GUIDELINES FOR EFFECTIVE SCORE REPORTING

CSE Technical Report 326

Pamela R. Aschbacher & Joan L. Herman

UCLA Center for Research on Evaluation,
Standards, and Student Testing

July 1991

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Introduction

Serious concerns about low student achievement in this era of limited resources have led to an increased demand for data to guide and justify costly investments in our educational system. To meet this need, there has been a vast increase in the quantity of test information collected at all levels—local, state, and national. More states than ever before are conducting statewide testing programs, an increase from 39 to 47 states over the past five years alone. More tests are being given at more grade levels and in more content areas than ever before, with particular increase in writing, science, social studies, arts, and physical education (Aschbacher, 1989). With the advent of performance-based testing, still more assessment data will be available. But is more necessarily better? How can we ensure that this wealth of information demanded and now available is worth the time and expense of collecting it?

A comprehensive answer to these questions requires that a number of issues be considered: what information is truly needed, how to collect it, how to analyze and interpret it, and how to present it to appropriate audiences. A fair amount of attention has been given to the first three issues (witness the many professional articles, meetings, and discussions devoted to the National Assessment of Educational Progress alone), and rightly so, for they are very important issues. However, the fourth issue, appropriate presentation of results, has received far less attention. Yet the extent to which test data are clearly communicated and understood is critically important in educational decision-making. For example, the recent debate over the "Lake Wobegon effect" (i.e., the reality, possible causes, and implications of whether the vast majority of states and school districts nationwide are reporting above-average results on nationally normed standardized tests) illustrates how lack of precision and incomplete explanations in reporting practices can lead to misinterpretations, inaccurate claims, sharp criticisms, and inappropriate solutions (Linn, Graue, & Sanders, 1990).

Research has shown that the educators to whom written assessment reports are targeted often do not make much use of them and rely instead on personal conviction, informal information, and on staff members' interpretations of test results to derive policy implications and evaluate programs (Herman, 1989). According to users, reports were not organized to answer their questions or drive the decisions they needed to make. Furthermore, characteristics of reports that users said would facilitate their use of available data (such as summaries, graphics, trends of targeted subgroups, relationships between test scores and instructional, student, or community variables) are missing from many typical district and state reports (Aschbacher, 1989). Such findings underscore the need to improve the effectiveness of assessment reports.

The purpose of this paper is to examine current practice in state reporting of assessment results and to provide guidelines for effective reporting, derived from the literature on cognitive psychology, communications, and information representation and decision-making, along with illustrations of exemplary practice. Both content and format concerns are addressed. The information presented below on current practice is based on 1984 and 1989 reviews of assessment reports from over 30 states (Burstein, Baker, Aschbacher, & Keesling, 1986; Aschbacher, 1989), and on results from a spring 1989 written survey sent by CRESST to all state testing directors. The guidelines for score reporting presented below are based on a review of literature for this project and on work done for a CRESST project on increasing the utility of information systems in schools (Herman, 1989).

What is the Extent of Current State Assessment Reporting?

In 1989, 47 states had statewide assessment programs. Of these 47, 30 states used standardized norm-referenced tests as part of their statewide assessment programs, and 31 used objectives-based or criterion-referenced tests for state assessment. These two groups were not mutually exclusive; 14 states used both types of tests statewide. In addition, 14 states had minimum competency exams for end-of-course or graduation requirements, and 32 had writing assessments.

Clearly, generating reports on all these testing programs to a number of different audiences is a significant task in many states. In addition to an overall report for the state education agency, over three-fourths of the states provide individualized reports to districts and to schools as well. Over half the states also prepare class reports and individual student reports for teachers, parents, and students. Nearly half the states also prepare special reports for their state legislatures, and about a third provide press releases or short reports to the media. In states that use two or more different types of tests (e.g., norm-referenced, objectives-based, and/or minimum competency), which are often administered at different times of the year and thus reported separately, the number of reports can be staggering. Fully three-fourths of the states prepare at least five to as many as nine different types of reports of testing results each year! Altogether, approximately 275 different types of state assessment reports were produced in 1989 alone.

What is the Nature of Current Practice in State Reporting?

There is an extremely wide range in the length and sophistication of state reports produced. The basic state report that nearly every state publishes illustrates this variability. On one end of the continuum are reports that consist of about 10 pages of computer printouts or other tables with little or no explanatory text, while on the other end are reports with over 200 pages of text, tables, graphs, additional graphics, and color printing.

The contents of a typical state report are presented in Figure 1 (see Appendix). The typical report provides a summary of results for a single test for three grade levels in several subject matter areas (usually reading, math, and language arts). Overall scores are reported in a few bar graphs, and district results are presented in voluminous tables. Breakdowns of results by district, ethnicity, sex, and socioeconomic status are typically included.

Despite the many tables of numbers, few reports provide the kind of detailed information necessary to interpret results most accurately and meaningfully. Most states report the average scale score or median percentile for norm-referenced tests or percent of students scoring above a cutoff point on criterion-referenced tests. Only a very few states, such as Washington and North Carolina, give a more comprehensive view of student achievement by reporting distributions of results (see Figure 2). Very few state reports provide careful explanations of the statistics used, define common terms like "average" or "national norm" that many readers misinterpret, give the year in which the national norming was performed, or describe which students were excluded from testing (Burstein, 1990).

The typical presentation of achievement trends over time simply provides test scores for all students tested at a given grade level for several years in a row. Note that with this method, changes in student demographics over time are confounded with changes in achievement, and few states analyze or discuss the relationship between changes in test results and demographics over time. Hawaii is the rare example of a state that provides test scores on roughly the same cohort of students at two different grade levels (i.e., the same class of students when they

were in the third and sixth grades). However, the report does not include information about changes in the composition of the cohort due to attrition or population mobility from one tested year to another.

There seems to be a slight trend for more states to present data analyzed by ethnicity, sex, and region now than five years ago; the proportion analyzing data by SES (about half) has remained the same (Burstein et al., 1986; Aschbacher, 1989). Aside from these four variables, the typical state report does not relate test score data to other student, teacher, school, or community characteristics that might provide a more comprehensive explanation of student achievement.

However, about two-fifths of the states do report auxiliarly information about students, schools, or communities to contextualize their descriptions of educational performance. Most states that do report such information include at least some data on students' and/or parents' behaviors, attitudes, and characteristics, such as languages spoken, special needs, career plans, amount of television watched, or attitudes toward school. Some states report data on the learning environment, such as curriculum offerings, course requirements, use of equipment and materials, and teacher qualification or practices. Only six states report data on the community, state, or national context, such as local, state, or federal education expenditures, percent of state population with high school diplomas, average per capita income in the region, or percent of the state's school age children who are handicapped.

The states that report auxiliary information incorporate it in their assessment reports in a variety of ways. A few states, such as North Carolina, provide interesting context data and discuss it in the text but do not directly relate it to test results (see Figure 3). Some other states discuss the relation of auxiliary data to achievement but do not provide graphics or illustrative statistics in the text. Maine's "Math: Interpretative Summary of Results" (see Figure 4) is a case in point. Only a few states, such as Utah, present the information graphically and also discuss in the text the relationships between test scores and contextual data (see Figure 5). Most of the states with data from teacher or student surveys, school records, and so forth simply report the results in a separate part of the report without tying them to the student achievement data. Some readers may draw inaccurate conclusions; others may be left with no sense of the "big picture" or how to combine the pieces of information to make appropriate decisions.

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

How Can Assessment Data Be Reported Most Effectively?

Although decisions about the content of a report or training materials may be guided by political as well as technical considerations, decisions about how to present the information should be informed by research and experience. However, for a number of reasons, state reports tend to be produced without benefit of formal feedback on the success of previous reports, other states' reporting experiences, or information from research on how readers perceive and comprehend data in written materials.

Most state and local education agency staff must produce reports within tight timelines and with limited resources, so they tend to rely on what has worked for them in the past—or at least has not caused significant problems. They do not routinely conduct studies of user satisfaction or other evaluations of their reporting practices. For example, only six percent of districts in one state conducted such evaluations (Barber et al., 1988).

Although many state education agencies share test items and other information about their testing programs with each other, few states routinely share their reports, their approaches to creating the reports, or their successes and failures in reporting. Even states that have closely shared test development strategies (such as Michigan and Illinois with their new reading tests) have not shared their methods of reporting the reading test results. No doubt the general lack of report sharing is a function of many variables: the different levels at which data are reported from state to state (e.g., student versus school), the small amount of time and resources within which reports must be generated, the tendency to repeat a tried-and-true recipe, and cuts in funding that have curtailed the routine mass-mailing of reports to other states.

The effectiveness of reports could be enhanced by attending to research as well as experience. In fact, research has demonstrated that a reader's comprehension and use of data is a function of the way the data are presented (MacDonald-Ross, 1977). For example, the visual and logical arrangement of data have an important effect upon recall and comprehension of information (Washburn, 1927).

The following guidelines for effective presentation of data are based on an understanding of the cognitive limitations and preferences of readers and the use of data in decision-making. Examples are drawn from reports of state assessment results or other materials distributed by state education agencies. Obviously, with the variety of purposes and audiences that exist across districts and states, there is no one best way to report assessment results. The recommendations presented here will need to be adapted to individual situations and political and logistical realities.

Guidelines for Effective Reporting of Assessment Results

1. Know the audience and the purpose.

Use the expectations of the audience to enhance credibility and utility of the findings. A report is most likely to be truly useful if it directly addresses the specific concerns and information needs of its audience. Research conducted by the Far West Lab (Mills, in draft) on the impact and utility of state reports suggests that information in a general report tends to be used less by superintendents, principals, teachers, and legislators, for example, than separate reports tailored specifically to each of these audiences. In weighing the value versus costs of creating separate reports and in determining the content of any one report, it is necessary to consider the needs of each significant audience for the testing information: what do they need or want to know? What decisions should the information inform or actions should it motivate or justify?

The information in the report will not be used if the readers do not believe it is useful or valid, so steps must be taken to present it in such a way that the audience will accept it. Research has shown that people tend to believe what they want to believe. They tend to ignore findings that are inconsistent with their expectations (Cousins & Leithwood, 1986), and they seem willing to accept data that challenges the status quo when they do not support current practice (Weiss & Bucuvalas, 1980). Thus, authors of assessment reports should consider the expectations of the audience when deciding how elaborately and carefully to lay out the logical analysis of results, or deciding which results need to be emphasized or given particular support in order to be attended to and believed. To help readers make their expectations for certain outcomes explicit so that the data they are reviewing will be used, the following strategies might be employed:

- prompt readers to ask a question of themselves related to possible findings before viewing the data;
- present tables and graphs with headings in the form of questions, such as "What do you see happening to reading scores over time?" or "Is there a relationship between students' attitudes towards math and their test scores?"

In addition to addressing readers' expectations, it is important to supply data with convincing credentials in order to avoid losing a reader who decides that the data are untrustworthy. The most important credential is a clear account of where the numbers came from and how they were manipulated so that the report will not be discounted for vagueness or statistical machinations. The explanation, whether in the text or a technical appendix, should leave the reader with the sense that the authors are careful with numbers and their interpretation, and are not afraid to reveal details of the data collection and analysis, such as who was not tested or how missing data was handled (Lowry, 1983).

Let the expertise and background knowledge of the audience guide the nature and amount of explanation and details included. Many reports tend to assume that the readers are very familiar with tests, statistics, and measurement, but this is often not the case. Policymakers and legislative staff, for example, must deal with many issues besides education, and many issues within education besides testing. Many readers are not particularly literate in assessment and may be unfamiliar with technical terms and various acronyms (Aschbacher, 1988). Many local board of education members are uncomfortable with statistical or technical reports and rely on district staff to interpret the results for them (Herman, 1989). In a study of parents' understanding of test results, about a third of the parents had not had more than a high school education, and three-fourths did not understand

how mastery of objectives was measured (Barber et al., 1988). Hence, it is important to include sufficient explanations for such audiences without using a condescending tone. For example, item response generated scores, standard deviation scores, and graphs summarizing "qualitative" information (such as percent of goals mastered when mastery means getting 75 percent correct on five out of seven objectives) should be reported with careful explanation as these may be easily misunderstood. In general, it is better to assume that readers know little than to assume they know too much. The readers can skip over information they know, but cannot supply details that they do not know yet must have in order to understand the results and make good decisions.

Use language appropriate to the audience; avoid jargon. In most score reports the prose is less than lively. While it is easy to take a cheap shot at their writing style, its impact on utility needs to be considered. The best-used reports are readable and interesting.

Typically, much technical language is embedded in explanatory materials, and it can be pretentious and obscure. On the other hand, technical terms are conducive to brevity as well as precision and can be appropriately used with some audiences (Estes & Wetmore, 1983; Ruben, 1989). Over-simplification or casual use of terms that have technical and lay definitions (such as "average," "norm," or "significant") can lead to misinterpretations. Footnotes or a glossary of technical terms, types of scores or statistics used, acronyms, and abbreviations may be useful to clarify meanings.

2. Keep It Simple

Present information in a small number of categories. To make the report more usable, minimize the information processing load on readers. Select a presentation format that will allow the reader to organize and conceptualize the information in terms of a small number of chunks or categories (five to nine). Miller (1956) found that humans are only capable of retaining about seven (plus or minus two) chunks of information in short-term memory.

Organize the report so that it can be read selectively, since many readers do not have time to digest the entire report and only will attend to parts of it. This may require some repeated brief references to critical details or definitions to prevent misunderstandings that could occur from reading only a limited portion of the report, such as the findings without the methods.

Use summaries. Summarize the main points of the report in a short executive summary or abstract at the beginning of a report as an advance organizer. This will provide a cognitive structure into which readers can incorporate the many details presented in the body of the report without getting lost in the verbiage or masses of tables. It can also quickly clarify what is and is not covered in the report so that the reader can zero in on topics of interest or turn to another report as needed.

Report users in Herman's study (1989) said that they greatly appreciated report summaries and preferred that they be located at the beginning. Surprisingly, only half of the state reports reviewed by Aschbacher (1989) contained summaries or abstracts, and many of these were located at the end of the report. Figure 6 displays the one-page executive summary contained in Alaska's 200-plus page state assessment report. Some states' summaries are longer and contain brief data charts as well as text.

Some state reports discuss data from a variety of testing programs and other sources, such as teacher and student questionnaires. An example of such a report is

What Is the Penny Buying for South Carolina?, a report on the assessment of the South Carolina Education Improvement Act of 1984. In this case, an initial summary (see Figure 7) organizes report highlights by chapter, and each subsequent chapter has its own overview as well. Such advance organizers in long, complicated reports are extremely valuable.

Be straightforward. This common sense guideline was underscored by research on how people process information in displays (Bettman & Kakkar, 1977; Ghani, 1981; Slovic & Lichtenstien, 1968). In general, these researchers found that people tend to select for use information that is easily processed, to resist making even simple transformations of information or data, and to discount or ignore information requiring inference from an explicit display. In other words, people tend to use only the information that is explicitly displayed and use it only in the form in which it is presented. Thus, a good display should provide information in the most straightforward and applicable form possible. In a similar vein, research on the recall of ideas from text indicates that readers remember explicitly-stated ideas better than ideas they have to infer from the text (Estes & Wetmore, 1983).

A small booklet produced by Massachusetts illustrates how weak skill areas may be effectively communicated to a general audience without requiring much transformation (see Figure 8). It presents three kinds of information together—general conclusions about a specific skill, statistics, and a sample item—rather than the typical method of presenting them in separate sections as is typical in large reports. The latter practice forces the reader to go through several steps—to locate, synthesize, and transform the information—in order to make full use of it. Most readers will not bother to do so.

Many state reports make little attempt to explore relationships among different kinds of data; yet, given the kinds of data there, it is clear that some of it was selected because of implied relationships among data such as class size, ethnicity, language background, and staff experience. Since many readers tend to make errors in interpreting statistics, much less teasing out cause-effect relationships, it would be better to explicitly investigate and articulate these relationships and prevent erroneous assumptions or inferences.

3. Be clear, accurate, comprehensive, and balanced.

Interpret statistics clearly; provide explanations. It is important to accompany data with explicit explanations or cautions that will counter common statistical misinterpretations. The four most common errors that people make when interpreting statistics are (according to Remus & Kottemann, 1986):

- users assume trends where in fact only random variation is occurring;
- they give too much weight to findings based on small or unrepresentative samples;
- they make incorrect cause-effect inferences from correlations or even independent events; and
- when faced with several sources of uncertainty, users simplify decision contexts by ignoring or discounting some of that uncertainty.

Readers of test score reports often confuse percent correct with percentile and have difficulty understanding scale and standard scores. For example, item response theory scores are sometimes presented in reports without any interpretation; since these are often seen as standard deviation scores by the uninitiated, this lack of explanation may lead to misunderstanding. Composite

scores should also be accompanied by an explanation of how they were derived. Even the term average is often used without careful definition and may be misleading.

Some state reports include relatively brief and straightforward explanations of the statistics presented and cautions in interpreting them. For example, the Colorado report distinguishes between percent correct and percentile rank and cautions how percentiles can be manipulated (see Figure 7). The Washington report carefully explains the type of score they use, how to interpret it, and how to interpret the box and whisker plots they present (see Figure 8 for an excerpt). The North Carolina report provides several pages of explanation of measures of central tendency and definitions of measurement terms (see Figure 11 for a portion).

Readers often misunderstand references to "the norm." When a report compares state or district results to the national norm, it is important to provide two pieces of information to lessen the exaggerated view of performance that can otherwise occur (Linn, Graue, & Sanders, 1990):

- emphasize the year in which the norms were obtained and explain the implications of using norms that are several years old or annual user-based norms;
- identify clearly the number of students excluded from testing, the proportion of the student population this represents, and the reasons for those exclusions.

Many reports lack explanatory materials accompanying the displays in tables or graphs. Often explanations of what population was tested, how scores were derived, and how data were collected and interpreted can be lengthy and informative, but they are usually set apart from the display in either an introduction or appendix. However, even when these are read by readers, the limits of short-term memory would suggest that the explanations have only a slight chance of being used. Hence, it is recommended that critical points be included near the data displays.

Many readers will find it useful for the report author to distinguish between statistical and practical significance when presenting test results. For example, one report contained the following text: "Student performance in both math and reading is quite good and has been improving over the last four years. This year 20 percent of our students scored in the above average stanine categories in reading. This is in contrast to 19 percent last year." It is not clear from the report whether this difference is statistically significant, but the language implies that it is an important difference. The reader might well wonder, however, if a one percent change is indicative of meaningful improvement.

Provide as rich a picture of the findings as the purpose and data allow. Dwelling on system-level average performance is simplistic and misleading (Burstein, 1990). Make use of the variety of interesting information available (such as that collected by many states on student background, resources, curriculum, and instruction) to provide a context to help explain and describe educational performance.

Apportion space in the report according to the importance of the ideas or results. Describe relatively unimportant findings and ideas as briefly as possible, and avoid using vivid examples, displays, or descriptions for them. Research on the comprehensibility of text indicates that the amount of space devoted to an idea and the vividness of its description are related to how easily it is recalled (Estes & Wetmore, 1983).

4. Use techniques to capture and focus the reader's attention.

Use visual techniques such as color coding, Q & A format, and graphics. Many readers of test results will have little time to peruse the report, so visual techniques will help highlight the most important findings and help them locate details in which they may be especially interested. For example, Utah uses heavy-weight colored title pages with an attractive logo to separate the sections of its report of state assessment results. The same color is repeated with good effect on many of the bar graphs in the report. South Carolina's report *What Is the Penny Buying for South Carolina?* (not the basic report on state assessment results) utilizes some simple but eye-catching graphics as borders for many of the graphs presented (see Figure 12).

A large body of research summarized by Bohle and Garcia (1988) indicates that color in print advertising and text has a strong effect on readers. In general, people prefer pages with color over black and white pages; the use of color may boost readership. Color is more important in gaining attention than is illustration, but while color is better for the fast grab or the quick appeal, black and white is better for a response needing more in-depth thinking. Bohle and Garcia's research found that color had a strong effect on eye movement when a person reads. They also found that color alone made a newspaper appear more interesting, pleasant, exciting, and powerful, but not more or less believable, ethical, or valuable.

Research suggests that color coding in graphs is most effective in multiple line graphs when many categories are included and highly discriminable colors are used (Schutz, 1961b; Shontz, Trum, & Williams, 1971). Most state and district reports reviewed, however, did not utilize color printing, probably due to cost.

Question-and-answer format can make information more inviting to the reader. Both Alaska and Michigan reports use this style to advantage (see Figures 13 & 14). For example, the Michigan handbook on its state assessment program has sections formatted in Q & A style that quickly and effectively draw the reader's attention to specific points the authors want to highlight, such as appropriate use of test results.

Use effective headings for sections, tables and graphs. Headings provide an explicit structure which assists the reader in integrating the information presented. To be helpful, headings need to be informative and unambiguous (Wright, 1977). Interesting and informative headings can pull a reader through a report that would not otherwise be read to its conclusion. Headings in the form of running commentary parallel to the text may be more useful than the typical format (Morris, Fitz-Gibbon, & Freeman, 1987).

Headings for graphs are one means of directing users' attention towards important data and clarifying what the graphs represent. All too often graph headings are too long and complicated to direct the reader's attention quickly. Instead of struggling to include all the important variables and modifiers in the heading, report authors could put some information in an explanatory footnote below the graph. Headings are more easily located when not in all capital letters (Poulton, 1967). See Figure 15 for an example of how such common practices could be improved.

Headings in the form of questions can draw readers into the report and prompt them to ask questions of themselves related to possible findings before viewing the data. For example, the *Illinois Schools Report to the Public* (not the general state assessment report) heads most of its graphs with questions and then accompanies each graph with an interpretative paragraph (see Figure 16). This

technique can help the readers make their expectations for certain outcomes explicit so that they are more likely to accept and use the data presented.

Use negative wording for special effect. A sentence with negative wording may be much stronger than its alternative affirmative wording (Wright, 1977). For example, saying "Never use a subject sample size smaller than five" is more emphatic than saying "Choose a sample size of five or larger." Furthermore, negative sentences are more effective at correcting a reader's expectations than positive sentences (Wason, 1965). However, because readers have an easier time with text that is written using simple, active, affirmative statements, negative wording should be used sparingly when emphasis is required (e.g., findings or directions that conflict with user expectations).

5. Suit format to purpose.

Use both numeric and adjectival descriptions of data. Numeric data are numbers reported as numerals or words (10 or ten). Nonnumeric data are adjectival and qualitative descriptions (e.g., extremely large or substantial decline; Bell, 1984). Research suggests that many people recall more accurately and better comprehend the overall meaning of adjectival information than numeric (Scammon, 1977). For example, "one-third of the students..." is remembered and understood better than "34 percent of the students..." On the other hand, the same research suggests that people tend to ask more questions and want more information when they are presented with adjectival rather than numeric data. Perhaps this effect occurs because numeric data is perceived to be more precise and accurate than adjectival information and allows for easier comparisons across values in a display (Cherry, 1966; Russo & Dosher as cited by Ghani, 1981).

However, numbers may lead readers to believe that the information is more precise than is truly the case or that there are more distinctive differences between individuals or groups than really exist. The conflict of providing readers with information in a form that will be more accurately understood (adjectival) or in a form that will be more readily accepted and more easily compared (numerical) will have to be resolved by development. Meanwhile, providing numeric and verbal descriptions of test results provides a means of balancing concerns with comprehensibility, credibility, and accuracy. For example, adjectives accompanying numerical categories (such as percentile ranks 1-25 being designated as "below average") are used in some reports.

When both numerical data and text are involved, Washburne's research (1927) suggests that numerical data should not be presented as embedded within text if more than two items are to be presented. Instead, separate tables or displays should be used.

Be consistent. Readers tend to resist changes in information representation styles (Ghani, 1981). Therefore, consider keeping the format of graphs and tables consistent throughout the report. However, use consistency reasonably, not slavishly. The type of graph used should be geared to the type of data to be displayed (see following section). Different formats within a report can help cue the reader that something conceptually different is being presented. They can also provide some interesting variety.

Select graphic format to fit the purpose. When considering which format to choose for tables and graphs, one should be aware that no one graphic format proves universally superior and that each format has its own domain of application (MacDonald-Ross, 1977). Washburne (1927) compared the effects of different formats and concluded that different formats foster the learning of different types of information. Research and experience suggests the following properties of

different formats (Croxton & Stein, 1932; Schutz, 1961b; Vernon, 1950; Washburne, 1927):

1. **bar graph**: best for static comparisons; facilitates easier and more accurate comparisons than pie charts, segmented bar charts, or three-dimensional figures (cubes).

2. **line graphs**: best for trends (dynamic comparisons); a single graph with multiple lines (if not too many and too confusing) is better at representing multiple trends than multiple graphs with single lines.

3. **pictographs**: give users rapid and striking general impressions.

4. **tables**: when presenting exact numbers, tables are one of the most important formats because of their compact nature. However, they have a serious weakness—they are abstract and therefore require more processing by the reader (MacDonald-Ross, 1977). Thus, employing a table involves a trade-off between exactness, compactness, and ease of use.

5. **pie charts**: best show portions comprising a whole.

6. **box and whisker plots**: show the spread of scores; compare the same group's high, average, and low scores on different tests or compare different groups on the same test, or compare one group's growth over time; analyze the effects of special programs on the high, average, or low students in the group (see Figure 8).

7. **frequency distributions**: show differences at all levels of performance, not just the mean or median, as usually reported.

Follow guidelines below for creating effective graphs and tables.

General Guidelines

- Label the graphs and tables so the reader can interpret each entry without further assistance; make use of titles, heads, and footnotes.
- Titles should be brief but contain enough information to distinguish one table from another and to help the reader decide whether to linger or move on.
- Do not write headings in all capitals. Use upper and lower case.
- Color coded bars and lines on graphs help readers compare information more easily than lines and bars in the same color.
- Include zero on the scale or use a break in the scale to clearly indicate some exclusion.
- A confidence band (determined using standard error of measurement) can indicate whether apparent differences are significant.

Bar Charts

- Information should be labeled directly on the chart rather than through a "key" or "legend."
- Horizontal bar charts leave more room for labels but may be more difficult to read.

Line Graphs

- Avoid clutter, both of lines and tic marks.
- Choose the range of tic marks to include the entire range of the data; avoid scale breaks.
- If scale breaks are necessary, do not connect numerical values on either side of a break.
- Choose a scale so that data fill up as much of the region as possible.
- Use a logarithmic scale when presenting multiplicative factors.
- Data lines should be thicker than grid lines.
- Each data line should be clearly labeled.

Pictographs

- Symbols should be self-explanatory and easily differentiated from one another.
- Quantity is better represented by increasing the number rather than the size of symbols.

Tables

- The more numbers to be compared, the fewer digits each should have.
- Round numbers as much as meaningfulness allows.
- Numbers to be compared should be adjacent; horizontal comparisons are easiest for up to six numbers; for seven or more, vertical comparisons are easier.
- Order rows and columns by size of numbers rather than alphabetical order, although lists of district results may be easiest to use if alphabetized.
- Set rows and columns compactly so that the eye can make easy comparisons rather than spacing them out across the page.
- Provide row and column averages.

Pie Charts

- Limit pie charts to five or fewer segments.
- Put labels inside segments if possible.

√ Summary Checklist for Effective Score Reporting

1. Know the audience and the purpose.

[] a. Use their expectations to enhance credibility and utility of the report

[] b. Use their expertise and background knowledge to guide explanations and details

[] c. Use language appropriate to the audience; avoid jargon

2. Keep it simple

[] a. Present information in a small number of categories

[] b. Use summaries

[] c. Be straightforward

3. Be clear, accurate, comprehensive, and balanced

[] a. Interpret statistics clearly and provide explanations

[] b. Provide as rich a picture as appropriate for the audience and purpose

[] c. Apportion space according to the importance of the ideas

4. Use techniques to direct reader's attention

[] a. Use visual techniques (e.g., graphics, colors)

[] b. Use effective headings

[] c. Use negative wording for special effect

5. Suit format to purpose

[] a. Use numeric and adjectival descriptions of data

[] b. Be consistent

[] c. Select graphic format to suit purpose

[] d. Follow guidelines for effective graphs and tables

References

Aschbacher, P.R. (1988, November). *Collaborating to make testing policies that work.* Presentation to the Education Staff Network at the Annual Meeting of the National Council of State Legislatures, Anapolis, MD.

Aschbacher, P.R. (1989, June). *Current trends in using and reporting state assessment results.* Paper presented at the Annual Assessment Conference of the Education Commission of the States, Boulder, CO.

Barber, B., Gadsden, V., Paris, S., Roeber, E., Checkoway, M., & Hull, L. (1988, April). *Current policies for reporting large scale assessment results to parents in the U.S.* Paper presented at the National Conference on Measurement in Education, New Orleans.

Bell, J. (1984). The effect of presentation form on the use of information in annual reports. *Management Science, 36*(2), 169-185.

Bettman, J., & Kakkar, P. (1977). Effects of presentation format on consumer information acquisition strategies. *Journal of Consumer Research, 3,* 233-240.

Bohle, R.H., & Garcia, M.R. (1988) Reader response to color halftones and spot color in newspaper design. *Journalism Quarterly, 64*(4) 731-739.

Bowman, J.P., & Branchaw, B.P. (1984) *Business report writing.* Chicago: Dryden.

Burstein, L. (1990). Looking behind the "average": How are states reporting test results? *Educational Measurement: Issues and Practice, 9*(3) 23-26.

Burstein, L., Baker, E.L., Aschbacher, P., & Keesling, W. (1986). *Using state test data for national indicators of educational quality: A feasibility study* (CSE Tech. Rep. No. 259). Los Angeles: UCLA Center for the Study of Evaluation.

Cherry, C. (1966). *On human communication.* Boston: MIT Press.

Cousins, J., & Leithwood, K. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331-364.

Croxton, F., & Stein, H. (1932). Graphic comparisons by bars, squares, circles, and cubes. *Journal of the American Statistical Association, 27,* 54-60.

Estes, T.H., & Wetmore, M.E. (1983). Assessing the comprehensibility of text. *Reading Psychology: An International Quarterly, 4,* 37-51.

Ghani, J. (1981). *The effects of information representation and modification.* Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

Herman, J.L. (1989, June). *Toward greater use of test and evaluation data.* Paper presented at the Annual Assessment Conference of the Education Commission of the States, Boulder, CO.

Lowry, I.S. (1983). *Designing readable and persuasive tables.* Santa Monica, CA: RAND.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990). *Comparing state and district test results to national norms: Interpretations of scoring "above the national average"* (CSE Tech. Rep. No. 308). Los Angeles: UCLA Center for Research on Evaluation, Standards, and Student Testing.

MacDonald-Ross, M. (1977). How numbers are shown: A review of research on the presentation of quantitative data in texts. *Audio Visual Communication Review, 25*(4), 359-409.

Maidment, R. (1988, September). Seven steps to better reports. *Management Solutions, 33*, 31-34.

Miller, G. (1956). The magical number seven plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63*(2), 81-97.

Morris, L.L., Fitz-Gibbon, C.T., & Freeman, M.E. (1987). How to communicate evaluation findings. In J.L. Herman (Ed.), *The program evaluation kit.* Newbury Park, CA: Sage.

Poulton, E. (1967). Searching for newspaper headlines printed in capital or lower case letters. *Journal of Applied Psychology, 51*, 417-425.

Remus, W., & Kottemann, J. (1986, December). Toward intelligent decision support systems: An artificially intelligent statistician. *Management Information Systems Quarterly, 10*(4), 403-418.

Ruben, D.H. (1989) Bibliotherapy: Practical considerations when writing for substance abuse readers. *Journal of Alcohol and Drug Education, 34*(3), 70-78.

Scammon, D.L. (1977). "Information load" and consumers. *Journal of Consumer Research, 4*(3), 148-155.

Shontz, W., Trum, G., & Williams, L. (1971). Color coding for information location. *Human Factors, 13*(3), 237-246.

Shutz, H.G. (1961b). An evaluation of formats for graphic trend displays: Experiment II. *Human Factors, 3*(2), 99-107.

Slovic, P., & Lichtenstein, S. (1968b). The importance of variance preferences in gambling decisions. *Journal of Experimental Psychology, 78*, 646-654.

Washburne, J. (1927). An experimental study of various graphic, tabular and textual methods of presenting quantitative material, part I. *Journal of Educational Psychology, 18*, 361-376.

Wason, P. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior, 4*, 7-11.

Weiss, C., & Bucuvalas, M. (1980). *Social science research and decision making.* New York: Columbia University Press.

Wright, P. (1977). Presenting technical information: A survey of research findings. *Instructional Sciences, 6*, 93-134.

Appendix

Figure 1

The "Typical" State Assessment Report

- 53 pages
- includes summary, text, overall results in a few bar graphs, tables of district data by grade and subject area
- for CRTs: percent achieving given objectives; for NRTs: median percentiles
- comparisons of state data to national norm
- test score trends over past 4-5 years
- test scores analyzed by ethnicity, sex, SES

Figure 2

Box and Whisker Plots to Report Distributions of Scores

Washington Statewide Assessment General Report, Fall 1987, p. 23

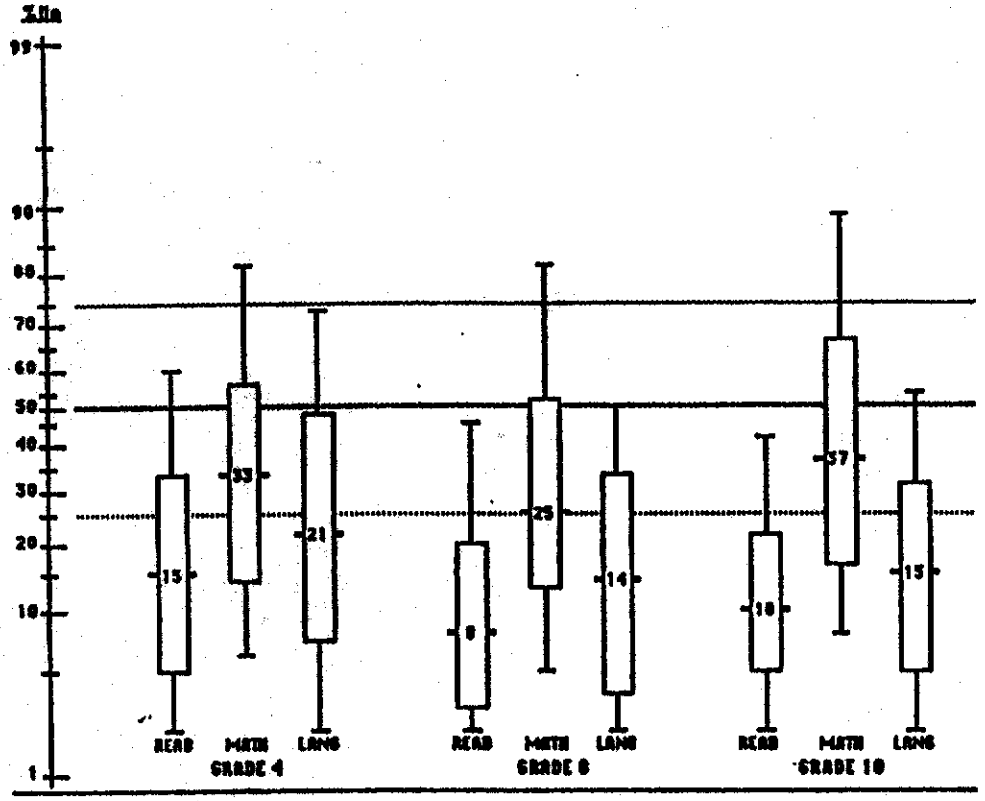
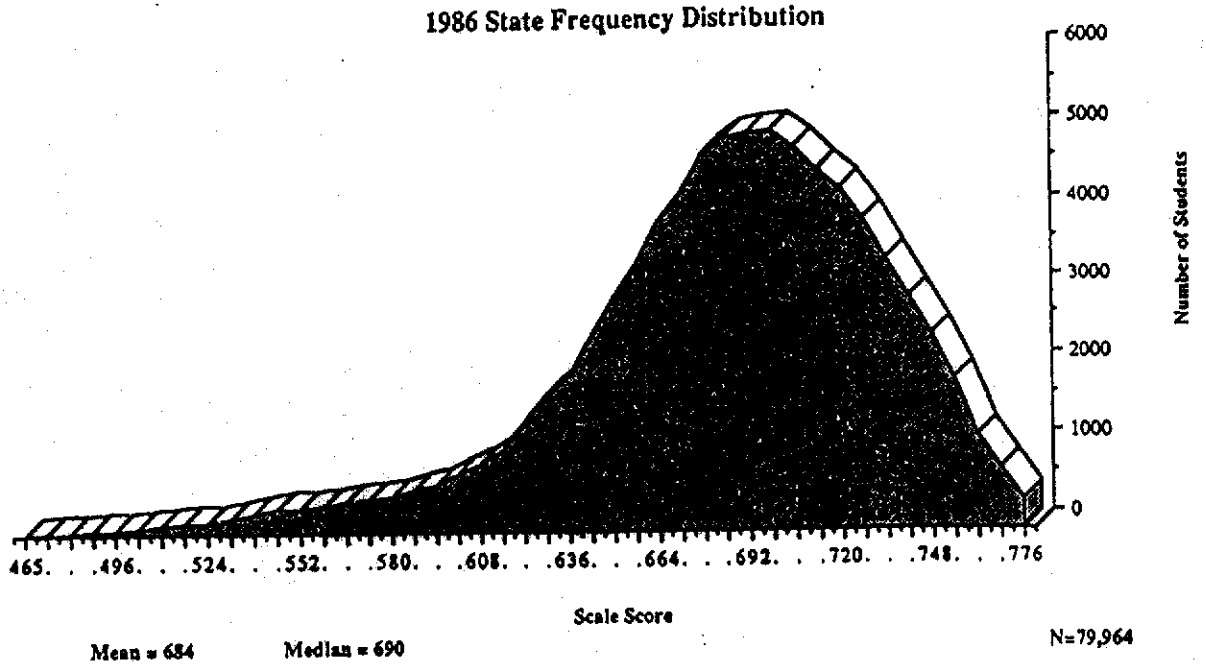


