Journal of Educational and Behavioral Statistics
Spring 1997, Vol. 22, No. 1, pp. 1–30

Improving Tabular Displays, With NAEP Tables as Examples and Inspirations

Howard Wainer
Educational Testing Service

Keywords: *graphical comprehension, NAEP, tabular displays*

The modern world is rich with data; an inability to effectively utilize these data is a real handicap. One common mode of data communication is the printed data table. In this article we provide four guidelines the use of which can make tables more effective and evocative data displays. We use the National Assessment of Educational Progress both to provide inspiration for the development of these guidelines and to illustrate their operation. We also discuss a theoretical structure to aid in the development of test items to tap students' proficiency in extracting information from tables.

In his 1786 atlas of England and Wales, William Playfair wrote of the increasing complexity of modern life. He pointed out that when life was simpler and data were less abundant, an understanding of economic structure was both more difficult to formulate and less important for success. But by the end of the 18th century, this was no longer true. Statistical offices had been established and had begun to collect data on which political and commercial leaders could base their decisions. Yet the complexity of these data precluded their easy access by any but the most diligent.

Playfair's genius was in surmounting this difficulty through his marvelous invention of statistical graphs and charts. The complexity of life within 18th-century Britain and the massiveness of available data were but trifles in comparison to today's complex network of data sources and topics. These data are being transformed into graphic forms at a breathless pace.

Today we have the need to clearly and accurately display summaries of huge amounts of information. Computing equipment, software, and electronic networks provide the means to summarize information and disseminate results. What we lack is a broad understanding of how best to do it. In this article we examine the data table as a communicative display and suggest four steps which, if followed, can allow tables to communicate better.

In this report we focus principally on the table because we heartily subscribe to the notion that although accurate data can help us to understand the world, they can help only

> if they are properly interpreted. There can be no assurance of a proper interpretation, however, unless the arrangement of the data on the printed page is clear, logical, complete, and properly focused. . . . Incidentally, it is our conviction, tested in experience, that language flows more easily and

> logically from the pen of him whose tabulated data reflect careful and precise thinking. (Walker & Durost, 1936, p. iii)

I have two goals in this article. The primary one is to provide concrete guidelines for constructing more communicative tables. My second goal is more limited. Because tables are used so often to communicate information, it is generally felt that the ability to understand tables is an important skill for a literate citizen. It is not uncommon to find tables used as stimuli within many tests of basic reading or mathematical skills. Thus my second goal is to illustrate how improved tabular presentation immediately lends itself to expanding the range of test questions that can be asked.

For reasons that will become clearer by the end of this article, we have chosen to intertwine a general discussion of tabular display with a discussion of the use of tables within the National Assessment of Educational Progress (NAEP)[1]—both as stimuli in test items and as communicative media in published reports.

To build any effective display we must have a firm notion of purpose. We cannot know what the best answers are unless we know what the questions are. Thus we must first understand what questions will be asked of the data. Any discussion of data display in the abstract is pointless. To aid in understanding the intended purposes for tabular displays, we shall look at the sorts of test questions that NAEP pairs with data tables. It does not seem far-fetched to assume that the sorts of questions NAEP experts believe children ought to be able to answer from tables are the same sorts that everyone else should be facile with.

This exercise is what naturally led us to the second goal, oftentimes intermingled in this report with the first, of expanding the range of test questions that can be posed when data are well displayed. The advice offered here comes from many sources, filtered through me. To the extent that I add anything to the wisdom of those who have preceded me on this path, it is in the attention devoted to depicting error.

Tabular Presentation

> Getting information from a table is like extracting sunlight from a cucumber. (Farquhar & Farquhar, 1891, p. 55)

The disdain shown by the two 19th-century economists quoted above reflected a minority opinion at that time. As commonly prepared, tables, spoken of so disparagingly by the Farquhars, remain to a large extent worthy of contempt.

Before we explore ways to improve a tabular display, it is wise to be explicit about the likely audience and goals of the display. In this report we examine tables within NAEP that are aimed at three separate audiences: children, the lay public, and education professionals. While it might seem that this diversity of audience and associated goals ought to yield quite different structures for their displays, it appears that the requirements of their

shared cognitive and perceptual apparatus dominate their differences in age, training, and interests. The sets of rules for table construction that emerge for the three groups are virtually identical in general structure and vary only because of constraints imposed by the increasing complexity of the data themselves.

Why are tables used to display data? All data displays, including tables, are used for one or more of four purposes:

(1) Exploration. Data can contain answers to questions that may be explicit in the viewer's mind or not. Data exploration answers explicit questions while posing questions previously unthought of.
(2) Communication. Once the data are explored they can be displayed to convey what has been discovered to a broader audience.
(3) Storage. Data are expensive to gather; once they have been gathered, it is usually imprudent to lose them. In the past they have been stored for future use in various sorts of data displays.
(4) Decoration. Data displays are often used to enliven a presentation. Indeed, conversations with reporters on the use of graphics invariably center around how to locate a display to attract the eye of the reader.

A principal tenet of effective data display is that before designing a display one must establish a hierarchy of purpose and not try to do too much. A display aimed at communication should not try to serve an archival purpose as well, since rules governing these two purposes are often antithetical.

The initial collection, and hence display, of most data sets begins with a data table. Thus any discussion of display should start with the table as the most basic construction. Rules for table construction are often misguided, aimed at the use of a table for data storage rather than data exploration or communication. The computer revolution of the past 30 years has obviated the need for archiving of data in printed tables, but rules for table preparation have not been revised apace with this change in purpose. Modern data storage is accomplished well on magnetic disks or tapes, optical disks, and other mechanical devices. Paper and print are meant for human eyes and human minds.

Helen Walker and Walter Durost (1936) provided a careful description of guidelines for the construction of statistical tables. Ehrenberg (1977) amplifies some of these rules to allow tables to become a still more effective multivariate display. Among his rules are (a) rounding heavily, (b) ordering and grouping the rows and columns by some aspect of the data, (c) framing the display with suitable summary statistics, and (d) spacing to aid perception. More recent work on effective tabular presentation (Clark, 1987; Wainer, 1992, 1993) elaborates and illustrates these simple rules for designing effective tables.

We shall begin this discussion with a more detailed statement and justification of these four rules of effective tabular display within the context of tabular displays in NAEP test items. When this is complete, we shall then go on to do the same thing for larger tables used principally for communication

in NAEP reports, although they are often expected to serve archival purposes as well.

Tables as Part of NAEP Items

Example 1: 1992 12th Grade Math, Questions 3 and 4

This example shows how rounding table entries makes a difference. The original table on which Questions 3 and 4 were based is:

POPULATIONS OF DETROIT AND LOS ANGELES
1920-1970

Year	City	
	Detroit	Los Angeles
1920	950,000	500,000
1930	1,500,000	1,050,000
1940	1,800,000	1,500,000
1950	1,900,000	2,000,000
1960	1,700,000	2,500,000
1970	1,500,000	2,800,000

The two questions (omitting the alternatives offered) were:

3. How many more people were living in Los Angeles in 1960 than 1940?
4. What was the first year listed in which the population of Los Angeles was greater than the population of Detroit?

If we round to two digits (the nearest hundred thousand) and tidy up the display a bit, we get:

Populations For Two Cities
Years: 1920-1970
Units: Millions

Year	Detroit	Los Angeles
1920	1.0	0.5
1930	1.5	1.1
1940	1.8	1.5
1950	1.9	2.0
1960	1.7	2.5
1970	1.5	2.8

The answer to Question 3 is clearly 1 million, and the answer to Question 4 is 1950. It awaits empirical verification whether this is easier than before revision, but my intuition (and my twelve-year-old son) certainly suggests so.

Why did I suggest rounding to two digits? Let us explore this in a discussion of the first rule of table construction:

Rule I. Round—a lot! This is for three reasons:

- Humans cannot understand more than two digits very easily.
- We can almost never justify more than two digits of accuracy statistically.
- We almost never care about accuracy of more than two digits.

Let us take each of these reasons separately.

Understanding. Consider the statement "This year's school budget is $27,329,681." Who can comprehend or remember that? If we remember anything, it is almost surely the translation "This year's school budget is about 27 million dollars."

Statistical justification. The standard error of most statistics is proportional to 1 over the square root of the sample size.[2] God did this, and there is nothing we can do to change it. Thus suppose we would like to report a correlation as .25. If we don't want to report something that is inaccurate, we must be sure that the second digit is reasonably likely to be 5 and not 6 or 4. To accomplish this we need the standard error to be less than .005. But since the standard error is proportional to $1/\sqrt{n}$, the obvious algebra ($1/\sqrt{n} \sim .005$, therefore $\sqrt{n} \sim 1/.005 = 200$) yields the inexorable conclusion that a sample size of the order of 200^2 or 40,000 is required to justify the presentation of more than a two-digit correlation. A similar argument can be made for most other statistics.

Who cares? I recently saw a table of average life expectancies that proudly reported the mean life expectancy of a male at birth in Australia to be 67.14 years. What does the 4 mean? Each unit in the hundredths digit of this overzealous reportage represents 4 days. What purpose is served in knowing a life expectancy to this accuracy? For most communicative (not archival) purposes, 67 would have been enough.

