the decision making process that transforms the individual graphs produced into displays of evidence. As suggested in the quotation at the opening of the paper, that process requires an understanding of the evidence to be displayed, the questions to be answered, and the persons who will be interpreting the evidence to answer to those questions.
The graphs produced for this application were displayed in a variety of arrangements and formats to address particular questions. In most cases, multiple graphs were displayed on a single page to allow easy comparisons or to tell a complete story about a single test. A single page might contain a variety of graphs displaying all of the relevant characteristics or statistics for a single test. Alternatively, the page might contain graphs showing the same statistic for a single test across years. Another alternative presentation is the display of graphs of multiple tests within across grade levels within a single year. In practice, after individual graphs are produced, there is considerable flexibility in producing various arrangements of the graphs, as needed. Ultimately, each display is designed to assist the persons interpreting the display to answer a specific question.

REFERENCES:

ACKNOWLEDGMENTS
SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the authors at:

Charles A. DePascale
Center for Assessment
PO Box 351
Dover, NH 03821-0351

Work Phone: 603-516-7900
Email: cdepascale@nciea.org
Web: www.nciea.org

Jennifer Dunn
Measured Progress
100 Education Way
Dover, NH 03820

Work Phone: 603-749-9102
Email: dunn.jennifer@measuredprogress.org
Web: www.measuredprogress.org

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *