DOCUMENT RESUME

ED 387 520 TM 023 814

AUTHOR Wainer, Howard
TITLE A Study of Display Methods for NAEP Results: I. Tables. Program Statistics Research. Technical Report No. 95-1.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-10
PUB DATE Mar 95
NOTE 51p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Access to Information; Data Analysis; Information Dissemination; *Information Utilization; National Surveys; *Tables (Data); Test Construction; *Test Results; Test Use
IDENTIFIERS *Data Display; *National Assessment of Educational Progress

ABSTRACT

The National Assessment of Educational Progress (NAEP) is an enormous and enormously ambitious project. It generates data of a richness and complexity beyond any simple survey. The broad utilization of the information it provides can be aided through the use of more evocative data displays. In this report, the uses to which data tables are put is examined, and ways are suggested in which the construction of tables can be modified to enable them to carry out their roles more efficaciously. A theoretical structure is also discussed to aid in the development of test items to tap students' proficiency in extracting information from tables. Five figures and 18 tables illustrate the discussion. (Contains 17 references.) (SLD)

RR-95-10

A Study of Display Methods for NAEP Results: I. Tables

Howard Wainer
Educational Testing Service

PROGRAM
STATISTICS
RESEARCH

Technical Report No. 95-1

Educational Testing Service
Princeton, New Jersey 08541

A Study of Display Methods for NAEP Results: I. Tables

Howard Wainer
Educational Testing Service

Program Statistics Research
Technical Report No. 95-1

Research Report No. 95-10

Educational Testing Service
Princeton, New Jersey 08541

March 1995

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

A study of display methods for NAEP results: I. Tables

Howard Wainer[1]
Educational Testing Service

Abstract

NAEP is an enormous and enormously ambitious project. It generates data of a richness and complexity beyond any simple survey. The broad utilization of the information it provides can be aided through the use of more evocative data displays. In this report we examine the uses to which data tables are put and suggest ways in which the construction of tables can be modified to enable them to carry out their role more efficaciously. We also discuss a theoretical structure to aid in the development of test items to tap students' proficiency in extracting information from tables.

1. Introduction

The most critical measure of any educational system is the performance of its students. But what yardstick should be used to accomplish this measure? The fact that modern education has many goals suggests that we must measure the extent of its success in a variety of ways. One important instrument of this measurement are the data gathered during the course of the National Assessment of Educational Progress (NAEP).

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 sampling and assessment methodology that is technically sophisticated. The results of NAEP are made available to both the professional and lay public continuously and with increasing frequency are cited as evidence in public debates about educational topics.

NAEP's results are complex, consisting, as they do, of:

(i) outcomes on achievement tests of complex character on a variety of subjects and

(ii) attitude and behavioral information from the children, teachers, and others associated with the children's schooling, as well as

(iii) 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. All data displays are used for one or more of four purposes:

1. Exploration — the data contain important messages, answers to questions that may be explicit in the viewer's mind or not. Looking at the data allows explicit questions to be answered and unthought of questions to be asked.

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 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 communi-

cation should not try to serve an archival purpose as well, since rules governing these two purposes are often antithetical.

In this report we focus on a single display format, the table, and examine ways to improve its performance for its various purposes. We do this because we heartily subscribe to the notion that although figures never lie, at least

> ."...*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

2. Tabular presentation

> "*Getting information from a table is like extracting sunlight from a cucumber*"
>
> Farquhar & Farquhar, 1891

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

Before exploring 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 appear 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 dominates their differences in age, training and interests. The sets of rules for table construction that emerges for each of the three groups are virtually identical in general structure and only vary because of constraints imposed by the increasing complexity of the data themselves.

Why are tables used to display data? 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 tables, but rules for table preparation have not been revised apace with this change in purpose.

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:

I. rounding heavily,

II. ordering rows and columns by some aspect of the data,

III. flanking the display with suitable summary statistics, and

IV. spacing to aid perception.

More recent work on effective tabular presentation (Wainer, 1992, 1993) elaborates and illustrates these simple rules for designing effective tables. Driving these rules is the orienting attitude that a table is for communication, not data storage. Modern data storage is accomplished well on magnetic disks or tapes, optical disks, or some other mechanical device. Paper and print are meant for human eyes and human minds.

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 smallish tables used principally for communication and for larger ones which seem to serve archival purposes as well.

2.1 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) we get:

POPULATIONS, IN MILLIONS, OF
DETROIT AND LOS ANGELES
1920-1970

	City	
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 to question 4 '1950'. It awaits empirical verification whether this is easier than before revision, but my intuition (and my ten 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:

i. Humans cannot understand more than two digits very easily.

ii. We can almost never justify more than two digits of accuracy statistically.

iii. We almost never care about accuracy of more than two digits.

Let us take each of these reasons separately.

Understanding. Consider the statement that "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 any statistic is proportional to one over the square root of the sample size. 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 ($1/\sqrt{n} \sim .005 \Rightarrow$

$\sqrt{n} \sim 1/.005 = 200$) 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 all 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 hundredth's 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. Equation (1) is taken from *State Court Caseload Statistics: 1976.*

$$\text{Ln(DIAC)} = -.10729131 + 1.00716993 \times \text{Ln(FIAC)} \qquad (1)$$

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

$$\text{DIAC} = .9\ \text{FIAC}. \qquad (2)$$

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. A colleague, who knows more about courts than I do, suggested that I 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 equation 1 can give statisticians a bad name[2].

Example 2. 1990 8th and 12th grade Science Assessment

Table 1

Original Table

	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

Any redesign task must first try to develop an understanding of purpose. The presentation of this data set 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 uses batteries up most quickly? The least quickly?

4. Are there any unusual interactions between equipment and battery brand?[3]

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 tabie to make your decision.*

Before going further I invite you to read Table 1 carefully and see to what extent you can answer these 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. Two useful ways to order the data are:

i. Size places — Put the largest first. 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.

ii. 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.

Table 2

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

Table 2 is a redone version of Table 1 in which batteries (rows) are ordered by battery life in a radio, longest lasting first. And 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 that,

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 is suitable, more often a median. But whatever is chosen it should be visually different than the individual entries and set spatially apart.

Table 3

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**

The summaries (means) surrounding Table 3 makes the row and column effects explicit. We now see that not only is the Never Die battery the best all around, but we have a measure of how much better. 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 between a table and 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 semi-graphical 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 their size. The rule then is

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 last version of Table 1 shown as Table 4.

Table 4

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**

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 a flashlight, which has twice the life again as would have 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 was transformed as we have in table 4. Should we transform the table? Structuring the table as I have is not based on the particular questions that were asked, but rather on general rules for all tables and would have been done in exactly the same way before seeing 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[4]. 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 vs. 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?

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 adhere 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 know?

In the concluding section of this paper 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

Question 9, associated with this table, asks

> *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 solve this 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[5], that orders the rows by the data rather than some arbitrary letter, removes the need for the second step. Also, by 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.

Revised Table

Ten Students' Test scores

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

In none of these examples was the preferred structure chosen on the basis of the specific questions asked. Each table was revised using the four rules specified, to the extent that each was needed. The fact that they then made answering the questions asked easier is a testament to the efficacy of the rules. It is my contention that we ought to prepare all data displays to be used as stimuli in a test item (tables in this instance) according to the highest standards. This will make most of the current crop of questions trivially easy, but will allow the test developer to ask deeper questions. I will discuss the character of such questions further in section 3.

2.2 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 sometimes saving space is 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 so as 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* (page 83).

This table, reproduced as Table 5 below, shows the average mathematics performance of 8th grade examinees from all participating jurisdictions in the 1992 state mathematics assessment as a function of parent's education. Also included are the percentages of examinees in each state whose parent's education is at each of the designated levels.

TABLE 212 Average Mathematics Proficiency by Parents' Highest Level of Education (continued)

PUBLIC SCHOOLS	Grade 8 - 1992									
	Graduated College		Some Education After High School		Graduated High School		Did Not Finish High School		I Don't Know	
	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)	257 (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 (0.8)	243 (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)	250 (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.

Included in parentheses are the standard errors of all figures presented.

Before attempting 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 ten 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.

Answers to all of these questions lie within the bounds of Table 2.12, but how easily can they be extracted? Can we ease the pain of this extraction through a change in the design of the table?

To allow easier data manipulation I have reformatted table 2.12 to separate better the levels of education and to place the standard errors in separately labeled columns.

Table 6

Average Mathematics Proficiency by Parents' Highest Level of Education
Grade 8 - 1992

PUBLIC SCHOOLS	Graduated College				Some Education After High School				Graduated High School				Did Not Finish High School				I don't Know			
	Percent of Students		Average Proficiency		Percent of Students		Average Proficiency		Percent of Students		Average Proficiency		Percent of Students		Average Proficiency		Percent of Students		Average Proficiency	
	%	se	θ	se	%	se	θ	se	%	se	θ	se	%	se	θ	se	%	se	θ	se
NATION	40	1.4	279	1.4	16	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	44	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	0.8	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	0.8	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	18	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
District of 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	268	1.9	25	1.0	246	1.8	6	0.5	242	3.5	18	0.8	246	2.1
Idaho	48	1.3	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	32	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	0.8	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	268	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	258	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	282	2.1
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	1.5	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	1.8	281	2.1	18	0.8	272	1.6	21	1.0	253	1.6	16	1.0	247	1.0	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

Placing standard errors in parentheses is a useful and time honored convention, but my purposes are served better this way. For the moment, allow me this intermediate modification. It will disappear by the time we're done. This redone table is shown as Table 6.

Let us begin the real work of the redesign by asking why would one 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[6]. The resulting table is shown as Table 7.

We have also moved the District of Columbia into the section of non states that also includes Guam and the Virgin Islands. All ordering is done within table section. Note that the key summaries are in **boldface type.**

Table 7

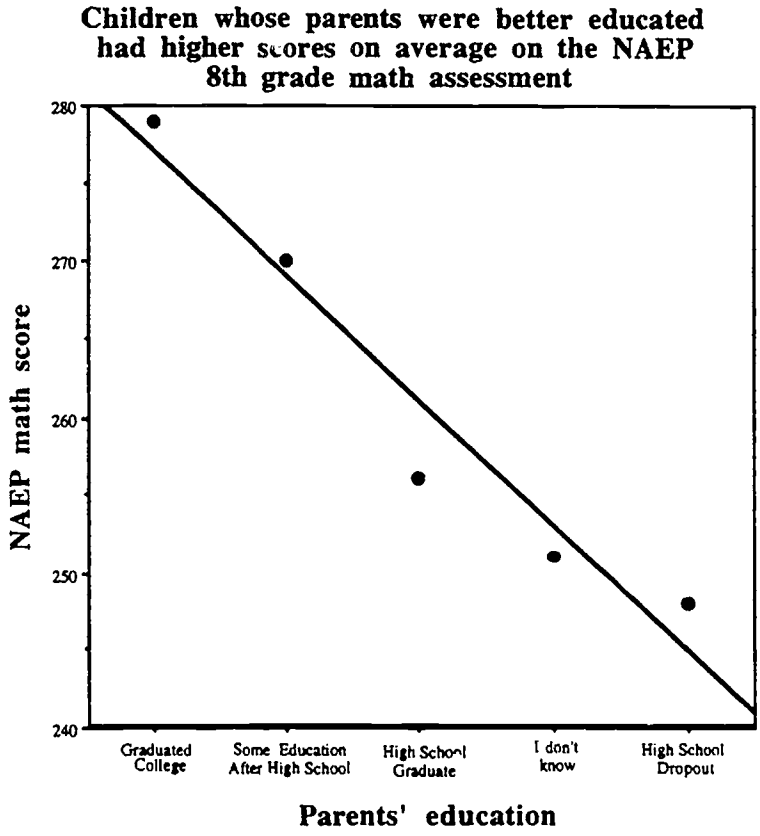
Average Mathematics Proficiency by Parents' Highest Level of Education
Grade 8 - 1992

PUBLIC SCHOOLS	Graduated College				Some Education After High School				Graduated High School				Did Not Finish High School				I don't Know				Average	
	Percent of Students		Average Proficiency		Percent of Students		Average Proficiency		Percent of Students		Average Proficiency		Percent of Students		Average Proficiency		Percent of Students		Average Proficiency			
	%	se	θ	se	%	se	θ	se	%	se	θ	se	%	se	θ	se	%	se	θ	se	θ	se
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	267	1.4
Central	44	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	274	2.5
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	267	3.9
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	267	2.7
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	258	2.4
STATES																						
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	283	1.4
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	283	1.5
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	282	1.6
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	278	1.5
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	2.1	278	1.4
Wisconsin	38	2.4	287	1.8	24	0.8	282	1.5	28	1.8	270	1.9	5	0.8	254	3.4	6	0.6	255	4.0	278	2.0
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	277	1.6
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	274	1.3
Wyoming	42	0.9	281	0.9	22	0.8	278	1.7	23	0.7	268	1.1	5	0.6	258	3.3	7	0.5	260	2.2	274	1.3
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	274	1.3
Connecticut	47	1.3	288	1.0	16	0.8	272	1.8	22	0.9	260	1.8	8	0.6	245	3.3	9	0.6	251	2.4	273	1.6
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	272	1.6
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	272	1.6
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	271	2.3
Pennsylvania	39	1.8	282	1.8	19	0.9	274	1.9	30	1.2	262	1.6	7	0.8	252	2.8	5	0.6	252	3.8	271	1.9
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	271	1.8
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	269	1.8
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	268	2.1
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	267	1.9
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	267	1.7
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	267	2.1
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	265	2.5
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	265	1.5
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	264	1.8
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	264	2.1
Texas	34	1.6	281	2.1	18	0.8	272	1.6	21	1.0	253	1.6	16	1.0	247	1.0	11	0.8	244	2.4	264	1.8
Delaware	39	1.2	274	1.3	18	1.0	266	2.3	30	1.0	251	1.7	8	0.5	248	4.0	8	0.9	248	3.4	262	1.9
Kentucky	28	1.4	278	1.6	19	0.8	267	1.6	32	0.9	254	1.6	15	0.9	248	1.7	6	0.4	242	2.8	261	1.7
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	260	2.2
South Carolina	37	1.4	272	1.5	16	0.7	266	1.7	31	0.9	248	1.4	9	0.6	248	2.1	7	0.3	247	3.0	260	1.7
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	259	2.1
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	259	1.8
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	259	1.5
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	258	2.0
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	258	1.5
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	258	1.7
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	257	1.9
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	256	1.9
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	251	2.2
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	249	2.2
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	246	1.8
OTHER JURISDICTIONS																						
Guam	26	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	235	2.0
District of 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	234	1.9
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	222	1.9

Table 7 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 4th and 8th 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 jurisdictions labeled 'territories.' 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.

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 heterogeneous in parental education. A small plot (below) of mean math performance against parents' education makes the quantitative aspect of this relationship clearer. This provides a reasonable answer to question 2.

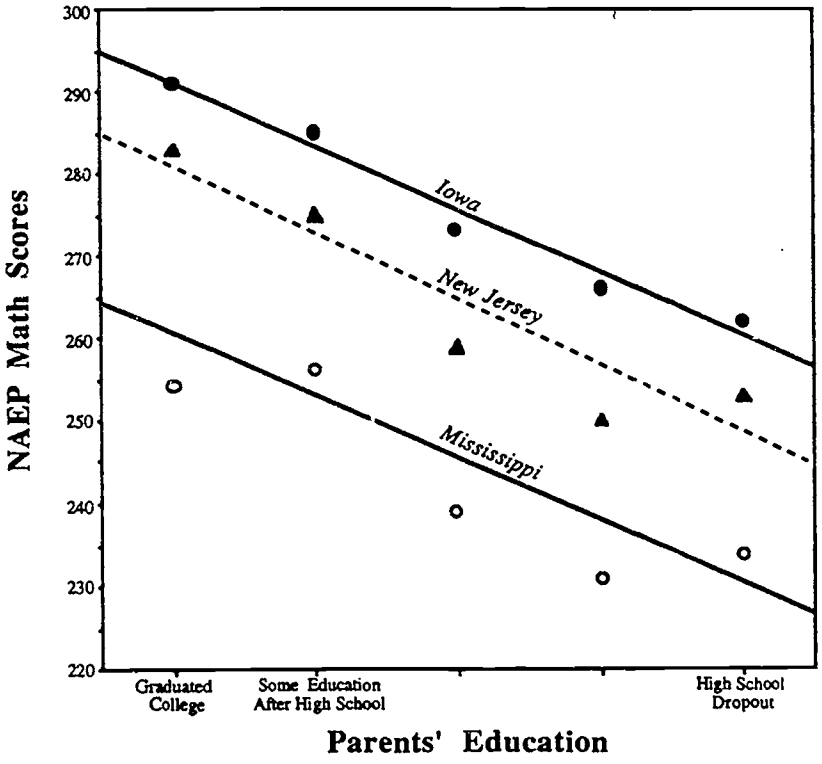
Figure 1



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 the one above for individual states. Such a graph, shown below as figure 2, contradicts the hypothesis that differences in states' overall performance is 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 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



Answering 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 hasn't that arithmetic 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 and thus 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 below as Table 8) segregates the standard errors into a separate table and substi-

tutes 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," multiples the first entry 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.[7]

Table 8

Average Mathematics Proficiency by Parents' Highest Level of Education
Grade 8 - 1992

PUBLIC SCHOOLS	Graduated College		Some Education After High School		Graduated High School		Did Not Finish High School		I don't Know		Mean
	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	
NATION	40	279	18	270	25	256	8	248	9	251	267
Central	44	283	20	273	26	264	4	***	7	258	274
Northeast	38	282	18	267	26	259	8	246	10	250	267
West	43	279	18	274	19	252	9	246	11	248	267
Southeast	35	270	17	263	28	249	12	248	8	248	258
STATES											
Iowa	44	291	21	285	25	273	4	262	5	266	283
North Dakota	54	289	18	283	19	271	3	259	5	272	283
Minnesota	48	290	21	284	22	270	3	256	7	268	282
Maine	40	288	22	281	26	267	6	259	5	266	278
Wisconsin	36	287	24	282	28	270	5	254	6	255	278
New Hampshire	48	287	17	280	24	267	6	259	7	262	278
Nebraska	48	287	20	280	24	267	4	247	6	256	277
Idaho	48	281	20	278	19	268	7	254	6	254	274
Wyoming	42	281	22	278	23	266	5	258	7	260	274
Utah	53	280	22	278	15	258	3	254	7	258	274
Connecticut	47	288	16	272	22	260	6	245	9	251	273
Colorado	46	282	19	276	21	260	6	250	7	252	272
Massachusetts	48	284	17	272	21	261	7	248	7	248	272
New Jersey	45	283	18	275	23	259	7	253	8	250	271
Pennsylvania	39	282	19	274	30	262	7	252	5	252	271
Missouri	36	280	22	275	29	264	8	254	6	252	271
Indiana	33	283	21	275	32	260	8	250	6	249	269
Ohio	37	279	19	272	32	260	7	243	5	249	268
Oklahoma	39	277	21	272	26	257	8	254	6	251	267
Virginia	41	282	18	270	24	252	9	248	8	251	267
Michigan	38	277	23	271	26	257	6	249	7	248	267
New York	44	277	18	271	23	256	6	243	10	240	265
Rhode Island	43	276	18	271	22	256	8	244	8	239	265
Arizona	36	277	22	270	21	256	10	245	12	248	264
Maryland	44	278	18	266	25	250	6	240	7	245	264
Texas	34	281	18	272	21	253	16	247	11	244	264
Delaware	39	274	18	266	30	251	6	248	8	248	262
Kentucky	28	278	19	267	32	254	15	246	6	242	261
California	39	275	18	266	17	251	10	241	16	240	260
South Carolina	37	272	16	268	31	248	9	248	7	247	260
Florida	39	268	19	266	24	251	8	244	10	244	259
Georgia	35	271	18	264	30	250	11	244	6	245	259
New Mexico	34	272	20	264	26	249	11	244	10	245	259
Tennessee	33	267	21	265	29	251	12	245	5	243	258
West Virginia	29	270	18	269	33	251	13	244	7	239	258
North Carolina	36	271	20	265	27	246	10	240	6	240	258
Hawaii	38	267	15	268	25	246	6	242	16	246	257
Arkansas	30	264	20	264	31	248	11	246	8	245	256
Alabama	33	261	18	258	29	244	13	239	7	237	251
Louisiana	32	256	20	259	30	242	10	237	7	236	249
Mississippi	36	254	16	256	29	239	13	234	7	231	246
Means	40	279	18	270	25	256	8	248	9	251	267
OTHER JURISDICTIONS											
Guam	28	246	13	244	27	229	10	224	22	226	235
District of Columbia	32	244	17	240	29	224	9	225	12	229	234
Virgin Islands	23	224	11	232	29	221	14	219	24	217	222
	Error terms for comparisons										
Max Std er of diff	3.4	3.5	2.1	3.4	2.5	3.5	1.4	6.4	1.7	6.4	3.5
40 Bonferr	11.0	11.3	6.8	11.0	8.1	11.3	4.5	20.7	5.5	20.7	11.3
820 Bonferroni	13.6	14.0	8.4	13.6	10.0	14.0	5.6	25.6	6.8	25.6	14.0

This table is not only a good deal clearer to look at, it is, for most prospective users, a good deal easier to use to make inferences about statistical significance of observed differences. As an example, note that all the observed differences between New Jersey and Iowa are statistically significant at the three highest levels of parental education, and marginally so at the lowest.

Last, there is no good reason remaining to combine mathematics achievement and percentage of children in each category into the same table. It seems to me that it would be clearer if they were separated, perhaps onto two tables on facing pages. To examine this we divided Table 8 into two parts, shown below as Table 9 and Table 10. Table 9 contains just mean mathematics proficiency; Table 10 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 10 are ordered by the state means from Table 9. 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 (e.g. **266**). Entries that are unusually small are boxed but unshaded (e.g., **240**). Thus in Table 9 we see that the average score of children whose parents had only some post high school education was unusually high. Similarly Nebraska and Connecticut's children of high school drop-outs scored unusually poorly.

Table 9

Average Mathematics Proficiency by Parents' Highest Level of Education
Grade 8 - 1992

	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**
	Central	283	273	264	***	258	**274**
	Northeast	282	267	259	246	250	**267**
	West	279	274	252	248	248	**267**
	Southeast	270	263	249	246	248	**258**
	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	Louisiana	256	259	242	237	236	**249**
41	Mississippi	254	256	239	234	231	**246**
	Means	**279**	**270**	**256**	**248**	**251**	**267**
	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 10

Percent of Children by Parents' Highest Level of Education
Grade 8 - 1992

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
Central	44	20	26	4	7
Northeast	38	18	26	8	10
West	43	18	19	9	11
Southeast	35	17	28	12	8
STATES					
Iowa	44	21	25	4	5
North Dakota	54	18	19	3	5
Minnesota	48	21	22	3	7
Maine	40	22	26	6	5
Wisconsin	38	24	28	5	6
New Hampshire	46	17	24	6	7
Nebraska	46	20	24	4	6
Idaho	48	20	19	7	6
Wyoming	42	22	23	5	7
Utah	53	22	15	3	7
Connecticut	47	16	22	6	9
Colorado	46	19	21	6	7
Massachusetts	48	17	21	7	7
New Jersey	45	18	23	7	8
Pennsylvania	39	19	30	7	5
Missouri	36	22	29	8	6
Indiana	33	21	32	8	6
Ohio	37	19	32	7	5
Oklahoma	39	21	26	8	6
Virginia	41	18	24	9	8
Michigan	38	23	26	6	7
New York	44	18	23	6	10
Rhode Island	43	18	22	8	8
Arizona	36	22	21	10	12
Maryland	44	18	25	6	7
Texas	34	18	21	16	11
Delaware	39	18	30	6	8
Kentucky	28	19	32	15	6
California	39	18	17	10	16
South Carolina	37	16	31	9	7
Florida	39	19	24	8	10
Georgia	35	18	30	11	6
New Mexico	34	20	26	11	10
Tennessee	33	21	29	12	5
West Virginia	29	18	33	13	7
North Carolina	36	20	27	10	6
Hawaii	38	15	25	6	16
Arkansas	30	20	31	11	8
Alabama	33	18	29	13	7
Louisiana	32	20	30	10	7
Mississippi	36	16	29	13	7
Means	40	18	25	8	9
OTHER JURISDICTIONS					
Guam	28	13	27	10	22
District of Columbia	32	17	29	9	12
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

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 two times the square root of the sum of the squared residuals) were then highlighted. Table 9 goes about as far as we might expect in displaying the results to answer all of the questions about achievement scores phrased earlier.

Last, we have combined the individual standard errors that were previously housed in the original table and piled them into Table 11. For consistency it probably would have been sensible to make up two tables of standard errors matching Tables 9 and 10, but we believe that this will be so rarely consulted that it wasn't worth the extra page. Users' experience will inform this judgment and we should be prepared to change the format if I am wrong.

Table 11

Standard Error of Average Mathematics Proficiency by Parents' Highest Level of Education
Grade 8 - 1992

PUBLIC SCHOOLS	Graduated College		Some Education After High School		Graduated High School		Did not Finish High School		I don't Know	
	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency
NATION	1.4	1.4	0.6	1.2	0.8	1.4	0.6	1.8	0.5	1.7
Central	2.7	2.9	1.4	1.6	1.7	2.3	0.7	[illegible]	0.8	3.8
Northeast	3.1	4.2	1.1	3.0	2.2	4.2	0.9	4.2	1.2	3.3
West	2.9	2.6	1.2	2.6	1.5	2.9	1.1	2.4	0.9	2.9
Southeast	1.9	1.9	0.8	2.0	1.4	1.9	1.6	4.2	1.0	4.3
STATES										
Iowa	1.4	1.2	0.8	1.5	1.1	1.3	0.4	2.4	0.4	2.8
North Dakota	1.2	1.1	0.7	1.9	1.3	1.7	0.5	4.5	0.5	2.8
Minnesota	1.3	1.0	0.9	1.8	0.9	1.8	0.4	4.2	0.6	3.0
Maine	1.5	1.4	1.0	1.5	1.1	1.1	0.5	2.7	0.5	2.6
New Hampshire	1.5	1.4	0.8	1.5	1.1	0.9	0.5	2.5	0.5	2.1
Wisconsin	2.4	1.8	0.8	1.5	1.8	1.9	0.6	3.4	0.6	4.0
Nebraska	1.5	1.2	1.0	1.6	1.2	1.7	0.5	3.3	0.6	3.8
Idaho	1.2	0.9	0.8	1.3	0.9	1.4	0.5	2.3	0.5	2.8
Utah	1.3	1.0	1.0	1.2	0.8	1.8	0.3	3.2	0.5	2.7
Wyoming	0.9	0.9	0.8	1.7	0.7	1.1	0.6	3.3	0.5	2.2
Connecticut	1.3	1.0	0.8	1.8	0.9	1.8	0.6	3.3	0.6	2.4
Colorado	1.2	1.3	0.9	1.6	0.9	1.5	0.6	2.4	0.5	2.6
Massachusetts	1.5	1.3	0.8	1.8	1.0	1.4	0.6	3.2	0.6	2.6
Missouri	1.3	1.7	0.9	1.5	1.0	1.6	0.7	2.4	0.5	2.9
New Jersey	1.6	1.8	0.8	2.1	1.2	2.5	0.6	3.8	0.7	3.9
Pennsylvania	1.8	1.6	0.9	1.9	1.2	1.6	0.8	2.8	0.5	3.8
Indiana	1.5	1.5	0.9	1.9	1.1	1.6	0.6	2.6	0.5	3.3
Ohio	1.4	1.8	0.7	1.6	1.1	2.3	0.6	2.6	0.5	4.5
Michigan	1.6	2.2	0.9	2.0	0.9	1.7	0.5	2.0	0.6	3.0
Oklahoma	1.4	1.5	0.9	1.9	1.0	1.7	0.7	2.9	0.5	4.3
Virginia	1.5	1.5	0.8	1.6	0.9	1.5	0.6	2.1	0.6	2.5
New York	1.8	1.9	1.1	2.4	1.0	2.5	0.8	4.2	1.0	3.8
Rhode Island	1.1	1.1	1.5	1.5	1.4	1.6	0.4	2.1	0.6	2.5
Arizona	1.5	1.5	1.0	1.5	0.9	1.6	0.7	2.5	0.8	2.7
Maryland	1.7	1.8	0.9	1.9	1.2	1.8	0.8	3.7	0.5	3.8
Texas	1.6	2.1	0.8	1.6	1.0	1.6	1.0	1.0	0.8	2.4
Delaware	1.2	1.3	1.0	2.3	1.0	1.7	0.5	4.0	0.9	3.4
Kentucky	1.4	1.6	0.8	1.6	0.9	1.6	0.9	1.7	0.4	2.8
California	1.8	2.0	1.0	2.1	0.9	2.1	0.9	2.2	1.1	2.9
South Carolina	1.4	1.5	0.7	1.7	0.9	1.4	0.6	2.1	0.3	3.0
Florida	1.5	1.9	0.7	1.9	1.1	1.8	0.7	2.7	0.7	3.2
Georgia	1.7	2.1	0.7	1.7	1.2	1.3	0.8	2.2	0.6	2.6
New Mexico	1.4	1.4	0.7	1.4	1.1	1.4	0.7	1.9	0.6	2.0
North Carolina	1.2	1.4	0.8	1.6	0.9	1.7	0.6	2.3	0.5	3.6
Tennessee	1.5	2.1	0.9	1.8	1.0	1.6	0.8	2.0	0.4	3.6
West Virginia	1.1	1.5	0.8	1.4	1.1	1.2	0.9	1.8	0.4	2.3
Hawaii	1.1	1.5	0.9	1.9	1.0	1.8	0.5	3.5	0.8	2.1
Arkansas	1.1	1.9	0.8	1.7	1.1	1.6	0.7	2.4	0.6	2.7
Alabama	1.6	2.5	0.7	2.0	1.1	1.8	0.9	2.0	0.6	2.9
Louisiana	1.4	2.5	0.9	1.8	1.3	1.6	0.7	2.4	0.6	3.7
Mississippi	1.7	1.6	0.7	2.0	1.4	1.6	0.8	1.8	0.6	2.8
OTHER JURISDICTIONS										
Guam	1.2	1.9	0.7	2.4	1.1	1.9	0.9	2.5	1.2	2.0
Dist. of Columbia	1.0	1.7	0.8	1.9	0.8	1.6	0.7	3.2	0.6	2.2
Virgin Islands	1.1	2.0	0.8	2.4	0.9	1.9	0.9	2.4	1.0	1.4
Maximum se of difference	3.4	3.5	2.1	3.4	2.5	3.5	1.4	6.4	1.7	6.4