The effects of too many digits is sufficiently pernicious that I would like to emphasize the importance of rounding with another short example. The following equation is taken from *State Court Caseload Statistics: Annual Report, 1976* (Court Statistics Project, 1976):

$$\ln (DIAC) = -.10729131 + 1.00716993 \times \ln (FIAC),$$

where *DIAC* is the annual number of case dispositions, and *FIAC* is the annual number of case filings. This is obviously the result of a regression analysis with an overgenerous output format. Using the standard error justification for rounding we see that to justify the eight digits shown we would

need a standard error that is of the order of .000000005, or a sample size of the order of 4×10^{16}. This is a very large number of cases—the population of China doesn't put a dent in it. The actual n is the number of states, which allows one digit of accuracy at most. If we round to one digit and transform out of the log metric we arrive at the more statistically defensible equation

$$DIAC = .9\ FIAC.$$

This can be translated into English as "There are about 90% as many dispositions as filings." Obviously the equation that is more defensible statistically is also much easier to understand. My colleague Al Biderman, who knows more about courts than I do, suggested that we needed to round further, to the nearest integer ($DIAC = FIAC$), and so a more correct statement would be "There are about as many dispositions as filings." A minute's thought about the court process reminds one that it is a pipeline with filings at one end and dispositions at the other. They must equal one another, and any variation in annual statistics reflects only the vagaries of the calendar. The sort of numerical sophistry demonstrated in the first equation can give statisticians a bad name.[3]

Example 2: 1990 8th and 12th Grade Science Assessment

Any redesign task must first try to develop an understanding of purpose. The presentation of the data set in Table 1 must have been intended to help the reader answer such questions as:

(1) What is the general level (in hours) of battery life for the brands chosen?
(2) How do the battery brands differ with respect to their life expectancies? What's the best one? The worst?
(3) What kinds of equipment use batteries up most quickly? Least quickly?

TABLE 1
Table paired with Items 21 and 22 on the 1990 8th and 12th grade science assessment

	Battery Life in Hours			
Battery Brands	Cassette Player	Radio	Flashlight	Portable Computer
Constant Charge	5	19	10	3
PowerBat	7	24	13	5
Servo-Cell	4	21	12	2
Never Die	8	28	16	6
Electro-Blaster	10	26	15	4

(4) Are there any unusual interactions between equipment and battery brand?

These are obviously parallel to the questions that are ordinarily addressed in the analysis of any multifactorial table—overall level, row, column, and interaction effects.

By characterizing the information in the table in this way we are able to explicitly lay out areas of questions that might be asked about these data in an effort to determine the extent to which students can understand data presented in a table. In fact, there were three questions that followed this table, but only one asked about the data, and it was parallel to Question 2:

21. On the basis of the information in the table, which brand do you think is the best all-purpose battery? (Assume all batteries cost the same.)

The next question asked about how the student made this determination:

22. Briefly explain how you used the information in the table to make your decision.

Before going further, I invite you to read Table 1 carefully and see to what extent you can answer the four questions. But don't peek ahead!

The entries in this table are already rounded, so we can go directly to the second rule of table construction:

Rule II. Order the rows and columns in a way that makes sense. Alphabetical order is almost never the best way to go. Three useful ways to order the data are:

- by size. Often we look most carefully at what is on top and less carefully further down. Put the biggest thing first. Also, ordering by some aspect of the data often reflects ordering by some hidden variable that can be inferred.
- naturally. Time is ordered from the past to the future. Showing data in that order melds well with what the viewer might expect. This is always a good idea.
- according to interest. If we are especially interested in comparing a particular set of rows or columns, put them adjacent to one another.

Table 2 is a redone version of Table 1 in which batteries (rows) are ordered by battery life in a radio, with the longest-lasting battery first. Types of equipment (columns) are ordered by how quickly they use up batteries, least voracious first. From this we see that by ordering by radio use we have also ordered for flashlights. There is some minor shuffling within the cassette player and computer columns. Now that the table is ordered, answering NAEP Question 21 is easy, as are most other main effect questions.

We can improve matters still further by remembering the third rule:

TABLE 2
First revision of battery life table (rows and columns ordered, extraneous lines removed)

Battery Brands	Battery Life in Hours			
	Radio	Flashlight	Cassette Player	Portable Computer
Never Die	28	16	8	6
Electro-Blaster	26	15	10	4
PowerBat	24	13	7	5
Servo-Cell	21	12	4	2
Constant Charge	19	10	5	3

Rule III. ALL is different and important. Summaries of rows and columns are important as a standard for comparison—they provide a measure of usualness. What summary we use to characterize *all* depends on the purpose. Sometimes a sum or a mean is suitable, more often a median. But whatever is chosen it should be visually different from the individual entries and set spatially apart.

The summaries (means) surrounding Table 3 make the row and column effects explicit. Now we not only see that the Never Die battery the best all around, but we have a measure of how much better it is. We also see that a computer uses batteries about 6 times as fast as a radio.

Can we go further? Sure. To see how requires that we consider what distinguishes a table from a graph. A graph uses space to convey information. A table uses a specific iconic representation. We have made tables more understandable by using space—making a table more like a graph. We can improve tables further by making them more graphical still. A semigraphical display like the stem-and-leaf diagram (Tukey, 1977) is merely a table in which the entries are not only ordered but are also spaced according to the size of the gaps between adjacent rows or columns. The rule then is:

TABLE 3
Second revision of battery life table (row and column means shown and emphasized)

Battery Brands	Battery Life in Hours				
	Radio	Flashlight	Cassette Player	Portable Computer	**Battery Averages**
Never Die	28	16	8	6	**15**
Electro-Blaster	26	15	10	4	**14**
PowerBat	24	13	7	5	**12**
Servo-Cell	21	12	4	2	**10**
Constant Charge	19	10	5	3	**9**
Usage averages	**24**	**13**	**7**	**4**	**12**

Rule IV. Add spacing to aid perception. If there is a clustering among rows or columns, space them so that they look clustered. To put this notion into practice, consider the next version of Table 1, shown as Table 4.

The rows have been spaced according to what appear to be significant gaps (Wainer & Schacht, 1978), and we see that batteries fall into two groups: three relatively strong batteries and two weaker ones. This yields a table that is about as good as we can do. Now we can see that a battery lasts about twice as long in a radio as in a flashlight, and about twice as long in a flashlight as in a cassette player. Moreover, we see clearly that the three best batteries yield about 50% more life than the two worst.

This brings us to an interesting issue. NAEP Questions 21 and 22 could be answered trivially if the table were transformed as we have done in Table 4. Should we transform the table? The way in which we have structured the table is not based on the particular questions that were asked, but rather on general rules for all tables. We would have done it in exactly the same way had we not seen the questions. This transformation merely follows a set of rules that characterizes good practice. The original table was flawed in that it didn't conform to standards of good practice.

Basing a characterization of an examinee's ability to understand a data display on a question paired with a flawed display is akin to characterizing someone's ability to read by asking questions about a passage full of spelling and grammatical errors whose sentences were ordered haphazardly. What are we really testing?

One might say that we are examining whether or not someone can understand what is de facto "out there." I have some sympathy with this view, but what is the relationship between the ability to understand illiterate prose and the ability to understand proper prose? If we measure the former, do we know anything more about the latter? Yet how often do we encounter well made displays in the everyday world? Should we be testing what is, or what should be?

TABLE 4
Third revision of battery life table (rows spaced to accentuate battery clusters)

Battery Brands	Battery Life in Hours				**Battery Averages**
	Radio	Flashlight	Cassette Player	Portable Computer	
Never Die	28	16	8	6	**15**
Electro-Blaster	26	15	10	4	**14**
PowerBat	24	13	7	5	**12**
Servo-Cell	21	12	4	2	**10**
Constant Charge	19	10	5	3	**9**
Usage averages	**24**	**13**	**7**	**4**	**12**

A more practical problem is that if a display is properly constructed, most commonly asked questions are easily answered. That is the nature of graphics and human information processing ability. It is harder to ask nontrivial questions of a well constructed table. This is not an isolated issue. I will discuss it further in the conclusion of this article.

While we cannot hope to resolve these issues here, I would like to add one vote toward testing literacy with prose that is correctly composed and testing numeracy with data displays that conform to accepted standards of good practice. If we do otherwise we may be able to connect our test with common practice, but is that what we wish to do?

In the concluding section of this article I will discuss the kinds of questions that can be constructed and suggest a theoretical structure that will aid in future tests of this sort.

Example 3: 1992 4th, 8th, and 12th Grade Math Assessment

Original Table

Ten Students' Test scores

Student	Score
A	88
B	65
C	91
D	36
E	72
F	57
G	50
H	85
I	62
J	48

Revised Table

Ten Students' Test scores

Student	Score
C	91
A	88
H	85
E	72
B	65
I	62
F	57
G	50
J	48
D	36
Mean	**65**

Question 9, associated with the above table, is as follows.

9. The table above shows the scores of 10 students on a final examination. What is the range of these scores? (then four options)

To answer this question, one needs to know that the range is the difference between the largest and the smallest entries, find them, and then subtract them. A properly prepared table, which orders the rows by the data rather than some arbitrary letter, removes the need for the second step.[4] Also, introducing spaces where there are data gaps (invisible in the original table) provides the opportunity to ask other, deeper questions about the structure of these data.

Big Tables in NAEP Reports

NAEP reports are often mother lodes of information, but sometimes it takes a considerable amount of effort to mine that information. One reason that such effort is required is the format of the data presentation. It appears that saving space is sometimes viewed as a more important goal than effective communication. Let us examine a single large table from one major NAEP report and see how the application of the aforesaid four rules can increase its comprehensibility. The table chosen shares enough of its characteristics with other tables to allow one example to be broadly generalizable.

Example 4: Table 2.12 From Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States

This table, reproduced as Table 5, shows the average mathematics performance of eighth-grade examinees from all participating jurisdictions in the 1992 state mathematics assessment as a function of parent's education. Also included is the percentage of examinees in each state whose parents' education is at each of the designated levels. As is customary, the standard errors of all figures are presented in parentheses.

Before we attempt to revise this table, it is wise to consider its likely purpose. Why would anyone want to see data like these? What sorts of questions would such data answer? How easily could the reader of this table answer the same sorts of questions that were asked of children in the assessment? How hard is it to answer a question analogous to Question 21 about what is the best all-purpose battery (What is the best performing state?)? Or one analogous to Question 9 about the range of scores among 10 children (What is the range of performances among the 41 participating states?)? Any redesign should allow such obvious questions to be answered easily.

