

Alternative Displays for Communicating NAEP Results: A Redesign and Validity Study

Howard Wainer

Educational Testing Service

Ronald K. Hambleton

University of Massachusetts at Amherst

Kevin Meara

University of Massachusetts at Amherst

Five displays, chosen from the NAEP 1994 Reading: A First Look, were redesigned. The redesign was informed by the principles developed and enunciated in Wainer's 1997 popular text Visual Revelations. After the redesign was completed a survey of educational policymakers was done in which substantive questions were asked about the content of the various displays. Each redesign was paired with the original and were assigned randomly to one of two survey forms. We found that, on average, the redesigns yielded both more accurate and faster answers to the questions asked. The more difficult the question the greater the disparity between the original format and the redesigned one.

The National Assessment of Educational Progress (NAEP) is a large scale survey of US primary and secondary education. It employs a stratified three-stage clustered sampling design for students in various age and grade groups, and a complex partially balanced incomplete block design for the administered items. The questionnaire items are classified into attitude and cognitive items. In addition there are background questions that all examinees are requested to answer. Performance is characterized with an estimated parameter representing the mean of the distribution of a group of examinees' locations on the ability measured by the assessment.

The data are complex and multivariate. The published reports typically present small univariate excerpts from the vast ramifying reality of the entire database; behind these reports stand data almanacs with page after page of cross-tabulations of outcomes for each question stratified by each of the background variables. Even the reports intended primarily for the use of policymakers are, in our view, simultaneously too much and not enough. They are too much because it is difficult to go

This research was supported by the National Center for Education Statistics, contract R999B60018 to the Educational Testing Service, Howard Wainer, Principal Investigator. The authors are pleased to have the opportunity to acknowledge this help. Any opinions expressed here are those of the authors and do not necessarily reflect that of the National Center for Education Statistics or the US Government. We are also grateful to a number of state departments of education for providing participants and making local arrangements for the study: Connecticut, Massachusetts, New Jersey, North Carolina, Pennsylvania, and Rhode Island. Without their cooperation the study could not have been completed. Last we would like to thank Alex Sedlecek for her helpful comments on an earlier draft. Of course, as always, any errors that remain are solely the responsibility of the second author.

through all of these results and obtain a coherent idea of the main findings. They are not enough because stratifying by only one variable at a time is often insufficient to enable interested individuals to obtain the insights that are needed to fully utilize the information. Unfamiliar reporting scales and statistics further complicate the interpretation and use of the results for policymakers.

The value of this survey is directly related to the extent to which its results are understood and disseminated. At the same time, the disparity in resources allocated to the gathering of the data when compared to disseminating them is striking. There are at least as many pitfalls associated with data display as with test construction, sampling, and analysis (Hambleton & Slater, 1995, 1996; Wainer, 1984, 1990a, 1990b, 1991a, 1991b, 1991c, 1991d, 1992a, 1992b, 1992c, 1993a, 1993b, 1993c, 1993d). This situation can be helped enormously through a careful consideration of the modes of presentation that are used in this dissemination, and empirical verification of their usefulness for their intended audience. It is toward such a careful consideration and testing that this research was aimed.

The Study

Two earlier NAEP-based projects examined the efficacy of NAEP data displays. One, reported on by Hambleton and Slater (1996), presented education policymakers as well as education professionals with existing NAEP displays within the context of a NAEP report and measured the extent of their understanding. A second, whose results were reported by Wainer (1996a, 1996b, 1997a, 1997b, 1997c), examined many of the same displays found wanting in the Hambleton & Slater survey and redesigned them following well-established principles of design.

This study has a simple concept. We used the principles exemplified in Wainer's previous studies and the Hambleton and Slater findings to redesign a set of displays whose formats are representative and broadly used in many NAEP reports up through the end of 1994. We then subjected these new designs, paired with the same data in their original format in an expanded version of Hambleton and Slater's survey. The goal was to see to what extent the redesigns aided in comprehensibility. Speed of response to the questions was added as another important variable since policymaker time spent reading reports of test results is typically very limited.

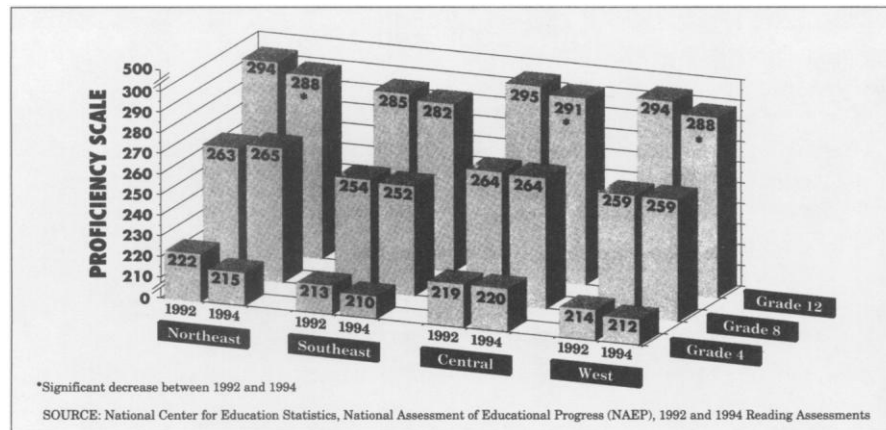
In choosing displays for this study, attention was focused on the report: *NAEP 1994 Reading: A First Look* (Williams, Reese, Campbell, Mazzeo, & Phillips, 1995). This report is representative of many that are aimed at a broad audience of education policymakers and education professionals, as well as containing some tables and graphs whose features previous research has shown to be amenable, in prospect, to improvement through redesign.

The Original Displays, the Redesigns, and the Principles Used in the Redesign

The first display

The first display chosen (see display 1-O) depicts the average reading proficiency by grade and region for two separate years. Three goals of such a display must surely be to:

Figure 3. Average Reading Proficiency by Grade and by Region — NAEP 1992 and 1994



DISPLAY 1-O. A three dimensional bar chart showing the average reading proficiency by grade and region for two years. This was Figure 3 in Williams et al., 1995.

- (i) show the changes over those two years by region,
- (ii) show the differences among regions within grade, and
- (iii) show the differences among the grades.

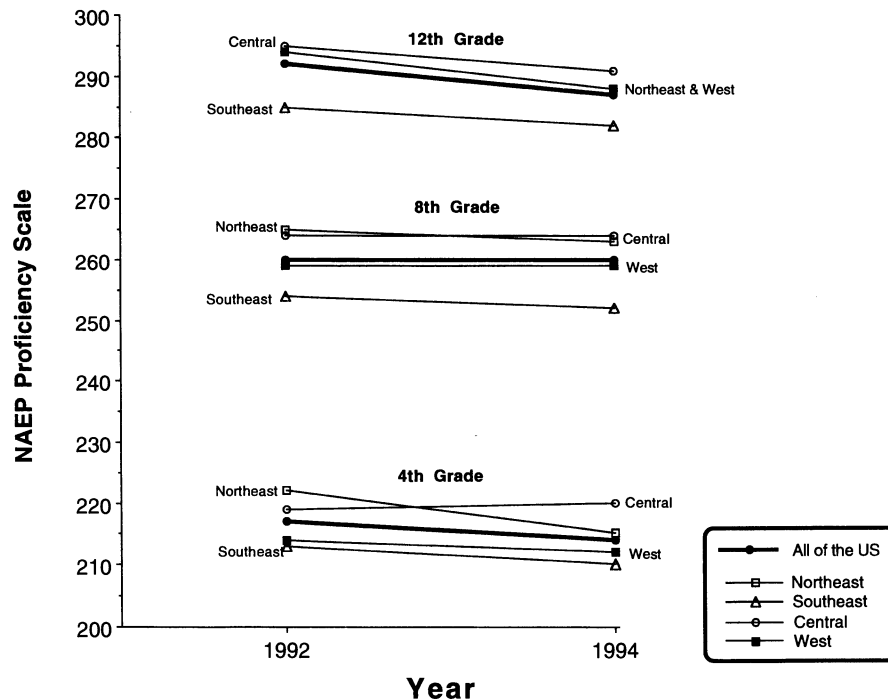
While the inclusion of the actual values associated with each makes retrieving those values trivial, it contributes to the necessity of “reading” the display rather than “seeing” it (to use Bertin’s [1983] useful distinction). “Seeing” a display allows deep structure to be internalized without recourse to linguistic decoding. It is thus much faster and provides a longer lasting (more memorable) mental image. A view of display 1-O conveys the image that the scores are vastly different as students progress from 4th to 8th and on to 12th grade. Not much else seems obvious.

The alternative, shown as display 1-R, is vastly simplified. The three dimensional bars are replaced by points, and the change over time is represented by a line connecting those points. Questions that might be phrased that are associated with the three goals stated above are easily answered. The decline among 4th and 12th graders is clearer, as is the relatively poor performance of the Southeast region. The differences among the three grade levels remain clear and emphasize the larger gains made between 4th and 8th grade than between 8th and 12th. Such effects, even now that we know to look for them, are still not obvious in the original display.

The second display

The second display chosen (see display 2-O) is a composite display consisting of two tables. Together they show reading performance for each grade by race and parents’ education level. The principal points of such a display are obviously to show the differences in performance by race and by education level. The point of placing them together is to provide context, for the interpretation of the observed differences of either set of results is aided through its juxtaposition with the other.