Thus we have found that by separating variables into separate tables that are only tangentially related, once some important summaries are calculated yields tables of 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 some electronic data base 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.[8] Its principal drawback is that alphabetic ordering 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 2.12, are constructed and ordered by overall performance (instead of alphabetically), we have made finding a particular state a bit more difficult[9]. 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 Table 9. Then have a small alphabetically ordered locator table that connects state names to row numbers in the empirically ordered tables.

Table 12

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

2.3 Small Tables in NAEP reports

We have seen, in Section 2.1, that even tables which contain only a few numbers can still be made more comprehensible through the application of four simple rules. We have also seen, in section 2.2, that these same rules offer more help in larger tables. Before designing any table we must first determine the questions that will plausibly be asked of it and include in that table only those data that facilitate answering those questions.

Table 13 is an example of a table of modest size, whose principal purpose must surely be to communicate information about the differences in students' reading performance in different parts of the country, in the changes in children's performance over 8 years of schooling, and interactions between growth rate and geographic location. Usually questions about change over time are better answered in a graph than a table, but with only three time points a table may not be a disastrous place to begin. One such revision is shown in Table 14. In it we have replaced the standard errrors with an estimate of the maximum standard error of the difference, emphasized the average proficiency, deemphasized the percentage of students in each section (probably could have been omitted entirely), eliminated irrelevant spaces, and ordered the regions by their overall mean proficiencies. These changes improve the table's comprehensibility considerably. It is now clear that the West and Southeast perform worse in 4th grade and that the West catches up by 12 grade[10]. The students in the Southeast are as far behind the rest of the country in 12th grade as they were in 4th.

TABLE 2.3 Average Reading Proficiency and Achievement Levels by Region, Grades 4, 8, and 12, 1992 Reading Assessment

	Percentage of Students	Average Proficiency	Percentage of Students At or Above			
			Advanced	Proficient	Basic	Below 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)	4(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(1.2)	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.

Table 14

Grade 4	Average Proficiency	Percentage of students at or above			Below Basic	Percentage of Students
		Advanced	Proficient	Basic		
Northeast	223	7	31	63	37	21
Central	221	4	26	63	37	27
West	215	4	24	56	44	28
Southeast	214	4	21	54	46	23
Grade 8						
Northeast	263	3	31	71	29	22
Central	264	2	31	73	27	25
West	260	2	27	68	32	28
Southeast	254	1	22	63	37	25
Grade 12						
Northeast	293	4	40	76	24	24
Central	294	3	40	79	21	26
West	292	4	38	77	23	27
Southeast	284	2	28	68	32	23
Maximum Standard error of difference	4.4	2.4	4.8	4.7	4.7	1.5

There are two aspects of this table that are worthy of further discussion. First the construction emphasizes geographic comparisons rather than time trends. Second, the use of cumulants to characterize the proportion of children at each level masks one phenomenon. An alternative construction (shown in Table 15) provides us with a different view. By grouping together the three grade levels within geographic region we see that average proficiency increases about 40 points from 4th to 8th grade and about 30 points from 8th to 12th grade. This is approximately true for all four regions.

By replacing the cumulant percentages at each level with the actual percent we see a rather remarkable effect. Specifically that there is about an 8% increase in children at the Basic level in 8th over what is found in 4th grade and a concomitant decrease in the percentage of 8th grade children at 'Below Basic.' Then we see about an 8% increase in 12th graders at 'Proficient and Advanced' over what was seen in 8th grade. This two step movement of children from 'below basic' to 'basic' in the 4th to 8th grade period and from 'basic' to 'proficient' in the 8th to 12th grade period was entirely invisible in the original table. Because it is not visible in the mean proficiency scores, the cause of this effect is probably definitional.