More generally, for this table, as with most two-way displays, the questions that can be answered are based on the factors presented, to wit:

(1) How did the children in each of the jurisdictions perform in math? Which states did the best? Which the worst? How much variation is there among the states? How does my state compare with others like it? With the nation as a whole? What is the clustering among the states?
(2) What is the relationship between parental education and children's math performance?
(3) Does parental education have the same effect in all jurisdictions?

In addition, there are questions parallel to these dealing with the percentage of children at each parental education level.

(4) How well educated are the parents of these children in each of the jurisdictions? Which states have the best educated parents? Which the worst? How much variation is there among the states? How does my

state compare with others like it? With the nation as a whole? What is the clustering among the states?

(5) Which level of parental education is most common? Which is least? How much parental education is "typical"?

(6) Does the distribution of parental education have the same shape in all jurisdictions?

After answering the above questions, we would like to be able to know which differences we observe are possible artifacts of sampling fluctuation and which represent real differences in the populations of interest.

TABLE 5

Original Table 2.12 from Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States *(p. 83): Average mathematics proficiency by parents' highest level of education*

	Grade 8 - 1992									
	Graduated College		Some Education After High School		Graduated High School		Did Not Finish High School		I Don't Know	
PUBLIC SCHOOLS	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency
NATION	40 (1.4)	279 (1.4)	18 (0.6)	270 (1.2)	25 (0.8)	256 (1.4)	8 (0.6)	248 (1.8)	9 (0.5)	251 (1.7)
Northeast	38 (3.1)	282 (4.2)	18 (1.1)	267 (3.0)	26 (2.2)	259 (4.2)	8 (0.9)	246 (4.2)	10 (1.2)	250 (3.3)
Southeast	35 (1.9)	270 (1.9)	17 (0.8)	263 (2.0)	28 (1.4)	249 (1.9)	12 (1.6)	246 (4.2)	8 (1.0)	248 (4.3)
Central	42 (2.7)	283 (2.9)	20 (1.4)	273 (1.6)	26 (1.7)	264 (2.3)	4 (0.7)	*** (***)	7 (0.8)	258 (3.8)
West	43 (2.9)	279 (2.6)	18 (1.2)	274 (2.6)	19 (1.5)	252 (2.9)	9 (1.1)	248 (2.4)	11 (0.9)	248 (2.9)
STATES										
Alabama	33 (1.6)	261 (2.5)	18 (0.7)	258 (2.0)	29 (1.1)	244 (1.8)	13 (0.9)	239 (2.0)	7 (0.6)	237 (2.9)
Arizona	36 (1.5)	277 (1.5)	22 (1.0)	270 (1.5)	21 (0.9)	256 (1.6)	10 (0.7)	245 (2.5)	12 (0.8)	248 (2.7)
Arkansas	30 (1.1)	264 (1.9)	20 (0.8)	264 (1.7)	31 (1.1)	248 (1.6)	11 (0.7)	246 (2.4)	8 (0.6)	245 (2.7)
California	39 (1.8)	275 (2.0)	18 (1.0)	266 (2.1)	17 (0.9)	251 (2.1)	10 (0.9)	241 (2.2)	16 (1.1)	240 (2.9)
Colorado	46 (1.2)	282 (1.3)	19 (0.9)	276 (1.6)	21 (0.9)	260 (1.5) >	6 (0.6)	250 (2.4)	7 (0.5)	252 (2.6)
Connecticut	47 (1.3)	288 (1.0) >	16 (0.8)	272 (1.8)	22 (0.9)	260 (1.8)	6 (0.6)	245 (3.3)	9 (0.6)	251 (2.4)
Delaware	39 (1.2)	274 (1.3)	18 (1.0)	268 (2.3)	30 (1.0)	251 (1.7)	6 (0.5)	248 (4.0)	8 (0.9)	248 (3.4)
Dist. Columbia	32 (1.0)	244 (1.7)	17 (0.8)	240 (1.9)	29 (0.8)	224 (1.6)	9 (0.7)	225 (3.2)	12 (0.6)	229 (2.2)
Florida	39 (1.5)	268 (1.9)	19 (0.7)	266 (1.9)	24 (1.1)	251 (1.8)	8 (0.7)	244 (2.7)	10 (0.7)	244 (3.2)
Georgia	35 (1.7)	271 (2.1)	18 (0.7)	264 (1.7)	30 (1.2)	250 (1.3)	11 (0.8)	244 (2.2)	6 (0.6)	245 (2.6)
Hawaii	38 (1.1)	267 (1.5)	15 (0.9) <	266 (1.9)	25 (1.0)	246 (1.8)	6 (0.5)	242 (3.5)	16 (0.8)	246 (2.1) »
Idaho	48 (1.2)	281 (0.9)	20 (0.8)	278 (1.3)	19 (0.9)	268 (1.4) >	7 (0.5)	254 (2.3)	6 (0.5)	254 (2.8)
Indiana	33 (1.5)	283 (1.5)	21 (0.9)	275 (1.9)	32 (1.1)	260 (1.6)	8 (0.6)	250 (2.6)	6 (0.5)	249 (3.3)
Iowa	44 (1.4)	291 (1.2) >	21 (0.8)	285 (1.5)	25 (1.1)	273 (1.3)	4 (0.4)	262 (2.4)	5 (0.4)	266 (2.8)
Kentucky	28 (1.4)	278 (1.6) »	19 (0.8)	267 (1.6)	32 (0.9)	254 (1.6)	15 (0.9)	246 (1.7)	6 (0.4)	242 (2.8)
Louisiana	32 (1.4)	256 (2.5)	20 (0.9)	259 (1.8)	30 (1.3)	242 (1.6)	10 (0.7)	237 (2.4)	7 (0.6)	236 (3.7)
Maine	40 (1.5)	288 (1.4)	22 (1.0)	281 (1.5)	26 (1.1)	267 (1.1)	6 (0.5)	259 (2.7)	5 (0.5)	266 (2.6)
Maryland	44 (1.7)	278 (1.8)	18 (0.9)	266 (1.9)	25 (1.2)	250 (1.8)	6 (0.8)	240 (3.7)	7 (0.5)	245 (3.8)
Massachusetts	48 (1.5)	284 (1.3)	17 (0.8)	272 (1.8)	21 (1.0)	261 (1.4)	7 (0.6)	248 (3.2)	7 (0.6)	248 (2.6)
Michigan	38 (1.6)	277 (2.2)	23 (0.9)	271 (2.0)	26 (0.9)	257 (1.7)	6 (0.5)	249 (2.0)	7 (0.6)	248 (3.0)
Minnesota	48 (1.3) >	290 (1.0) >	21 (0.9)	284 (1.8)	22 (0.9) «	270 (1.8) >	3 (0.4)	256 (4.2)	7 (0.6)	268 (3.0)
Mississippi	36 (1.7)	254 (1.6)	16 (0.7)	256 (2.0)	29 (1.4)	239 (1.6)	13 (0.8)	234 (1.8)	7 (0.6)	231 (2.8)
Missouri	36 (1.3)	280 (1.7)	22 (0.9)	275 (1.5)	29 (1.0)	264 (1.6)	8 (0.7)	254 (2.4)	6 (0.5)	252 (2.9)
Nebraska	46 (1.5)	287 (1.2)	20 (1.0)	280 (1.6)	24 (1.2)	267 (1.7)	4 (0.5)	247 (3.3)	6 (0.6)	256 (3.8)
New Hampshire	46 (1.5)	287 (1.4)	17 (0.8)	280 (1.5)	24 (1.1)	267 (0.9) »	6 (0.5)	259 (2.5)	7 (0.5) >	262 (2.5)
New Jersey	45 (1.6)	283 (1.8)	18 (0.8)	275 (2.1)	23 (1.2)	259 (2.5)	7 (0.6)	253 (3.8)	8 (0.7)	250 (3.9)
New Mexico	34 (1.4)	272 (1.4)	20 (0.7)	264 (1.4)	26 (1.1)	249 (1.4)	11 (0.7)	244 (1.9)	10 (0.6)	245 (2.0) »
New York	44 (1.8)	277 (1.9)	18 (1.1)	271 (2.4)	23 (1.0)	256 (2.5)	6 (0.8)	243 (4.2)	10 (1.0)	240 (3.8)
North Carolina	36 (1.2)	271 (1.4) >	20 (0.8)	265 (1.6) >	27 (0.9) <	246 (1.7)	10 (0.6)	240 (2.3)	6 (0.5)	240 (3.6)
North Dakota	54 (1.2) >	289 (1.1)	18 (0.7)	283 (1.9)	19 (1.3)	271 (1.7)	3 (0.5)	259 (4.5)	5 (0.5)	272 (2.8)
Ohio	37 (1.4)	279 (1.8)	19 (0.7)	272 (1.6)	32 (1.1)	260 (2.3)	7 (0.6)	243 (2.6)	5 (0.5)	249 (4.5)
Oklahoma	39 (1.4)	277 (1.5)	21 (0.9)	272 (1.9)	26 (1.0)	257 (1.7)	8 (0.7)	254 (2.9)	6 (0.5)	251 (4.3)
Pennsylvania	39 (1.8)	282 (1.6)	19 (0.9)	274 (1.9)	30 (1.2)	262 (1.6)	7 (0.8)	252 (2.8)	5 (0.5)	252 (3.8)
Rhode Island	43 (1.1)	276 (1.1)	18 (1.5)	271 (1.5)	22 (1.4)	256 (1.6)	8 (0.4)	244 (2.1)	8 (0.6)	239 (2.5)
South Carolina	37 (1.4)	272 (1.5)	16 (0.7)	268 (1.7)	31 (0.9)	248 (1.4)	9 (0.6)	248 (2.1)	7 (0.3)	247 (3.0)
Tennessee	33 (1.5)	267 (2.1)	21 (0.9)	265 (1.8)	29 (1.0)	251 (1.6)	12 (0.8)	245 (2.0)	5 (0.4)	243 (3.6)
Texas	34 (1.6)	281 (2.1) >	18 (0.8) >	272 (1.6)	21 (1.0)	253 (1.6)	16 (1.0)	247 (1.7)	11 (0.8)	244 (2.4)
Utah	53 (1.3)	280 (1.0)	22 (1.0)	278 (1.2)	15 (0.8)	258 (1.8)	3 (0.3)	254 (3.2)	7 (0.5)	258 (2.7)
Virginia	41 (1.5)	282 (1.5)	18 (0.8)	270 (1.6)	24 (0.9)	252 (1.5)	9 (0.6)	248 (2.1)	8 (0.6)	251 (2.5)
West Virginia	29 (1.1)	270 (1.5)	18 (0.8)	269 (1.4)	33 (1.1) <	251 (1.2)	13 (0.9)	244 (1.8)	7 (0.4)	239 (2.3)
Wisconsin	38 (2.4)	287 (1.8)	24 (0.8)	282 (1.5)	28 (1.8)	270 (1.9)	5 (0.6)	254 (3.4)	6 (0.6)	255 (4.0)
Wyoming	42 (0.9)	281 (0.9)	22 (0.8)	278 (1.7)	23 (0.7)	266 (1.1)	5 (0.6)	258 (3.3)	7 (0.5)	260 (2.2) »
TERRITORIES										
Guam	28 (1.2)	246 (1.9)	13 (0.7)	244 (2.4)	27 (1.1)	229 (1.9)	10 (0.9)	224 (2.5)	22 (1.2)	226 (2.0)
Virgin Islands	23 (1.1)	224 (2.0)	11 (0.8)	232 (2.4)	29 (0.9)	221 (1.9)	14 (0.9)	219 (2.4)	24 (1.0)	217 (1.4)