Figure 2 (cont'd)

Reporting Score Distributions

North Carolina's Report of Student Assessment, 1986-1988, p. 17

Figure1
California Achievement Tests
Grade 3 Total Battery



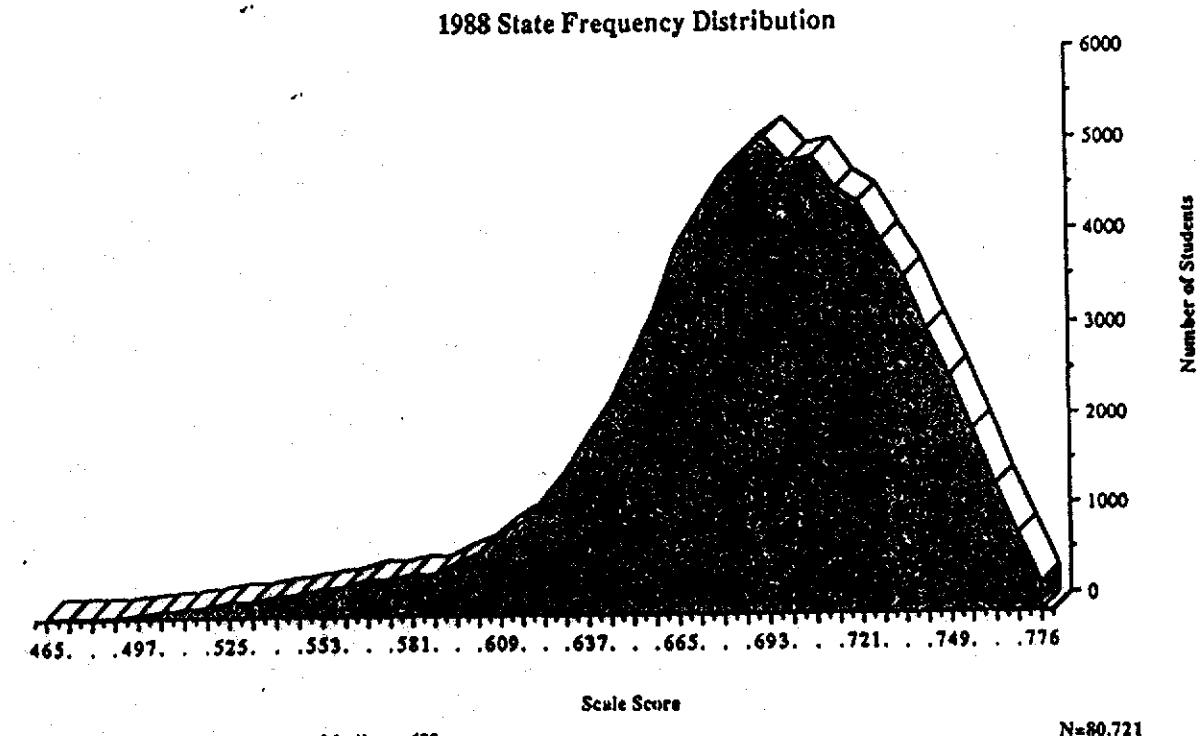


Figure 3

Table Presentation of Indicator Data

North Carolina Annual Testing Program Basic Skills:

Report of Student Performance, 1986-1988, p. 9

Table 3

COMPARISON OF NORTH CAROLINA AND THE NATION ON SELECTED EDUCATIONAL AND RELATED STATISTICS

		North Carolina	United States
Average Per Capita Income (How North Carolina Ranks Educationally Among the Fifty States)	1985	$11,617	$13,867
	1980	7,819	9,521
	1970	3,207	3,921
	1960	1,574	2,223
	1950	.....	.....
Per Capita State and Local Government Expenditures for Local Schools (including Capital Outlay) (*U.S. Department of Commerce, Bureau of Census, Government Finances, 1981) (**How North Carolina Ranks Educationally Among the Fifty States)	**1985	473.93	552.85
	*1980	350.80	410.28
	**1970	142.87	184.35
	**1960	65.11	84.27
	1950	.....	.....

Since the performance of North Carolina students is being compared with a national norm, a comparison of the state with the nation on selected important variables may contribute to a better understanding of the North Carolina results. Table 3 compares North Carolina to the nation on per capita income, average educational level, and per capita expenditures for public schools.

During the past twenty years, North Carolina has remained one of the poorer states in terms of per capita income. While the real difference between North Carolina and the national per capita average income has increased ($649 in 1960 to $1,702 in 1980), the relative difference has decreased from 41 percent to 22 percent. A review of state and local expenditures for education also reveals that North Carolina has historically been below the national average. During this time, the educational level of adults has increased greatly in North Carolina and the nation and at about the same rate. The number of handicapped students receiving services from public schools has also increased, particularly in the last decade. To bring North Carolina's per pupil rate of expenditure for the public schools to the national rate (when it has been about twenty percent less during this time) would require $510 for each of some 1,121,357 students or just under 575 million dollars ($571,892,070).

Figure 4

Text Discussion of Indicator Data

Maine Educational Assessment Mathematics: Interpretive Summary of Results, p. 5

Teacher Training

Students whose mathematics teachers have a strong mathematics background perform considerably better than students of teachers without such a background. Similarly, students perform better when their teachers are actively involved in professional development activities. Unfortunately, the majority of mathematics teachers in Maine do not participate in such activities on a regular basis and, at the lower grades, teachers with strong math backgrounds are rare.

Fifty-two percent of the grade 4 teachers have taken one or fewer courses in mathematics education. At grade 8, one-half of the math teachers do not consider themselves primarily mathematics teachers. One-fourth of the teachers at grade 11 have taken no courses in mathematics education and another 28 percent have had only one such course. It is not particularly surprising that so few of the teachers at the lower grades have a strong math background, and the large number of teachers at the upper grades who have had little or no training in math education may be related to the length of service for much of the teaching staff in the state of Maine. These teachers probably received their credentials before methods courses were required. Many of these teachers appear to be doing very little to improve their teaching skills and to increase their awareness of current research in mathematics education.

More than two-thirds of the mathematics teachers at grade 11 are not members of NCTM. Approximately one-half of the teachers at all three grade levels have not attended a workshop or conference, or taken a class in mathematics or mathematics education in the past two years. Schools are not providing mathematics inservice regularly. Three-fourths of the teachers at grades 4 and 8 report that little or no mathematics inservice has been provided by their schools in the past two years.

Figure 5

Graphic and Textual Presentation of Indicator Data

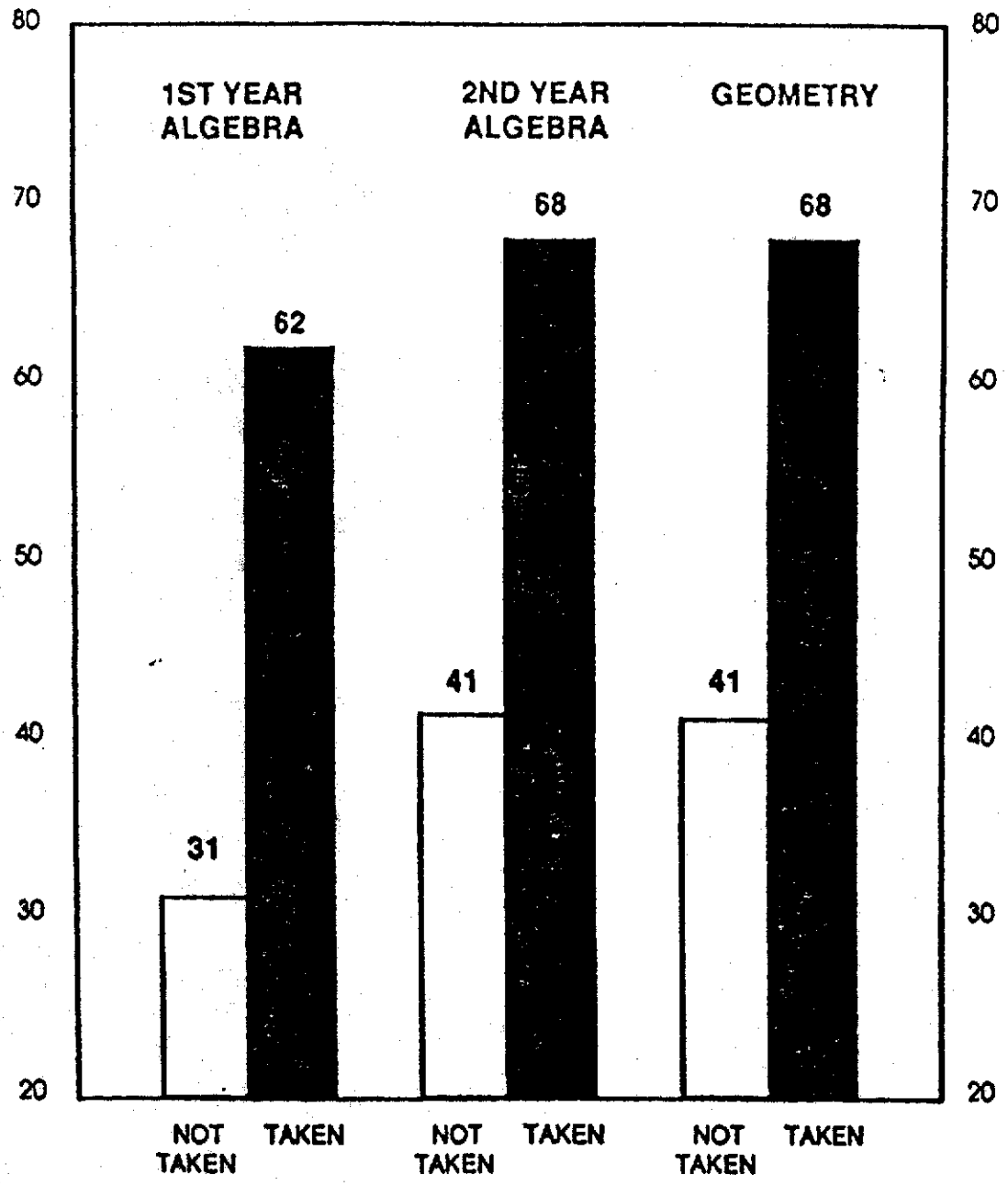
Utah Statewide Educational Assessment: General Report, 1987, p. 40, 42-43