Table 15

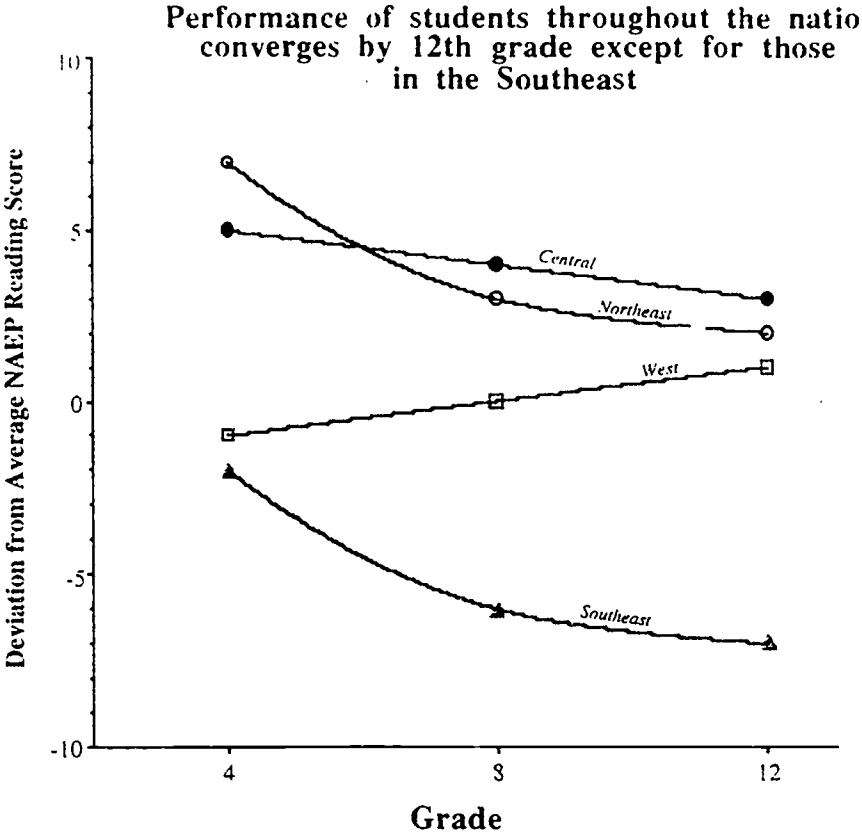
		Percentage of students at		
Northeast	**Average Proficiency**	Advanced and Proficient	Basic	**Below Basic**
Grade 4	**223**	31	32	**37**
Grade 8	**263**	31	40	**29**
Grade 12	**293**	40	36	**24**
Central				
Grade 4	**221**	26	37	**37**
Grade 8	**264**	31	42	**27**
Grade 12	**294**	40	39	**21**
West				
Grade 4	**215**	24	32	**44**
Grade 8	**260**	27	41	**32**
Grade 12	**292**	38	39	**23**
Southeast				
Grade 4	**214**	21	33	**46**
Grade 8	**254**	22	41	**37**
Grade 12	**284**	28	40	**32**
Maximum Standard error of difference	**4.4**	4.8	4.7	4.7

Can this display be improved still further? Yes. One way would be to break it up into smaller displays. For example, the average proficiencies are best shown as a two way table by themselves (see Table 16 below). 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 a much more evocative image could be obtained by subtracting out the grade effects and the grand mean and then plotting the residuals. This makes clear just how different the Southeast region is (see figure below) and exposes some trends in the residuals.

Table 16

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

Figure 3



In the same way, the percentage distributions can be shown as a two-way table, but they aren't interesting enough to bother plotting.

Table 17

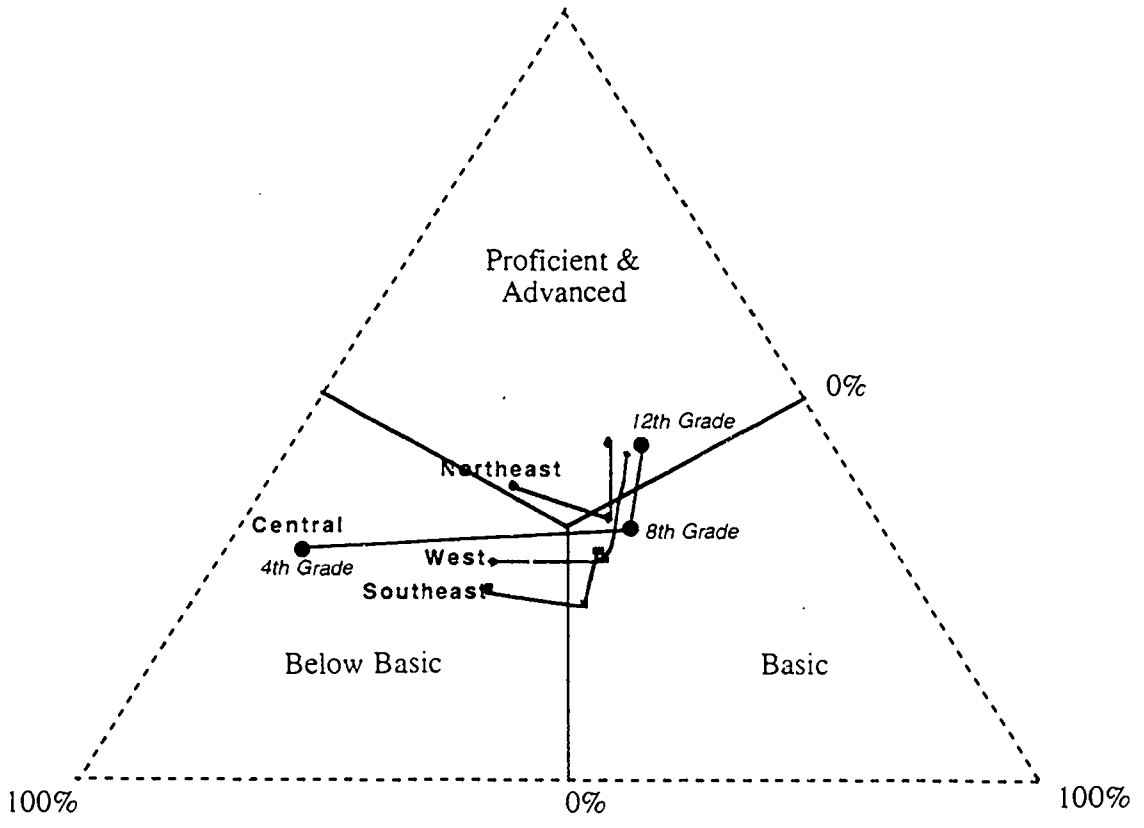
Percentage of students

	Grade 4	Grade 8	Grade 12	**Average**
Northeast	21	22	24	**22**
Central	27	25	26	**26**
West	28	28	27	**28**
Southeast	23	25	23	**24**

The data entries showing the percentage of children at each proficiency level are well shown with a trilinear plot (Wainer, 1994a). We see that the line that represents each section of the country extends horizontally from 4th to 8th grade indicating that the modal category shifts from 'Below Basic' to 'Basic' during that time period. The line makes a right angle turn moving upward toward 'Proficient & Advanced' as the modal category in 12th grade. While this phenomenon is discernible within the revised Table 15, once we know to look for it, it cannot be missed in this figuration. One is reminded of John Tukey's (1977) famous aphorism that a "good graph forces us to see what we weren't expecting."

Figure 4

1992 NAEP Reading Assessment for 4th, 8th & 12th grade in the four sections of the country



3. Discussion and Conclusions

There are no good or poor graphs or good or poor tables. Rather, some constructions answer the questions one is entitled to ask and 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 section 2. 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?

We have seen in section 2.3 that even though a table is a two-way array, it can sometimes be used to capture three- or more-way data. This yields more possible questions about the additional main effects and the various sorts of higher interactions. The goal of effective display is to ease the viewer's task in answering these questions. We have found that ordering, rounding, summarizing and spacing wisely 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 (combining two or more tables into one) and communicating clearly. It has been our experience that often breaking up complex displays sensibly communicates more efficiently, em for em, than a large compound table.