The percentages for parents' highest level of education may not add to 100 percent because some students responded "I don't know." »The value for 1992 was significantly higher than the value for 1990 at about the 95 percent certainty level. «The value for 1992 was significantly lower than the value for 1990 at about the 95 percent certainty level. These notations indicate statistical significance from a multiple comparison procedure based on the 37 jurisdictions participating in both 1992 and 1990. If looking at only one state, then > and < also indicate differences that are significant. Statistically significant differences between 1990 and 1992 for the state comparison samples for the nation and regions are not indicated.

Answers to all of these questions lie within the bounds of Table 5, but how easily can they be extracted? Can we ease the pain of this extraction through a change in the design of the table?

Let us begin the real work of the redesign by asking why one would want to include the percentages in each educational category in the same table as the mathematics proficiency, as opposed to placing them in their own table on a facing page. The major reason is that the percentages are important for calculating state means. Such means are given in other tables, but it would seem good practice (remember Rule III) to include them here. Once they are calculated, they provide a sensible variable on which to order the states (rather than the alphabet—Rule II). Once this ordering has been accomplished we can see apparent gaps in the states' performance. A natural visual metaphor for these data gaps is to include matching physical gaps.[5] The resulting table of mean proficiencies by state is shown as Table 6. The percentage distributions of parental education elided from this table have been set aside for a parallel table; we shall return to these data shortly.

We have also moved the District of Columbia into the section of nonstates that also includes Guam and the Virgin Islands. All ordering is done within table section. Note that the key summaries are in boldface type. For ease of manipulation the standard errors have temporarily been removed from their parentheses. They will shortly be removed altogether.

Table 6 allows us to answer some of the questions phrased initially quite easily, especially those dealing with the relative performance of the states (Question 1). The usual finding of Midwestern states having the highest average performance and the Southern states the lowest is seen immediately. Moreover, we see that there is a 37-point difference between the highest states and the lowest. Interpreting 37 points is helped by remembering that there is an average increase of 12 NAEP points/year between fourth and eighth grade in math. Thus, the 37-point difference can be interpreted as corresponding to about a three-year difference in average performance between the best and worst performing states. This increases to more than four years when one's gaze shifts to the three "other jurisdictions." The gaps depicted help keep our eyes from blurring while examining such a large table, and they also provide rough groupings that may be suggestive of explanatory hypotheses. Note that the data about the various sections of the country on the top of Table 5 have been removed entirely. This was done because they were the products of a different survey and have larger standard errors than the individual states from which those sections are composed. Their inclusion just added an unnecessary source of confusion.

Examining the average proficiency for the nation at each education level reveals the unsurprising result that children whose parents are better educated score higher in mathematics. In addition, it appears that children who don't know their parents' education perform slightly better than children whose parents did not finish high school. This is suggestive of a grouping somewhat

TABLE 6

Reformatted version of Table 5 in which standard errors are in separately labeled columns, categories of parental education are separated, average state performance is shown, and rows are ordered and spaced by average performance

PUBLIC SCHOOLS	Graduated College		Some Education After High School		Graduated High School		Did Not Finish High School		I don't Know		Average	
	θ	se	θ	se	θ	se	θ	se	θ	se	θ	se
Nation	**279**	**1.4**	**270**	**1.2**	**256**	**1.4**	**248**	**1.8**	**251**	**1.7**	**267**	**1.4**
States												
Iowa	291	1.2	285	1.5	273	1.3	262	2.4	266	2.8	**283**	**1.4**
North Dakota	289	1.1	283	1.9	271	1.7	259	4.5	272	2.8	**283**	**1.5**
Minnesota	290	1.0	284	1.8	270	1.8	256	4.2	268	3.0	**282**	**1.6**
Maine	288	1.4	281	1.5	267	1.1	259	2.7	266	2.6	**278**	**1.5**
Wisconsin	287	1.4	280	1.5	267	0.9	259	2.5	262	2.1	**278**	**1.4**
New Hampshire	287	1.8	282	1.5	270	1.9	254	3.4	255	4.0	**278**	**2.0**
Nebraska	287	1.2	280	1.6	267	1.7	247	3.3	256	3.8	**277**	**1.6**
Idaho	281	0.9	278	1.3	268	1.4	254	2.3	254	2.8	**274**	**1.3**
Wyoming	281	0.9	278	1.7	266	1.1	258	3.3	260	2.2	**274**	**1.3**
Utah	280	1.0	278	1.2	258	1.8	254	3.2	258	2.7	**274**	**1.3**
Connecticut	288	1.0	272	1.8	260	1.8	245	3.3	251	2.4	**273**	**1.6**
Colorado	282	1.3	276	1.6	260	1.5	250	2.4	252	2.6	**272**	**1.6**
Massachusetts	284	1.3	272	1.8	261	1.4	248	3.2	248	2.6	**272**	**1.6**
New Jersey	283	1.8	275	2.1	259	2.5	253	3.8	250	3.9	**271**	**2.3**
Pennsylvania	282	1.6	274	1.9	262	1.6	252	2.8	252	3.8	**271**	**1.9**
Missouri	280	1.7	275	1.5	264	1.6	254	2.4	252	2.9	**271**	**1.8**
Indiana	283	1.5	275	1.9	260	1.6	250	2.6	249	3.3	**269**	**1.8**
Ohio	279	1.8	272	1.6	260	2.3	243	2.6	249	4.5	**268**	**2.1**
Oklahoma	277	1.5	272	1.9	257	1.7	254	2.9	251	4.3	**267**	**1.9**
Virginia	282	1.5	270	1.6	252	1.5	248	2.1	251	2.5	**267**	**1.7**
Michigan	277	2.2	271	2.0	257	1.7	249	2.0	248	3.0	**267**	**2.1**
New York	277	1.9	271	2.4	256	2.5	243	4.2	240	3.8	**265**	**2.5**
Rhode Island	276	1.1	271	1.5	256	1.6	244	2.1	239	2.5	**265**	**1.5**
Arizona	277	1.5	270	1.5	256	1.6	245	2.5	248	2.7	**264**	**1.8**
Maryland	278	1.8	266	1.9	250	1.8	240	3.7	245	3.8	**264**	**2.1**
Texas	281	2.1	272	1.6	253	1.6	247	1.0	244	2.4	**264**	**1.8**
Delaware	274	1.3	268	2.3	251	1.7	248	4.0	248	3.4	**262**	**1.9**
Kentucky	278	1.6	267	1.6	254	1.6	246	1.7	242	2.8	**261**	**1.7**
California	275	2.0	266	2.1	251	2.1	241	2.2	240	2.9	**260**	**2.2**
South Carolina	272	1.5	268	1.7	248	1.4	248	2.1	247	3.0	**260**	**1.7**
Florida	268	1.9	266	1.9	251	1.8	244	2.7	244	3.2	**259**	**2.1**
Georgia	271	2.1	264	1.7	250	1.3	244	2.2	245	2.6	**259**	**1.8**
New Mexico	272	1.4	264	1.4	249	1.4	244	1.9	245	2.0	**259**	**1.5**
Tennessee	267	2.1	265	1.8	251	1.6	245	2.0	243	3.6	**258**	**2.0**
West Virginia	270	1.5	269	1.4	251	1.2	244	1.8	239	2.3	**258**	**1.5**
North Carolina	271	1.4	265	1.6	246	1.7	240	2.3	240	3.6	**258**	**1.7**
Hawaii	267	1.5	266	1.9	246	1.8	242	3.5	246	2.1	**257**	**1.9**
Arkansas	264	1.9	264	1.7	248	1.6	246	2.4	245	2.7	**256**	**1.9**
Alabama	261	2.5	258	2.0	244	1.8	239	2.0	237	2.9	**251**	**2.2**
Louisiania	256	2.5	259	1.8	242	1.6	237	2.4	236	3.7	**249**	**2.2**
Mississippi	254	1.6	256	2.0	239	1.6	234	1.8	231	2.8	**246**	**1.8**
Other Jurisdictions												
Guam	246	1.9	244	2.4	229	1.9	224	2.5	226	2.0	**235**	**2.0**
District of Columbia	244	1.7	240	1.9	224	1.6	225	3.2	229	2.2	**234**	**1.9**
Virgin Islands	224	2.0	232	2.4	221	1.9	219	2.4	217	1.4	**222**	**1.9**

heterogeneous in parental education. A small plot of mean math performance against parents' education (Figure 1), with a rough reference line drawn in, makes the quantitative aspect of this relationship clearer and provides a reasonable answer to Question 2.