The number of mathematics courses taken was related to students' achievement on the CTBS Mathematics test. In fact, students' Mathematics Total scores were very highly correlated (r = .61) with their scores on the Mathematics Class Experience scale. The magnitude of this relationship is exemplified in the results displayed in Exhibit 19. As can be seen from this exhibit, students who had taken certain mathematics courses had much higher Mathematics Total scores than students who had not. For example, students who had taken first year algebra typically had scores at the 62nd national percentile, while students who had not taken this course typically had scores only at the 31st percentile. Students who had taken second year algebra typically scored at the 68th percentile, while those who had not taken this course scored at only the 41st percentile. Similarly, students who had taken geometry typically scored at the 68th percentile while students who had not taken this course typically scored at only the 41st percentile. This same impact is shown in the mean raw scores displayed at the bottom of this exhibit. To some extent, the observed impact of taking more mathematics courses is inflated, because it is the more talented students who are encouraged to take more demanding mathematics courses. These students would be expected to get higher scores in any case. However, this phenomenon accounts for only a part of the observed difference in scores, especially because so many students are now taking these mathematics courses. To a considerable extent, the practice of taking more mathematics courses generally

Figure 5 (cont'd)

EXHIBIT 19

RESULTS ON CTBS MATHEMATICS TOTAL SCORE FOR STUDENTS WHO HAVE TAKEN VS. HAVE NOT TAKEN SPECIFIC MATHEMATICS COURSES, GRADE 11



	1st Year Algebra		2nd Year Algebra		Geometry	
	Not Taken	Taken	Not Taken	Taken	Not Taken	Taken
MEAN RAW SCORE	43.3	64.3	50.7	68.3	50.8	68.0

Figure 6

Sample Executive Summary

Basic Skills Performance of Alaska's Students

Achievement Test Results for 1984-85; 1985-86; 1986-87; and 1987-88, p. iii

EXECUTIVE SUMMARY

This report, entitled Basic Skills Performance of Alaska's Students, is the second annual report published by the Department of Education about the achievement of Alaska's public school students. The first report, released in February 1988, reported each of Alaska's 55 school districts' nationally standardized achievement test results for the 1985, 1986 and 1987 school years.

This year's report follows a similar format, but also adds the results of each school district's 1988 standardized achievement test results.

This year's report also updates written narratives describing the nature of each district's students, classrooms, graduation requirements and testing programs.

The test results are reported by district and for the state as a whole and are grouped into four grade levels: kindergarten through third grade; fourth through sixth grades; seventh and eighth grades; and the four high school years.

The 55 school districts administered tests from seven major national tests available from test publishers: the California Achievement Test; the Comprehensive Test of Basic Skills, the Iowa Test of Basic Skills/Test of Academic Proficiency, the Metropolitan Achievement Test, the SRA Achievement Series; the Stanford Achievement Test; and the Survey of Basic Skills. Each district administered only one of those tests, making comparisons between school district test scores misleading.

In 1988, test results were available for 68,618 students in Mathematics, 67,676 students in Reading Comprehension, 66,878 students in Language Arts, 54,346 students in Mathematics Problem Solving, and 27,314 students in Reading.

The major conclusions of the report are:

1. In 1987-88, the average scores of Alaskan students were consistently higher than the national average in virtually every content area.

2. In virtually every content area, the majority of Alaskan students scored above the national average.

3. The patterns found in the 1987-88 results were very similar to those found in the previous three years: the performance of Alaska's students on average has been consistently higher than the national average and the majority of Alaska's students scored above the national 50th percentile.

4. Within districts, the trends in student performance has varied. Some districts scored consistently high over the four year period, while others scored consistently low. Districts scored low where there were high concentrations of students whose families are poor and who live in homes where English is a second language.

5. On a statewide basis test scores remained consistent with previous years. In other words, statewide scores remained virtually the same. But, changes within grades and content areas for individual school districts were mixed. Approximately, 40% of the scores increased, 38% decreased, and 22% stayed the same.

The strength of the report is the individual district profiles for the years 1984-85, 1985-86, 1986-87 and 1987-88. Because of the wide variation of local testing practices and the differences in tests across the State, comparisons between districts and to the State average should not be made.

Future plans for statewide testing call for the adoption of a uniform test of reading, mathematics and language arts. The test will be administered for the first time in the fall of 1989. The results of the new testing program will be published in January, 1990.

As tests do not measure all important outcomes of schooling, future plans also call for the development of a broader based accountability system that collects more information about students, faculty, schools and school districts. This will include information on student dropouts and post-high school performance and the development of more in-depth individual district profiles.

Figure 7

Sample Report Highlights

What Is the Penny Buying for South Carolina?

Assessment of the Third Year of the South Carolina Education Improvement Act of 1984, p. 5

Report Highlights

In 1987, three years after the passage of the Education Improvement Act of 1984, many programs have been implemented statewide, while others are being implemented or are in the pilot-testing stage. This report examines indicators such as trends in student performance, student, teacher, and parent opinions, program implementation, and studies by outside groups in order to assess the status and impact of the EIA.

Many positive changes in education in South Carolina may be attributed to the EIA and related legislation, but problem areas remain, and certain programs need refinement. The highlights of this assessment report, presented below, provide an overview of the status and impact of the EIA through the end of school year 1986-87.

CHAPTER 1: ACADEMIC ACHIEVEMENT

- Performance on the BSAP and the CTBS improved for all grades and in six areas, only writing scores in grades six and eight did not improve.
- Test scores of students who have usually been "left behind" by reform, students of lower socioeconomic status, improved in relation to other students' performance.
- Scholastic Aptitude Test (SAT) total scores continued to improve, while the national scores remained unchanged. However, the percentage of students scoring above 600 points on the mathematics and verbal portions of the test remained about half of the national figures.
- Student attendance remained high at 96%, an average of 7.2 days absent; this figure placed the state fourth nationally.
- The dropout rate of students prior to high school graduation remained between 20-25%.

CHAPTER 2: SERVICES TO STUDENTS

- The half-day programs for four-year-olds had a positive effect on certain groups of children, and effectiveness was highly related to degree of implementation.
- Impressive gains were made by students in the 1985-86 remedial and compensatory programs (statewide average NCE gains ranged from 7.92 to 4.42). Visits to 20 elementary school compensatory programs found that principals and program and nonprogram teachers were nearly unanimous in their enthusiasm for the program.
- Enrollment in the gifted and talented programs in grades three through eight, where the majority of students are served, ranged from 7.9% to 11.6% of the student population. Almost 10,000 more students were served by these programs in 1986-87.
- Of the students who completed vocational education programs in 1985-86, 43% were employed in areas related to their vocational program training, an increase from the 37% reported last year. An additional 30% were in a related area of training in the military or in higher education, a figure similar to last year's rate.
- The percentage of examination scores qualifying students for college credit on the Advanced Placement Examinations increased to 51% in 1987, up from 48% the previous year. The number qualifying for credit was greater than the total number of exams taken in 1984, prior to the EIA.

Figure 8

Presentation Not Requiring Transformation by Reader

School Achievement: How Massachusetts Students and Their Schools Measure Up

Report on the 1988 Massachusetts Educational Assessment Program

- Although students can identify specific terms or concepts when they are presented in textbook examples, when given incidents from current events, they fail to make the connection.

  Half of the students in grade 8 and sixty percent in grade 12 were familiar with the concept of checks and balances and could identify ways in which this control works. However, they were not as skilled in identifying applications in actual situations, as illustrated in the example.

% students gr. 12	In 1987, the Senate blocked President Reagan's attempts to have Robert Bork appointed to the Supreme Court. The action was an application of
26	A. the power of veto.
5	B. minority dissent.
31	*C. checks and balances.
39	D. judicial review.

Figure 9

Explanation of Common Statistical Misunderstanding

The Results of the Fall 1988 Colorado Student Assessment Program: Special Analysis of Achievement by Public School Finance Categories, p. 2

To provide a profile of the "average" Colorado student at each of the three grades, the results presented in this report represent the average national percentile rank for students of the mean standard score in each learning area covered by the basic test battery. The mean standard score is simply the arithmetic average of the standard scores for all students included in the group.

A percentile rank of 50 indicates that the score is higher than the score obtained by fifty percent of the students taking the same test nationwide as a part of the norming or standardization sample. It does NOT mean that 50 percent of the items have been answered correctly.