Why is this true? To understand a graph involves two distinct phases of perception:

1. What are the components of data that are being reported?

2. What are the relations among them?

The first phase is easy if the horizontal and vertical components are unitary (e.g., Level of Proficiency vs. Region). It becomes more difficult if they are not (e.g., Level of Proficiency and Average Proficiency and Percent vs. 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, it aids comprehension by keeping the components of data simple and presenting tables paired. This was illustrated in section 2.2 when we separated the percentage distributions from the achievement scores into separate tables.

3.1 Error

The treatment of error is similar. It is often critical to know the standard errors of any reported statistic. But why? Sometimes it is to help in determining if an entry is close enough to zero to be ignored. Or in many cases, to allow us to compare any two figures in the table. If the latter, the standard errors are only important in so much as they allow us to compute the standard error of the difference between the two means of interest. We suggested simplifying the table by removing the standard errors to another table and replacing them with an estimate of the upper bound of the standard error of the difference. This we calculated using the two largest values of standard errors in each column. Of course this number can only be used 'as-is' if one only plans on a single test per column. Otherwise it needs to be increased to control the type I error on the set of tests that you perform. To aid in this task we boosted it up to represent two possible circumstances. The first is comparing one state with all others. This yields the possibility of 40 separate tests. The Bonferroni inequality then suggests multiplying the standard error of the difference by 3.2 to yield a .05 bound. The second possibility is to make all 820 (41 choose 2) pairwise comparisons. This requires multiplying by 4.0 for a .05 bound. We thus replaced each column of 41 individual standard errors with these three numbers. We believe that this is both more parsimonious and more useful. Of course, some users may find the bounds we calculated to be too rough and require more precision. They will need to go back to the original standard errors and perhaps use more powerful procedures for multiple comparisons. But we believe that this sort of use is sufficiently rare to warrant the simplification we have proposed.

3.2 Measuring numeracy

In section 2.1 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 helped to establish both how tables ought to be prepared as well as how easy it is to answer some questions from well-prepared tables.

However, 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 in section 2.1 and reproduced again here as Table 18. This is identical to Table 4 shown earlier except that four unusual entries have been indicated by boxing them in. A shaded box indicates a '2 hours too high' entry; an unshaded box means '2 hours too low' entry. Thus we would have expected the Never Die battery to last only 26 hours in a radio and a Constant Charge battery to last 21.

Table 18

	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**

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 in section 2.1 are some typical examples. We can do better with the guidance of a formal theory of graphic communication. What follows is an expansion of a theory proposed more than a decade ago (Wainer, 1980).

3.2.1 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) three levels of questions are:

- Elementary level questions involve data extraction, e.g., "How long does a Servo-Cell last in a Cassette Player?"
- Intermediate level questions involve trends seen in parts of the data, e.g., "How much longer is a battery likely to last in a radio than in a portable computer?"
- Overall level questions involve an understanding of the deep structure of the data being presented in their totality, usually comparing trends and seeing groupings, e.g., "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 (e.g., a trend - level 2) and then we look for points that do not con-

form to this trend (level 1). Such questions are hard to answer from a raw table (i.e. Table 1) but are trivial in Table 17 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[11].

The epistemological basis of this formulation was clearly stated by the Harvard mathematician and philosopher Charles Sanders Peirce (1891). He felt that all things could be ordered into monads, dyads, and triads, which he often characterized as firstness, secondness and thirdness.

Firstness considers a thing all by itself, for example redness. Secondness considers one thing in relation to another, for example a red apple. Thirdness concerns two things 'mediated' by a third, for example an apple falling from a tree. The tree and the apple are linked by the relation 'falling from.' Peirce applied firstness, secondness and thirdness to every branch of philosophy. There is no need, he argued, to go on to fourthness or fifthness and so on, because in almost every case these higher relations can be reduced to combinations of firstness, secondness and thirdness. On the other hand, genuine thirdness can no more be reduced to secondness than can genuine secondness to firstness.[12]

Peirce traces the origins of this architecture of theory to Kant's *Critique of Pure Reason*, but enough is uniquely Peirce's to credit him as its progenitor. One can think about it linguistically as firstness being like an noun, secondness as adjective noun combinations, and thirdness including a verb. Once again we can see that each level cannot be constructed from a lower one, and that we have no need for a concept of fourthness or more. How does this apply to the measurement of numeracy?

Reading a table at the intermediate level is clearly different than 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 (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 mixing together different dependent variables in the same display.

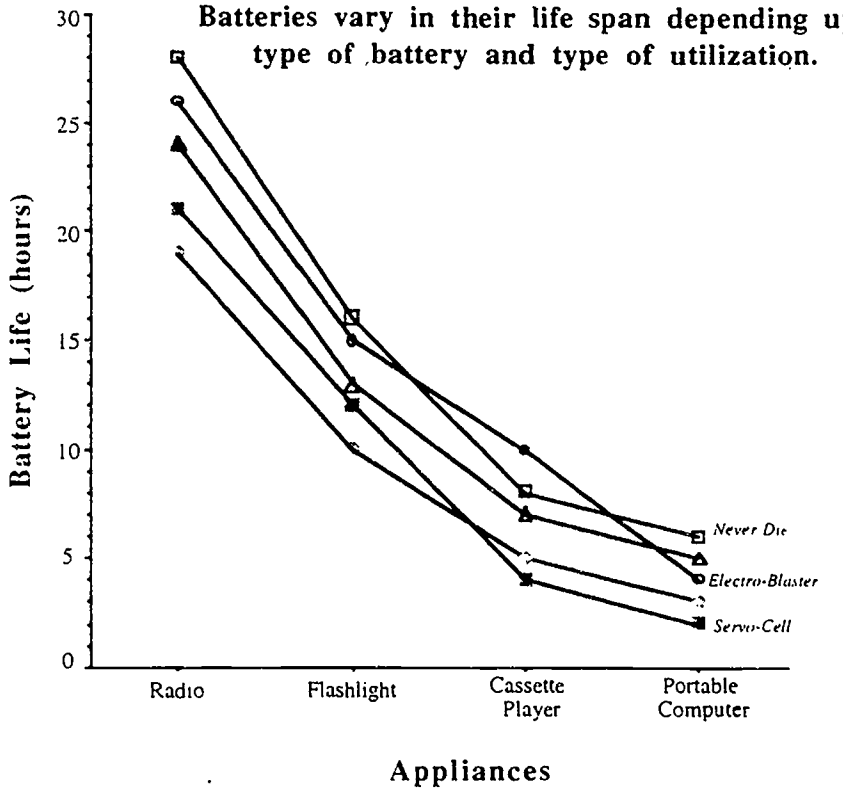
I hope that this brief introduction conveys a sense of how this formal structure can make it easier to construct tests of numeracy, and to understand better which characteristic of numeracy we are measuring. Of course, to ask questions at higher levels requires data of sufficient richness to support them, as well as tables clear enough for the quantitative phenomena to show through[13]. It is much more difficult to answer second or third level questions from Table 18 than from Table 9. It is also easier to see trends, and deviations from them, with a different display format altogether (see Figure 4 below). Once again we see that the format we choose must be based upon our purpose in constructing

the display. While elementary level questions are best answered with a table, level 2 and 3 questions are 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 non tabular 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 unalterable 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;[14] a better table of the same data ought to make the item easier. Similarly a more numerate audience makes a table appear more efficacious.