Scanning down the first column of the table shows that the higher-scoring states also tend to have a greater proportion of children coming from homes with a parent who was a college graduate. But even among just these children (conditioning on parents' education), there is still a 37-point difference between the highest- and lowest-scoring states. This is part of an answer to the third kind of question, although more complete answers can be built by constructing graphs like Figure 1 for individual states. Such a graph, shown as Figure 2, contradicts the hypothesis that differences in states' overall performance are due to differences in parents' education. Aside from being mildly startling in its own right, this result reduces still further the need to include the percentage of children in each parental education category within this table.

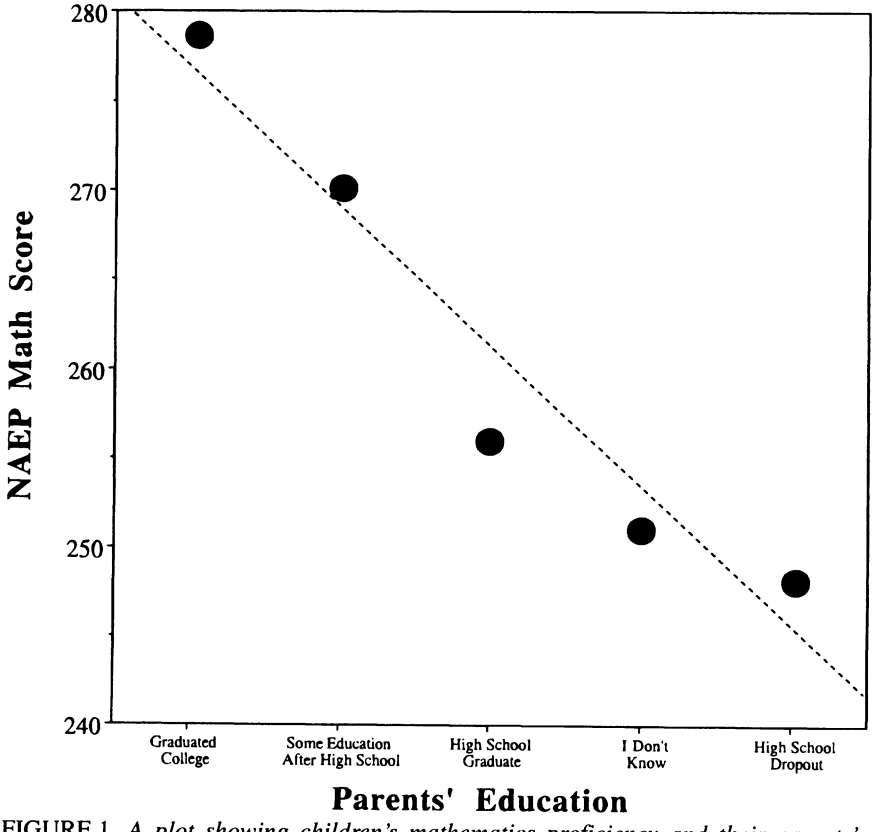

FIGURE 1. *A plot showing children's mathematics proficiency and their parents' education. A reference line drawn in shows roughly the relationship between the (mostly) ordered categories of parental education and children's performance.*

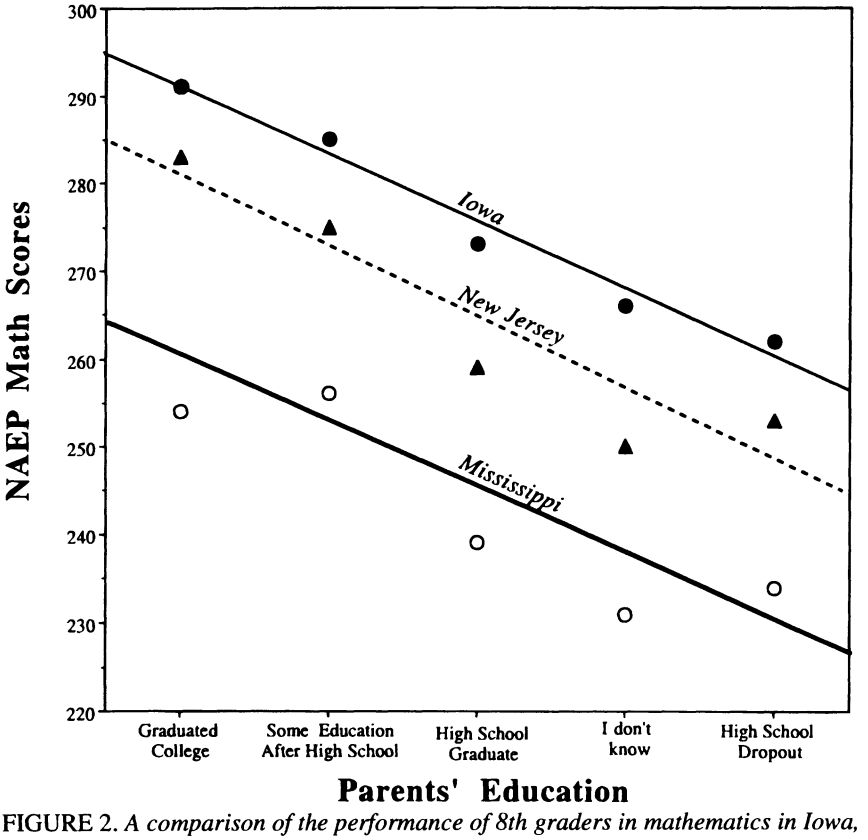

FIGURE 2. *A comparison of the performance of 8th graders in mathematics in Iowa, New Jersey, and Mississippi, shown as a function of their parents' education. The difference in the performance of children in mathematics by state is not solely due to differences in parental education.*

What About the Standard Errors?

Questions about the statistical significance of these observed differences can be answered after doing a little arithmetic on the standard errors included within the table. A natural question to ask is why that arithmetic hasn't already been done by the generators of the table. One possible answer to this question is that there are too many plausible questions of statistical significance that might be asked to calculate all of the possible error terms. But, playing devil's advocate, couldn't some conservative error term be calculated that would save all of the clutter introduced by the many columns of standard errors? The answer to this, simply put, is yes. And the next version of this table (shown as Table 7) segregates the standard errors into a separate table and substitutes instead (for quick and dirty significance judgments) three estimates of the standard error of the difference between any two entries in that column. The first is an upper bound on the standard error of the difference.

This is obtained by multiplying the largest value of the standard error in that column by $\sqrt{2}$. The second entry, labeled *40 Bonferroni*, is the first entry multiplied by 3.2. This is obtained from the Bonferroni inequality and based on the idea that a user is interested in making comparisons of his/her own state with each of the others. This controls the family of tests protection beyond the .05 level. The last entry, labeled *820 Bonferroni*, is the first entry multiplied by 4.0 and controls the family of tests significance for someone who compares each state with all others. It is likely that this last estimate is unnecessary, since anyone expecting to make that many comparisons will almost surely want the tighter error bounds constructed from the individual standard errors and perhaps use more powerful procedures for multiple comparisons (e.g., Benjamini & Hochberg, 1995).[6]

A table augmented with these error summaries, but relieved of the burden of individual accompanying standard errors, is not only a good deal clearer to look at but, for most prospective users, a good deal easier to use for making inferences about statistical significance of observed differences.

Next, while there was no good reason to combine mathematics achievement and percentage of children in each category into the same table, these percentage distributions are important in their own right. It was just that their presentation was clearer after they were separated into two tables. To examine this, consider the two variables, shown as Tables 7 and 8. Table 7 contains just mean mathematics proficiency; Table 8 just the distribution of children across levels of parental education. It appears that the benefits associated with housing both of these variables within the same table are too few to offset the increases in perceptual complexity that accrue by mixing them. It seems, however, worthwhile to keep them contiguous. Thus we would recommend placing them on facing pages. Note that the states in Table 8 are ordered by the state means from Table 7. This facilitates comparisons between the two tables. It also raises the interesting question of whether the increased ease of comprehension yielded by ordering a table by its contents is more than offset by the increased difficulty in making comparisons across tables ordered in different ways. This issue will be discussed further at the end of this section.

On both tables we have highlighted unusual entries by putting them in **boldface type** and boxing them in. Entries that are unusually large are also shaded. Entries that are unusually small are boxed but unshaded. We have also appended a positive or negative sign as a further reminder of the direction of the entry's variation. Thus in Table 7 we see that the average score of children whose parents had only some post–high school education was unusually high in West Virginia. Similarly, Nebraska's and Connecticut's children of high school dropouts scored unusually poorly.