Percentile ranks represent an ordinal scale of measurement. Their purpose is to rank or order students. Percentile ranks should NOT be added, subtracted, or averaged. For example, it is inappropriate to add the five scores reported for a student groups at grades 4 and 7, divide by five, and claim to have an "average" percentile rank or composite score.

Composite scores for student groups have not been included in this report because they do not provide meaningful information about achievement in a specific learning area or major components of a learning area. However, the composite scores are available upon request from the Program for Educational Quality office.

Figure 10

Explanation of Scores and Interpretation

Washington Statewide Assessment: General Report Grades 4, 8, and 10—Fall 1988, p. 3-4

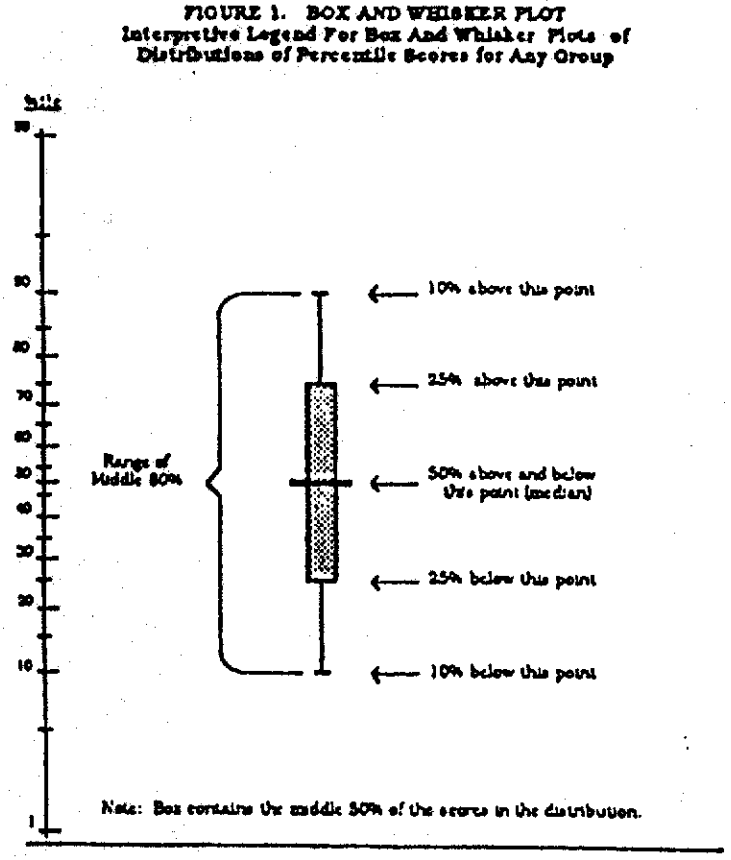
The median national percentile score describes the rank of the middle or average Washington student's score compared with the scores of students in the MAT6 national norm group. For example, a Washington median national percentile score of 62 means that, on a particular test, the middle or average student in Washington scored as high or higher than 62 percent of the students in the MAT6 norm group. Said another way, the score of the middle student in Washington (50% of Washington's scores are higher, 50% are equal to or lower) was selected as the "typical" score to represent Washington students' performance generally, compared to the national norm group performance.

By itself, a median national percentile score, or any "average" score, provides limited, and frequently misleading, information about a group's performance. Therefore, the following pages report Washington students' performance using a special graphic called a "box and whisker" plot. The box and whisker plot does report the median national percentile score for each group, but it also describes more completely <u>all</u> of the students' scores.

With the box and whisker plot, one can, for example, see how high the highest and low the lowest scores were--the "spread" of scores. One can also readily compare the same group's high, average, and low scores on different tests or the high, average, and low scores of different groups on the same test. Still further, one can analyze the same group's growth over time, as well as, see the effects of special programs on the high, average, or low students in the group.

Figure 1 is an interpretive legend for the box and whisker plots that follow. In the box and whisker plots, the median national percentile score for each group is printed in the box at the horizontal bar which locates the median score relative to the vertical percentile scale.

Figure 10 (cont'd)



By way of comparison, the box and whisker in Figure 1 also describes the performance of the MAT6 national norm group on all subtests. That is, the national norm group's median percentile score is 50 for all tests. The "box" for the national norm group describes the middle 50 percent of the scores in the norm group ranging from a percentile rank of 75 down to 25. The "whiskers" for the national norm group always extend up from the box to 90 and down to 10. To the extent that Washington's performance differs from the MAT6 norm performance, the box and whisker plots will reflect those differences through the range of scores.

Figure 11

Explanation of Technical Terms

North Carolina Annual Testing Program Basic Skills:

Report of Student Performance, 1986-1988, p. 175

Normal Curve Equivalents The normal curve equivalent (NCE) scale, ranging from 1 through 99, coincides with the national percentile scale at 1, 50, and 99. NCEs have many of the same characteristics as percentile ranks but have the additional advantage of being based on an equal-interval scale. That is, the difference between two successive scores on the scale has the same meaning throughout the scale. This property allows meaningful comparisons between different achievement test batteries and between different tests within the same battery. NCEs obtained by different groups of students on the same test or test battery can be compared by averaging the scores for the groups.

Grade Equivalents Whereas percentile ranks, normal curve equivalents, and stanines indicate relative standing within a particular grade level, grade equivalents (GE) extend across grade levels. The scale for grade equivalents ranges from 0.0 through 12.9, representing the thirteen years of school (K through 12) and the ten months in the traditional school year. September is taken as the beginning of the school year (.0); October is represented on the scale as .1, November as .2, and so on until June (.9). A grade equivalent represents the grade and month in school of students in the norm group whose test performance is *theoretically equivalent* to the test performance of given student. For the California Achievement Tests, grade equivalents are based on standardization administrations in October (.1) and May (.8). Grade equivalents for other t the school year are interpolated (mathematically estimated) from these two points.

To illustrate the calculation of grade equivalents, suppose that a vocabulary test i during November to a norming group of sixth-grade students and that the me obtained is 724. Then the grade equivalent for a scale score of 724 on that test 6 representing Grade 6 and .2 representing the month of November. Any s scale score of 724 on that vocabulary test would be given a grade equiva skills.

Caution should be exercised in interpreting grade equivalents. If grade equivalent of 4.8 on a reading test, it does not mean th reading skills that are taught in the school district during t means only that the student's performance on this test is performance of students in the norm group who had c

Figure 12

Sample Graphics

What Is the Penny Buying for South Carolina?

Assessment of the Third Year of the South Carolina Education Improvement Act of 1984, p. 12 & 57

GRAPH 1-1 — COGNITIVE SKILLS ASSESSMENT BATTERY
Percent of Five-Year-Old Students Tested as "Ready" for First-Grade Work

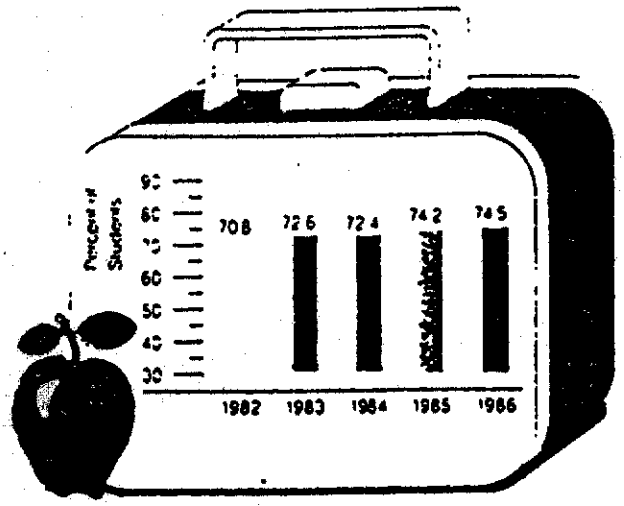


TABLE 3-4 — RELATIVE IMPORTANCE OF CRITERIA IN THE SELECTION OF PRINCIPALS

Rank Order 1987*	Superintendents
1	Interview performance with superintendent and school board (1)
2	Previous job experience (2)
3	Superintendent or board recommendation (3)
4	Assessment Center Report (6)
5	Length and quality of previous service (4)
6	Recommendation from special committees (8)
7	Letters of recommendation (7)
8	Undergraduate or graduate degrees (5)

Figure 13

Question and Answer Section

Michigan Educational Assessment Program Handbook, p. 15

- Can newspaper reporters or real estate agents, for example, use MEAP test results alone to provide a complete and accurate picture of a school or district?

No. Because both of these groups of people have influence on the decisions citizens make...

Figure 14

Questions and Answer Format

Basic Skills Performance of Alaska's Students

Achievement Test Results for 1984-85; 1985-86; 1986-87; and 1987-88, p. 2

Do all of Alaska's school districts use the same test?

No. Seven major tests are used in Alaska. These tests are:

California Achievement Test
Comprehensive Test of Basic Skills
Iowa Test of Basic Skill/Test of Achievement and Proficiency
Metropolitan Achievement Test
SRA Achievement Series
SRA Survey of Basic Skill
Stanford Achievement Test

Why are there a variety of tests in use in Alaska?

Districts select or continue to use a test series for three kinds of reasons: practical, technical, and alignment.

Practical reasons: Tests differ in their format, administration procedures, costs and publisher's services. Districts often continue to use a test series over time to maintain a long term data base and because they are satisfied with the level of service they are receiving from a testing company.

Technical reasons: Tests differ in the composition of their norming group. For example, some tests are normed in predominantly urban settings, while other tests can provide separate norms for specific subgroups of students. Districts review the characteristics of the norming group and the kinds of normative data that are available in choosing their tests.

Alignment reasons: There are differences in content emphasis among the various standardized achievement tests. Districts review the items and objectives of tests to determine how well the content and emphasis of the tests match the content and emphasis of their local curriculum.