Figure 5



3.3 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. We believe that the archival purpose is anachronistic and so focused our attention on rules for building tables to facilitate their efficacy as communicative devices. We found that the same four rules apply whether aimed at the simplest tables used as stimuli within the assessment or 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. One must be careful not to try to do too much, for oftentimes what is best for one purpose is antithetical to another -- clear communication and archival completeness are two purposes with divergent requirements that come immediately to mind.

Of foremost importance is the notion that we are typically not looking at a table for elementary purposes. To become involved in a problem and to understand it is to shift from an elementary reading to a more global one. The construction of efficacious data displays aims to promote this transition; allowing a reader the graceful change from spectator to participant.

4. References

Bertin, J. (1973). *Semiologie Graphique*. The Hague: Mouton-Gautier. 2nd Ed. (English translation done by William Berg & Howard Wainer and published as *Semiology of Graphics*, Madison, Wisconsin: University of Wisconsin Press, 1983.)

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.

Finn, C. E. (June 15, 1994). Drowning in Lake Wobegone. *Education Week*, P. 31,35.

Gardner, M. (1978). On Charles Sanders Peirce: Philosopher and Gamesman, *Scientific American, 240,* 18-26.

Peirce, C. S. (1891, January). The architecture of theories. *The Monist,* 161-176.

State Court Caseload Statistics: 1976 (1976). Williamsburg, VA: National Center for State Courts.

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. (1990). Measuring graphicacy. *Chance, 3,* 52 & 58.

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

Wainer, H. (1993). Tabular presentation. *Chance, 6(3),* 52-56.

Wainer, H. (1994a) Using Trilinear Plots for NAEP State Data. Technical Report 94-6. Educational Testing Service, Princeton, NJ.

Wainer, H. (1994b). On the Academic Performance of New Jersey's Public School Children: I. Fourth and Eighth Grade Mathematics in 1992, *Education Policy Analysis Archives*, *2(10)*, Entire issue.

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

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*. Technical Report #40. Department of Statistics: The University of Chicago.

Table Captions:

Table 5. Original Table 2.12 from *Data compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States* (page 83).

Table 6. Reformatted version of Table 5 in which standard errors are in separately labeled columns and categories of parental education are separated.

Table 7. Table 6 with average state performance shown, rows ordered and spaced by average performance.

Table 8. Table 7 with individual standard errors replaced by conservative estimates.

Table 9. A revision of Table 8 including only average student mathematics proficiency; unusual entries are highlighted and a state locator index inserted.

Table 10. A revision of Table 8 including only percentage distribution of parental education; unusual entries are highnlighted.

Table 11. The standard errors of the various state means ordered and spaced as in Table 7.

Table 12. Alphabetically ordered locator table of the states, to be used in case of an emergency loss of any particular jurisdiction.

Table 13. Original Table 2.3 from *NAEP Reading, Report Card for the Nation and the States* (page 89).

Table 14. Reformatted version of Table 13 in which standard errors are replaced with a conservative estimate of the standard error of the difference, rows are ordered by Average Proficiency, and columns are reordered.

Table 15. Reformatted version of Table 14 in which years are grouped together and the cumulant percentages of children at each proficiency level are replaced by the actual percentages.

Table 16. Regional proficiencies removed and reformatted to emphasize two way structure.

Table 17. Regional percentages removed and reformatted to emphasize two way structure.

Table 18. Revision of Table 4 highlighting unusual entries.

Footnotes

---

[1]This work was sponsored by the National Center for Education Statistics through contract number R999B40013 to the Educational Testing Service, Howard Wainer, Principal Investigator. Although I am pleased to express my gratitude for this support, I must re-express 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 also 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 than Brent Bridgeman, John Mazzeo and Keith Reid-Green 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.

[2]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.

[3]This question ought to be inappropriate in this context. It is hard to imagine a physical reason why a a battery that lasts longer in a cassette recorder would not last longer in a flashlight. But this is the table that was on the test. The only plausible interpretation of interactions here is that of error.

[4]As a silly, but compelling example, consider the following released NAEP reading passage from the 1992 Reading Assessment
(P. 285)
I AM ONE
I am only one.
But still I am one.
I cannot do everything,
But still I can do something:

And because I cannot do everything
I will not refuse to do the something that I can do.

-- Edward Everett Hale

One can imagine inventing up some questions to test a student's understanding of this simple poem, but suppose we reordered the rows alphabetically?
I AM ONE (lines ordered alphabetically)

And because I cannot do everything
But still I am one.
But still I can do something;
I am only one,
I cannot do everything,
I will not refuse to do the something that I can do.

This certainly makes it harder. Also, are we still testing the same construct? And does the same scoring rubric still hold? To add to this silliness, suppose we followed the usual dictates of table preparation that insist that we not "waste space". We might get

I AM ONE (lines ordered alphabetically and then spaced "efficiently")

And because I cannot do everything But still I am one. But still I can do something; I am only one, I cannot do everything, I will not refuse to do the something that I can do.

We can surely go further in this direction. Suppose we order all the words alphabetically and then space things out efficiently. We would arrive at:

I AM ONE (words ordered alphabetically and then spaced "efficiently")

am am And because But But can can cannot cannot do do do do do everything everything, I I I I I I I nct one, one only refuse something something; still still that the to will.

This shows us how much redundancy there was in the original poem. We can now easily remove it and fix up flawed capitalization yielding,

I AM ONE (words ordered alphabetically, redundancies removed and then spaced "efficiently")

Am and because but can cannot do everything, I not one, only refuse something; still that the to will.

Surely, all will agree that a test built of such a mish-mash might be testing something, but it would have little to do with a student's proficiency at reading and understanding poetry. If ordering alphabetically and spacing for 'efficiency' is silly with words (and it is) why do we even consider doing it with data?

[5]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 gradebooks which are alphabetized for ease of data entry and have a second version for retrieval. This discussion is about retrieval.

[6]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)

[7]Choosing the maximum may be too conservative for many users. Two alternatives that may be considered are: (i) 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. (ii) Replace MAX(se) with a more average figure (i.e. $\sqrt{\sum_{k=1}^{n} se_k^2 / n}$). 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, but will leave its resolution to other accounts.

[8]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.

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

[10] Making longitudinal inferences from cross-sectional data is always risky. In this instance the 'catching-up' of the west may instead be a falling back of the 4th and 8th grade students in the west due to the influx of a large number of immigrant children into those grades. This possible artifact would be eliminated if these data were standardized to a fixed population mix. One example of this (Wainer, 1994b) was the subject of considerable attention recently (Finn, C. E. (June 15, 1994))

[11]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.

[12]This paragraph is a pretty close paraphrasing of a description by Martin Gardner (1978, p. 23)

[13]*Purgamentum init, exit purgamentum.*

[14]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.