The determination of which entries were unusual was made by fitting a simple additive model to the data and examining the residuals. Those residuals that stuck out excessively (more than 2 times the square root of the mean of the squared residuals) were then highlighted. Table 7 goes about as far as

TABLE 7

Revision of Table 6 with individual standard errors replaced by conservative estimates, unusual entries highlighted, and a state locator index inserted

	PUBLIC SCHOOLS	Graduated College	Some Education After High School	Graduated High School	Did Not Finish High School	I Don't Know	Mean
	Nation	**279**	**270**	**256**	**248**	**251**	**267**
	States						
1	Iowa	291	285	273	262	266	**283**
2	North Dakota	289	283	271	259	272	**283**
3	Minnesota	290	284	270	256	268	**282**
4	Maine	288	281	267	259	266	**278**
5	Wisconsin	287	282	270	254	255	**278**
6	New Hampshire	287	280	267	259	262	**278**
7	Nebraska	287	280	267	**247 -**	256	**277**
8	Idaho	281	278	268	254	254	**274**
9	Wyoming	281	278	266	258	260	**274**
10	Utah	280	278	258	254	258	**274**
11	Connecticut	288	272	260	**245 -**	251	**273**
12	Colorado	282	276	260	250	252	**272**
13	Massachusetts	284	272	261	248	**248 -**	**272**
14	New Jersey	283	275	259	253	250	**271**
15	Pennsylvania	282	274	262	252	252	**271**
16	Missouri	280	275	264	254	252	**271**
17	Indiana	283	275	260	250	249	**269**
18	Ohio	279	272	260	243	249	**268**
19	Oklahoma	277	272	257	254	251	**267**
20	Virginia	282	270	252	248	251	**267**
21	Michigan	277	271	257	249	248	**267**
22	New York	277	271	256	243	**240 -**	**265**
23	Rhode Island	276	271	256	244	**239 -**	**265**
24	Arizona	277	270	256	245	248	**264**
25	Maryland	278	266	250	240	245	**264**
26	Texas	281	272	253	247	244	**264**
27	Delaware	274	268	251	248	248	**262**
28	Kentucky	278	267	254	246	242	**261**
29	California	275	266	251	241	240	**260**
30	South Carolina	272	268	248	248	247	**260**
31	Florida	268	266	251	244	244	**259**
32	Georgia	271	264	250	244	245	**259**
33	New Mexico	272	264	249	244	245	**259**
34	Tennessee	267	265	251	245	243	**258**
35	West Virginia	270	**269 +**	251	244	239	**258**
36	North Carolina	271	265	246	240	240	**258**
37	Hawaii	267	266	246	242	246	**257**
38	Arkansas	264	264	248	**246 +**	245	**256**
39	Alabama	261	258	244	239	237	**251**
40	Louisiania	256	259	242	237	236	**249**
41	Mississippi	254	256	239	234	231	**246**
	Other Jurisdictions						
42	Guam	246	244	229	224	226	**235**
43	District of Columbia	244	240	224	225	229	**234**
44	Virgin Islands	224	232	221	219	217	**222**
				Error terms for comparisons			
	Max Std error of diff	3.5	3.4	3.5	6.4	6.4	**3.5**
	40 Bonferroni	11.3	11.0	11.3	20.7	20.7	**11.3**
	820 Bonferroni	14.0	13.6	14.0	25.6	25.6	**14.0**

TABLE 8

A parallel of Table 7 including instead percentage distribution of parental education

	PUBLIC SCHOOLS	Graduated College	Some Education After High School	Graduated High School	Did Not Finish High School	I Don't Know
	Nation	**40**	**18**	**25**	**8**	**9**
	States					
1	Iowa	44	21	25	4	5
2	North Dakota	54 +	18	19	3	5
3	Minnesota	48	21	22	3	7
4	Maine	40	22	26	6	5
5	Wisconsin	38	24	28	5	6
6	New Hampshire	46	17	24	6	7
7	Nebraska	46	20	24	4	6
8	Idaho	48	20	19	7	6
9	Wyoming	42	22	23	5	7
10	Utah	53 +	22	15 -	3	7
11	Connecticut	47	16	22	6	9
12	Colorado	46	19	21	6	7
13	Massachusetts	48	17	21	7	7
14	New Jersey	45	18	23	7	8
15	Pennsylvania	39	19	30	7	5
16	Missouri	36	22	29	8	6
17	Indiana	33	21	32	8	6
18	Ohio	37	19	32	7	5
19	Oklahoma	39	21	26	8	6
20	Virginia	41	18	24	9	8
21	Michigan	38	23	26	6	7
22	New York	44	18	23	6	10
23	Rhode Island	43	18	22	8	8
24	Arizona	36	22	21	10	12
25	Maryland	44	18	25	6	7
26	Texas	34	18	21	16	11
27	Delaware	39	18	30	6	8
28	Kentucky	28 -	19	32	15	6
29	California	39	18	17 -	10	16
30	South Carolina	37	16	31	9	7
31	Florida	39	19	24	8	10
32	Georgia	35	18	30	11	6
33	New Mexico	34	20	26	11	10
34	Tennessee	33	21	29	12	5
35	West Virginia	29 -	18	33	13	7
36	North Carolina	36	20	27	10	6
37	Hawaii	38	15	25	6	16
38	Arkansas	30 -	20	31	11	8
39	Alabama	33	18	29	13	7
40	Louisiania	32 -	20	30	10	7
41	Mississippi	36	16	29	13	7
	Other Jurisdictions					
42	Guam	28	13	27	10	22
43	District of Columbia	32	17	29	9	12
44	Virgin Islands	23	11	29	14	24
			Error terms for comparisons			
	Max Std error of diff	3.4	2.1	2.5	1.4	1.7
	40 Bonferroni	11.0	6.8	8.1	4.5	5.5
	820 Bonferroni	13.6	8.4	10.0	5.6	6.8

we might expect in displaying the results to answer all of the questions about achievement scores phrased earlier.

Last, the individual standard errors that were previously housed in the original table have been combined and piled into two tables of standard errors matching Tables 7 and 8. These are available from the author at hwainer@ets.org. I believe that these will be so rarely consulted that it isn't worth using up extra pages here. Future experience will inform this judgment, and I am prepared to change the format if I am wrong.

Thus we have found that separating variables that are only tangentially related into separate tables yields increased comprehensibility. Once the separation is completed, the tables should be structured according to the four rules specified earlier. The questions posed at the beginning of this section, which characterize the most plausible reasons why anyone would want to see these data, are all answered more easily from these revised tables.

What about order? Clearly, if we wish to compare data values on different variables from the same set of states it is often helpful if those data are ordered in the same way in those different tables. This is currently accomplished by ordering all tables alphabetically. Is this a good idea? I think that there are several alternatives. The most attractive one to me is to order each table as an independent entity, to be looked at and understood on its own. Secondary analyses that require combining information from several tables should be done from a different data source than the table; almost surely it should be some electronic database that would allow easy subsequent manipulations. But if we are to think of the tables as the first available archive, there may be an argument for ordering all tables on a similar topic in the same way, so that various pieces of information about a particular state can be picked out easily. If so, alphabetical ordering is only one possibility among many. Is it the best one? Alphabetical ordering has only one thing going for it: It makes locating a specific state easier.[7] Its principal drawback is that it usually obscures the structure that the table was constructed to inform us about. If a set of tables like those that grew out of Table 5 are constructed and ordered by overall performance (instead of alphabetically), we have made finding a particular state a bit more difficult.[8] I believe that this is a small cost in comparison to the gain in comprehensibility. But even this can be ameliorated through the inclusion of a "locator table." All we need to do is number the jurisdictions in the table sequentially from 1 to 44, as was done in the first column of Tables 7 and 8, and then have a small, alphabetically ordered locator index table (Table 9) that connects alphabetically ordered state names to row numbers in the empirically ordered tables.

Compound Tables

Table 7 is a rectangular array showing a single dependent variable, mean mathematics proficiency, as a function of two independent variables, parents'

TABLE 9
Alphabetically ordered locator index table of the states in Tables 7 and 8, to be used in case of an emergency loss of any particular jurisdiction

State	Position	State	Position
Alabama	39	New York	22
Arizona	24	New Jersey	14
Arkansas	38	New Mexico	33
California	29	New Hampshire	5
Colorado	12	North Dakota	2
Connecticut	11	North Carolina	36
Delaware	27	Ohio	18
Florida	31	Oklahoma	19
Georgia	32	Pennsylvania	15
Hawaii	37	Rhode Island	23
Idaho	8	South Carolina	30
Indiana	17	Tennessee	34
Iowa	1	Texas	26
Kentucky	28	Utah	10
Louisiana	40	Virginia	20
Maine	4	West Virginia	35
Maryland	25	Wisconsin	6
Massachusetts	13	Wyoming	9
Michigan	21		
Minnesota	3	**Other Jurisdictions**	
Mississippi	41	District of Columbia	43
Missouri	16	Guam	42
Nebraska	7	Virgin Islands	44

education and geographic location. As we have seen, when properly designed such a table can be a clear and evocative communicator of information. Unfortunately, clear design is too commonly abandoned in favor of compound tables when multivariate or multilevel data are to be displayed. Such tables are very hard to understand. In fact, in a survey of education policymakers, Hambleton and Slater (1995) found that such compound tables were the most frequently misunderstood of any data display in NAEP executive summaries. In one such display, parallel in structure to Table 10, more than half of the education professionals answered a simple data extraction question incorrectly.