What are the kinds of skills that are tested on standardized achievement tests?

Standardized tests typically assess student performance in the basic skill areas of reading, mathematics and language arts. These areas are often considered to be the tools of learning. Often these general areas are broken down further into subskill areas like reading comprehension and mathematics problem solving.

Reading tests often include *vocabulary* items that ask students to choose the correct word from a list of words and *reading comprehension* items that ask students to read a paragraph and then answer questions about what they have read. In primary grades, *reading* tests often include items which ask students to listen to two words and tell whether they are the same or different and to find letters of the alphabet.

Mathematics tests often include *computation* items which ask students to add, subtract, multiply or divide numbers; items which test how well students understand the number system and the terms used in mathematics (*concepts*), for example, asking students which of four numbers is not a prime number; and items which ask students to read a short story problem and choose the correct answer (*problem solving*).

Language Arts tests often include *spelling* items that ask students to choose words that are misspelled; *capitalization* items that ask students to identify words that should be capitalized; *punctuation* items that assess students' ability to use commas, question marks, periods, etc.; and *usage* items which ask students to demonstrate their skills in using pronouns, verbs, adjectives, etc.

While reading, mathematics and language arts are the general areas most often tested, each standardized test takes a different approach to what specific objectives should be tested, how many items should be included, and what grades should be tested with the same test.

How accurate are standardized tests? Are they culturally biased?

Test publishers devote a large amount of resources to the preparation of their standardized tests. Usually the tests provide highly reliable results and are based on a careful scrutiny of basic skills curriculum across the country.

Standardized tests are less accurate for students within an age/grade grouping whose performance is either very high or very low. For these groups of students, test results are less reliable.

Figure 15

Simple, Direct is Better

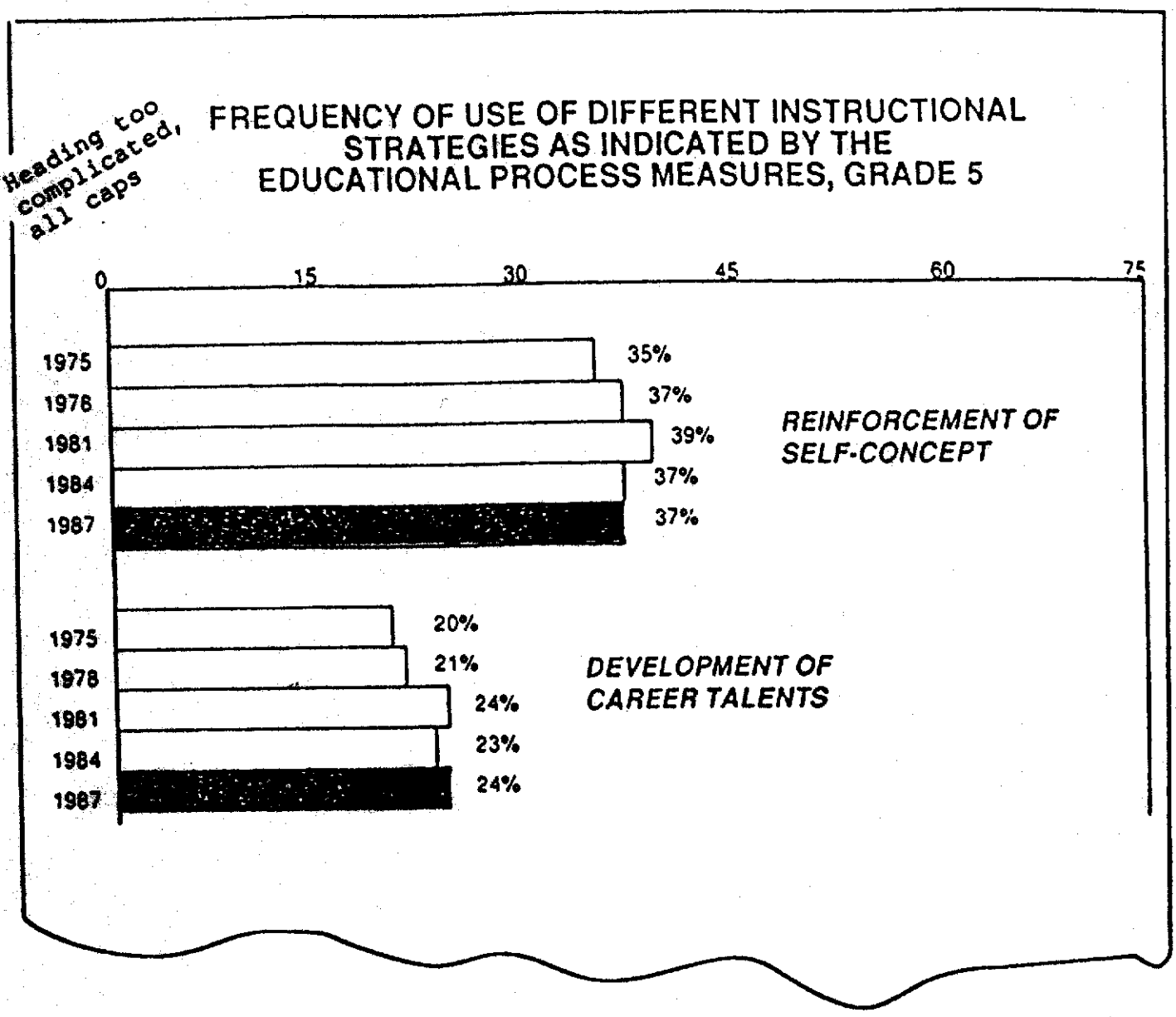


Figure 15 (cont'd)

Simpler heading, upper & lowercase, footnote explains

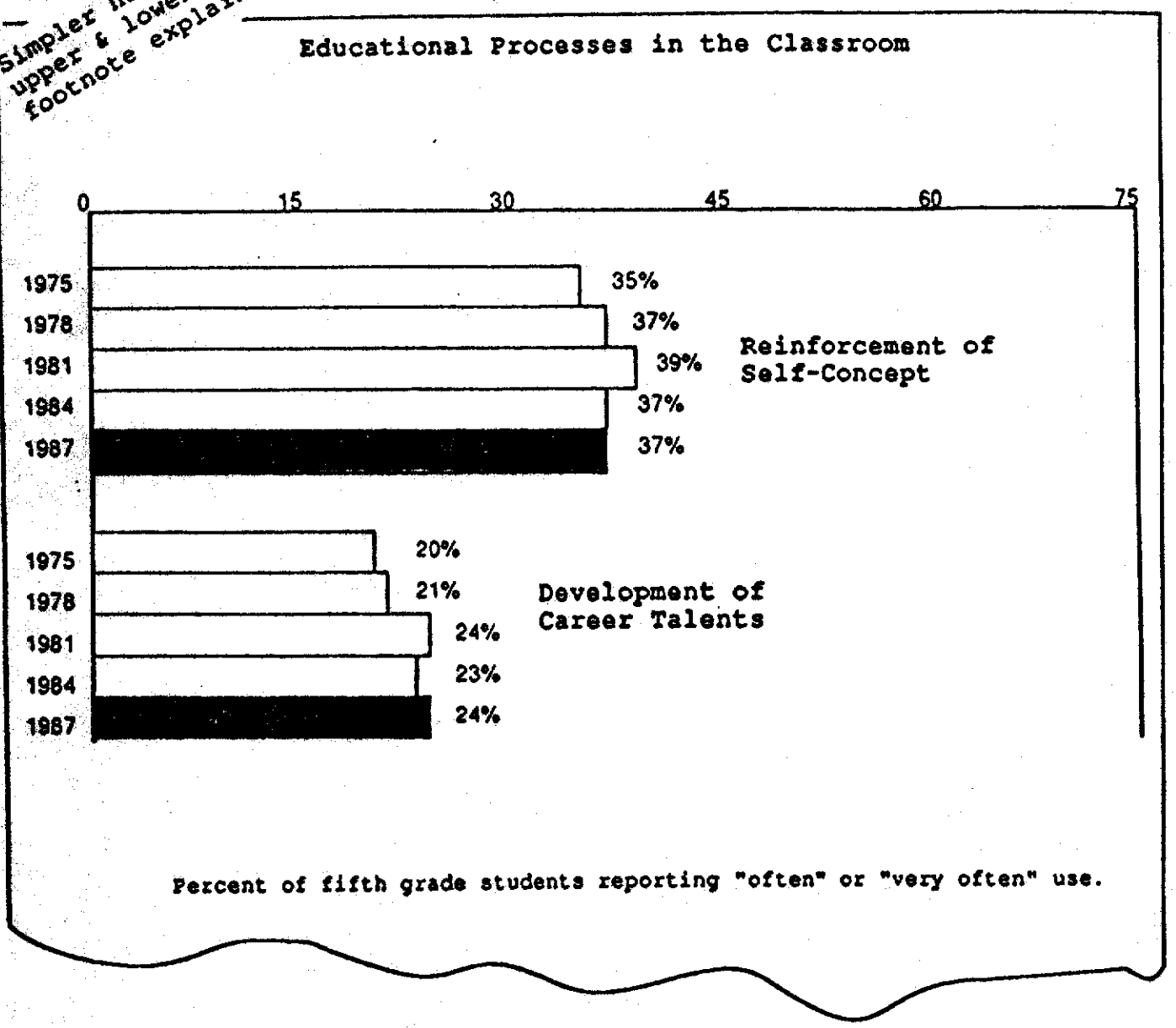


Figure 16

Graph Headings: Question Format

Performance Profiles: Illinois Schools Report to the Public School Report Card, 1986, p. 13