To illustrate the negative effects of a compound table, consider Table 10, a somewhat tidied up version of Table 2.3 from the 1992 NAEP Reading

TABLE 10

Reproduction (with blank lines removed) of Table 2.3 from NAEP Reading Report Card for the Nation and the States *(p. 89): Average reading proficiency and achievement levels by region*

	Percentage of Students	Average Proficiency	Percentage of Students At or Above			Below Basic
			Advanced	Proficient	Basic	
Grade 4						
Northeast	21(1.1)	223(3.7)	7(2.2)	31(4.1)	63(3.5)	37(3.5)
Southeast	23(1.0)	214(2.4)	4(0.7)	21(2.5)	54(3.2)	46(3.2)
Central	27(0.5)	221(1.4)	4(0.9)	26(2.1)	63(2.0)	37(2.0)
West	28(0.8)	215(1.5)	4(0.6)	24(1.4)	56(1.8)	44(1.8)
Grade 8						
Northeast	22(0.7)	263(1.8)	3(0.4)	31(1.9)	71(2.3)	29(2.3)
Southeast	25(0.5)	254(1.7)	1(0.4)	22(2.3)	63(1.8)	37(1.8)
Central	25(0.5)	264(2.2)	2(0.6)	31(2.4)	73(2.4)	27(2.4)
West	28(0.6)	260(1.2)	2(0.5)	27(1.4)	68(1.5)	32(1.5)
Grade 12						
Northeast	24(0.6)	293(1.3)	4(0.5)	40(1.6)	76(1.6)	24(1.6)
Southeast	23(0.6)	284(1.1)	2(0.3)	28(1.4)	68(1.4)	32(1.4)
Central	26(0.6)	294(1.1)	3(0.4)	40(1.6)	79(1.4)	21(1.4)
West	27(0.8)	292(1.6)	4(0.6)	38(2.2)	77(2.0)	23(2.0)

The standard errors of the estimated percentages and proficiencies appear in parentheses. It can be said with 95 percent certainty that for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference (see Appendix for details). Percentages may not total 100 percent due to rounding error.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 Reading Assessment

Assessment. This display can be improved by breaking it up into smaller displays. For example, the average proficiencies are best shown as a two-way table by themselves (see Table 11). This table shows quantitatively the modest size of the region effects and the much larger grade effects. There are no large interactions, and so no entries are boxed in. Of course, an even more evocative image could be obtained by subtracting out the grade means and then plotting the residuals. Such a table would make clear just how different the Southeast is. Similar tables containing the percentage of children at each of the NAEP reading levels could also be constructed.

Why is it that it is often easier to understand several simple displays than one compound one? To understand a display involves two distinct phases of perception (Bertin, 1973/1983) which are characterized by two questions: What are the components of data that are being reported? What are the relations among them?

The first phase is easy if the horizontal and vertical components are unitary, for example, grade level versus region. It becomes more difficult if they are not, for example, level of proficiency and average proficiency and percent versus region and grade. The second phase of perception is addressed by the four rules, but it is made more difficult if the first phase is complex. In this report we have suggested that unless the simultaneous presentation of multidimensional information is critical to understanding, comprehension is aided by keeping the components of data simple and presenting tables paired. This was illustrated earlier when we separated the percentage distributions and the achievement scores into separate tables.

Discussion and Conclusions

Just as there are no good or bad tests, neither are there good or poor graphs nor good or poor tables. Their value depends on the uses to which they are put. Some constructions answer the questions one is entitled to ask, and

TABLE 11

Reformatted version of Table 10 in which only regional proficiencies by grade level are included. It is arranged to emphasize the two-way structure of the data. It is bordered by means and main effects

Average Proficiency

Region	Grade 4	Grade 8	Grade 12	**Regional Means**	**Regional Effects**
Northeast	223	263	293	**260**	**4**
Central	221	264	294	**260**	**4**
West	215	260	292	**256**	**0**
Southeast	214	254	284	**251**	**-5**
Grade Means	**216**	**260**	**291**	**256**	
Grade Effects	**-40**	**4**	**35**		

others do not. By making the hierarchy of possible questions explicit, we emphasize the fact that one cannot look at a graph or table as one looks at a painting or a traffic signal. One does not passively read a graph; one queries it. And one must know how to ask useful questions.

What are the questions that can be asked? To some extent we explored this in the second section. In general they are the same questions that would be asked of data from any factorial experiment: What are the row effects? What are the column effects? What are the interactions? How do the rows and columns group as functions of these effects?

The goal of effective display is to ease the viewer's task in answering these questions. We have found that wisely ordering, rounding, summarizing, and spacing go a long way toward accomplishing this. In addition, we must confront the likely use of a table head-on before including various mixtures of variables into it. Adding extra stuff always affects comprehensibility, and we must make the triage decision between saving space by combining two or more tables into one and communicating clearly. It has been our experience that breaking up complex displays sensibly often communicates more efficiently, em for em, than a large compound table.

Measuring Numeracy

Earlier we showed that if tables that are used as stimuli within a test item are prepared properly, the questions associated with them are usually reduced in difficulty, often dramatically. This does not mean that the practice of asking such questions ought to be discontinued, any more than we advocate continuing to use poorly constructed tables to make such questions less trivial. The test's usefulness as a learning instrument would be enhanced if it served as a model for how tables ought to be prepared as well as illustrating the depth of information readily available from well prepared tables.

Well prepared tables will also allow us to construct questions that probe the deep structure of the data in a way that is too difficult with poorly prepared tables. What are such questions like? To answer this we need a little theory. And, to illustrate this theory, we will use the battery life item from the 1990 Science Assessment introduced earlier and reproduced here as Table 12. This is identical to Table 4, shown earlier, except that four unusual entries have been indicated by boxing them in. A shaded box with a positive sign indicates a higher than expected entry; an unshaded box with a negative sign means a lower than expected entry.

Ehrenberg (1977) calls the ability to understand data presented in a table "numeracy." This term may have broader application, but we shall use it in this narrow context for the nonce.

How can we measure someone's proficiency in understanding quantitative phenomena that are presented in a tabular way (an individual's numeracy)? Obviously there are NAEP test items written that purport to do exactly this; the items described earlier are some typical examples. We can do better

TABLE 12
Revision of Table 4 with unusual entries highlighted

	Battery Life in Hours				
Battery Brands	Radio	Flashlight	Cassette Player	Portable Computer	**Battery Averages**
Never Die	**28 +**	16	8	6	**15**
Electro-Blaster	26	15	10	4 -	**14**
PowerBat	24	13	7	5	**12**
Servo-Cell	21	12	4	2	**10**
Constant Charge	19 -	10	5	3 +	**9**
Usage averages	**24**	**13**	**7**	**4**	**12**

with the guidance of a formal theory of graphic communication (Wainer, 1980, 1992).

Rudiments of a Theory of Numeracy

Fundamental to the measurement of numeracy is the broader issue of what kinds of questions tables can be used to answer. My revisions of Bertin's (1973/1983) three levels of questions are:

- Elementary-level questions involve data extraction, for example, How long does a Servo-Cell last in a cassette player?
- Intermediate-level questions involve trends seen in parts of the data, for example, How much longer is a battery likely to last in a radio than in a portable computer?
- Overall-level questions involve the deep structure of the data being presented in their totality, usually comparing trends and seeing groupings, for example, Which two appliances show the same pattern of battery usage?, or, Which brands of batteries show the same pattern of battery life?

They are often used in combination; for example, Zabell (1976) referred to their use in the detection of outliers—unusual data points. To accomplish this objective we need a sense of what is usual (i.e., a trend at the intermediate level), and then we look for points that do not conform to this trend (the elementary level). Such questions are hard to answer from a raw table such as Table 1 but are trivial in Table 11, where such interactions (this time from an additive model) are highlighted.

Note that although these levels of questions involve an increasingly broad understanding of the data, they do not necessarily imply an increase in the empirical difficulty of the questions.[9]

Reading a table at the intermediate level is clearly different from reading a table at the elementary level; a concept of trend requires the notion of

connectivity. If the columns were not four appliances but instead four decreasing levels of parental education (as in Table 5), the idea of an increasing trend would be more meaningful. Comparing trends among different states likewise requires an additional notion of connectivity, but this time across the dependent variable (NAEP math scores). This connectedness is characterized by a common variable and emphasizes the inferential costs of mixing together different dependent variables in the same display.

I hope that this brief introduction conveys a sense of how this formal structure can make it easier to construct tests of numeracy, and to understand better which characteristic of numeracy we are measuring. Of course, to ask questions at higher levels requires data of sufficient richness to support them, as well as tables clear enough for the quantitative phenomena to show through. It is much more difficult to answer intermediate- or overall-level questions from Table 1 than from Table 12. It is also easier to see trends, and deviations from them, with a different display format altogether (see Figure 3). Once again we see that the format we choose must be based upon our purpose in

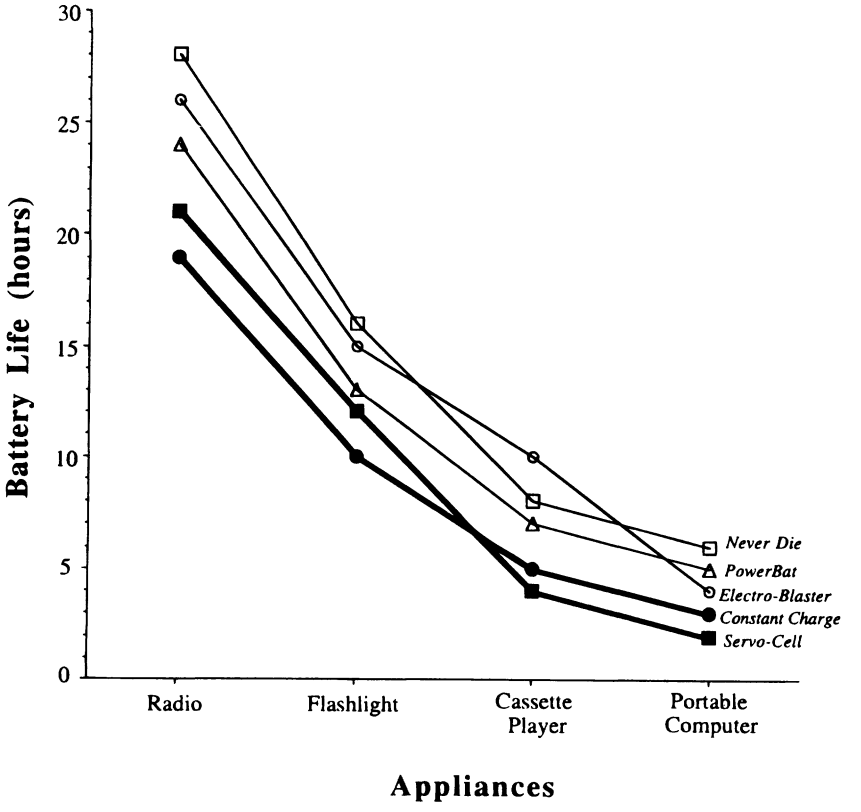

FIGURE 3. *A graph that emphasizes the large differences in battery life span among possible usages compared to the somewhat smaller differences among batteries.*

constructing the display. While elementary-level questions are best answered with a table, intermediate- and overall-level questions may be easier with a graph. However, as we have demonstrated, well prepared tables can be useful at higher levels.

My experience is that test items associated with tables tend to be questions of the first kind, although often they are compounded through the use of nontabular complexity. This is not an isolated practice confined to the measurement of numeracy. In the testing of verbal reasoning it is common practice to make a reasoning question more difficult simply by using more arcane vocabulary. This practice stems from the fact that it is almost impossible to write questions that are more difficult than the questioner is smart. When we try to test the upper reaches of reasoning ability, we must find item writers who are more clever still.

Of course, when we record a certain level of performance by an examinee on a table-based item, we can only infer a lower bound on someone's numeracy;[10] a better table of the same data ought to make the item easier. Similarly, a more numerate audience makes a table appear more efficacious.

Software

There is an enormous wealth of software available to make tables. I have found that the versatility of spreadsheet programs is especially useful. All tables in this article were prepared using Microsoft's EXCEL™ on a Macintosh computer. Such software allows pretty complete control of fonts, type sizes, borders, and shading. Ordering rows is trivial; ordering columns requires a little work. Transforming data from a spreadsheet table into the graph of your choice is easily accomplished within EXCEL™ for most common graphic forms. For more esoteric formats the data are easily moved into special-purpose programs.

The real power of spreadsheets emerges when calculations of some complexity are required. This lifts the spreadsheet from being merely handy to being essential for preparing tables. The identification of unusual data points in a two-way table requires calculating row and column effects and then subtracting them out. Determining which gaps in a univariate data string are likely to be worth emphasizing requires ordering the data, calculating the gaps as well as a vector of inverse logistic weights, and combining and summarizing them. All of these tasks can be done on the fly within a spreadsheet. Specially designed table software does not always measure up in this regard.

Before one can use this or any software on NAEP data, one must first extract those data from rather complex NAEP data files and transform them into a format acceptable to a spreadsheet. This formerly onerous task has been eased considerably through the development of some special-purpose software (NAEPEX) that is distributed with the NAEP Secondary-Use Data products. NAEPEX allows the user to define, extract, and analyze subsets

of NAEP data in a relatively painless manner. Further details about NAEPEX are contained in its user guide (Rogers, 1995).

Summing Up

Tables are used for many purposes within NAEP: as stimuli in test items, as containers to archive data, and as a communicative medium. Believing that the archival purpose is anachronistic, we focused our attention on rules for building tables to facilitate their efficacy as communicative devices. We found that the same four rules apply to the simplest tables used as stimuli within the assessment and to the most complex tables aimed at scientists. While the rules are objective and as such can be applied through a completely automatic procedure, human judgment and wisdom are still required. Before applying the rules, one must decide on the most likely prospective uses for the data in the table and include only those data that facilitate those uses.

Of foremost importance is the notion that we are typically not looking at a table to simply extract a number. To become involved in a problem and to understand it is to shift from extracting individual entries to understanding quantitative phenomena. The construction of efficacious data displays aims to promote this transition, allowing the reader a graceful change from spectator to participant.

Notes

This work was sponsored by the National Center for Education Statistics through Contact Number R999B40013 to the Educational Testing Service, Howard Wainer, Principal Investigator. Although I am pleased to express my gratitude for this support, I must reexpress the usual caveat that all opinions expressed here are those of the author and do not necessarily reflect the views of either NCES or the U.S. Government. I am delighted to be able to thank Jeremy Finn for his critical and constructive comments on this work as it developed. Of course, he shouldn't be held responsible for what has resulted from his good advice. I would also like to thank Brent Bridgeman, John Mazzeo, Keith Reid-Green, Linda Steinberg, and an unusually helpful associate editor and two sharp-eyed, thoughtful, but anonymous referees for their comments on an earlier draft. Last, my gratitude to John Tukey for his helpful suggestions on the choice of an error term for large tables. This article was abstracted from a considerably longer technical report (Wainer, 1995); readers interested in receiving a copy of that report can request it from Martha Thompson at Educational Testing Service (mthompson@ets.org).

[1]NAEP is a congressionally mandated survey of the educational achievement of American students and of changes in that achievement across time. This survey has been operational for nearly 25 years and utilizes technically sophisticated sampling and assessment methodology. The results of NAEP are made available to both the professional and lay public continuously and are cited with increasing frequency as evidence in public debates about educational topics.

NAEP's results are complex, consisting, as they do, of (a) outcomes on achievement tests of complex character on a variety of subjects; (b) attitude and behavioral information from the children, teachers, and others associated with the children's

schooling; and (c) detailed demographic information about the children who took the assessment instruments. These data are reported in a variety of ways that vary with the character of the data, their prospective audience, and the purposes of the data.

[2]One can easily construct pathological exceptions (e.g., a sample mean from a Cauchy distribution), but for most normal situations this is a pretty good general rule.

[3]I sometimes hear from colleagues that my ideas about rounding are too radical—that such extreme rounding would be "OK if we knew that a particular result was final. But our final results may be used by someone else as intermediate in further calculations. Too-early rounding would result in unnecessary propagation of error." Keep in mind that tables are for communication, not archiving. Round the numbers and, if you must, insert a footnote proclaiming that the unrounded details are available from the author. Then sit back and wait for the deluge of requests.

[4]Of course, teachers' grade books are usually alphabetical and so yield tables like the original. But I suspect many teachers (myself included) now use electronic grade books which are alphabetized for ease of data entry and have a second version for retrieval. This discussion is about retrieval.

[5]These gaps were determined to be largish through consideration of both their size and their location. A big gap in the tails is not as unlikely as one of similar size in the middle. In this instance we used inverse logistic weights on the gaps to adjust for location (Wainer & Schacht, 1978).

[6]Choosing the maximum may be too conservative for many users. Two alternatives may be considered. The first is shrinking the maximum inward based upon the stability of the estimates of the standard error. In this instance the standard errors are based on about 30 degrees of freedom. This would suggest some modest shrinkage. If the degrees of freedom were 3 or 300, quite different decisions would be reached. The second alternative is replacing MAX(se) with a more average figure—for example,

$$\sqrt{\sum_{k=1}^{n} se_k^2/n}.$$

This second alternative seems especially attractive in this instance, since the distribution of standard errors across states is not too far from the null distribution expected from a chi-square variable with 30 degrees of freedom. The issues surrounding the best choice of error term is a bit afield from our purpose, and so we shall be content to raise it and leave its resolution to other accounts.

[7]Although not that easy. I have discovered, to my chagrin, that the two-letter state abbreviations do not yield the same alphabetic ordering as the full state names.

[8]An especially difficult task is finding out that the state you are looking for did not participate in the assessment.

[9]Although one small empirical study among 3rd-, 4th-, and 5th-grade children (Wainer, 1980) showed that, on average, item difficulty increased with level and graphicacy increased with age.

[10]It is like trying to decide on Mozart's worth as a composer on the basis of a performance of his works by Spike Jones on the washboard.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289–300.

Bertin, J. (1983). *Semiology of graphics* (W. Berg & H. Wainer, Trans.). Madison: University of Wisconsin Press. (Original work published 1973)

Clark, N. (1987). Tables and graphics as a form of exposition. *Scholarly Publishing, 10*(1), 24–42.

Court Statistics Project. (1976). *State court caseload statistics: Annual report, 1976.* Williamsburg, VA: National Center for State Courts.

Ehrenberg, A. S. C. (1977). Rudiments of numeracy. *Journal of the Royal Statistical Society, Series A, 140*, 277–297.

Farquhar, A. B., & Farquhar, H. (1891). *Economic and industrial delusions: A discourse of the case for protection.* New York: Putnam.

Hambleton, R. K., & Slater, S. C. (1995). *Are NAEP executive summary reports understandable to policy-makers and educators?* (Research report). Amherst: University of Massachusetts.

Playfair, W. (1786). *The commercial and political atlas.* London: Corry.

Rogers, A. (1995). *NAEPEX: NAEP data extraction program user guide.* Princeton, NJ: Educational Testing Service.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wainer, H. (1980). A test of graphicacy in children. *Applied Psychological Measurement, 4*, 331–340.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*(1), 14–23.

Wainer, H. (1993). Tabular presentation. *Chance, 6*, 52–56.

Wainer, H. (1995). *A study of display methods for NAEP results: I. Tables* (Tech. Rep. No. 95-1). Princeton, NJ: Educational Testing Service.

Wainer, H., & Schacht, S. (1978). Gapping. *Psychometrika, 43*, 203–212.

Walker, H. M., & Durost, W. N. (1936). *Statistical tables: Their structure and use.* New York: Bureau of Publications, Teachers College, Columbia University.

Zabell, S. (1976). *Arbuthnot, Heberden and the* Bills of Mortality (Tech. Rep. No. 40). Chicago: The University of Chicago, Department of Statistics.

Author

HOWARD WAINER is Principal Research Scientist, Educational Testing Service, Princeton, NJ 08541; hwainer@ets.org. He specializes in statistical graphics and psychometrics.

Received March 6, 1995
Revision received August 28, 1995
Accepted October 26, 1995