Average Reading Proficiency by Grade and Region NAEP 1992 and 1994



DISPLAY 1-R. *The alternative we proposed replaces the pseudo 3-D bars with simple lines. The size of the differences among the three grade levels is clear, as is the relatively poor performance of the Southeast.*

Our goal in the redesign was to make the tables more like a graph by utilizing space to convey information (see display 2-R). One can see in the revised display what needed to be read in the original one; namely, that the variation in performance associated with differences in the level of parental education is about the same as the differences associated with race.

A fundamental canon of effective display is know the communicative purpose of the display and do not try to do too much. We felt that the principal goal of this display ought to be to show the variation in performance as a function of parental education and then to use these differences to place ethnic differences into a more meaningful context. To try to also include changes that the scores have undergone since the 1992 assessment seemed peripheral to the principal message. Hence we decided to omit that one bit of information here with the understanding that it could be carried in a separate display that focused specifically on changes over time.

TABLE 2 Average Reading Proficiency by Race/Ethnicity THE NATION'S REPORT CARD 1992 1994 Reading Assessment			
	1994		Change From 1992
	Percentage of Students	Average Proficiency	
Grade 4			
Total	100	214	-2
Race/Ethnicity			
White	69	224	-1
Black	15	187	-6
Hispanic	12	191	-10*
Asian	2	232	—
Pacific Islander	1	219	—
American Indian	2	201	-5
Grade 8			
Total	100	260	0
Race/Ethnicity			
White	70	268	0
Black	15	237	-1
Hispanic	11	240	-1
Asian	2	273	—
Pacific Islander	1	259†	—
American Indian	1	251	0
Grade 12			
Total	100	287	-5*
Race/Ethnicity			
White	73	294	-4*
Black	13	265	-8*
Hispanic	8	270	-9*
Asian	3	280	—
Pacific Islander	1	280†	—
American Indian	1	275†	—

Differences between two groups may be partially explained by other factors not included in this table.
 The NAEP reading scale ranges from 0 to 500.
 * The value for the 1994 assessment was significantly different from the value for 1992 at about the 95 percent confidence level.
 † Interpret with caution any comparisons involving this statistic. The nature of the sample does not allow accurate determination of the variability of this value.
 The standard errors for the 1994 national averages are between 0.7 and 1.0 scale score points. The standard errors for the 1994 race/ethnicity averages range from 0.7 to 7.4 points.
 — Due to significant changes in the wording of the race/ethnicity question between the 1992 and 1994 assessments, the 1992 results for Asian and Pacific Islander students are not comparable to 1994 results. Therefore, 1992 results for these two subgroups are not presented.
 *** Sample size is insufficient to permit a reliable estimate.
 Percentages may not total 100 percent due to rounding or, in the case of the race/ethnicity variable, because some students categorized themselves as "other."
 SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.

TABLE 4 Average Reading Proficiency by Parents' Education Level THE NATION'S REPORT CARD 1992 1994 Reading Assessment			
	1994		Change From 1992
	Percentage of Students	Average Proficiency	
Grade 4			
Total	100	214	-2
Parents' Education Level			
Graduated College	42	224	-1
Some Education After HS	8	223	1
Graduated HS	13	207	-5
Did Not Finish HS	4	188	-10
I Don't Know	34	206	-4
Grade 8			
Total	100	260	0
Parents' Education Level			
Graduated College	43	270	-1
Some Education After HS	20	266	0
Graduated HS	21	252	1
Did Not Finish HS	7	238	-5
I Don't Know	9	238	1
Grade 12			
Total	100	287	-5*
Parents' Education Level			
Graduated College	43	298	-3*
Some Education After HS	25	289	-5*
Graduated HS	21	277	-6*
Did Not Finish HS	7	266	-9*
I Don't Know	3	248	-10*

Differences between two groups may be partially explained by other factors not included in this table.
 The NAEP reading scale ranges from 0 to 500.
 * The value for the 1994 assessment was significantly different from the value for 1992 at about the 95 percent confidence level.
 The standard errors for the 1994 national averages are between 0.7 and 1.0 scale score points. The standard errors for the 1994 parents' education level averages range from 0.9 to 3.4 points.
 Percentages may not total 100 percent due to rounding.
 SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments.

DISPLAY 2-O. Two tables showing average reading proficiency by grade, parents' education, and ethnicity for 1992. These were Tables 2 and 4 in Williams et al., 1995.

The third display

The third display chosen (see display 3-O) is a table showing the reading proficiency among 4th graders, by state, for 1992 and 1994. The redesign (display 3-R) uses Wainer's (1997c, Chapter 10) precepts for tabular display. Since one obvious goal of such a display is to make comparisons among states (which states did best? worst? which ones are like mine?) it seemed sensible to ease this task by ordering the states by their overall performance (high to low) and by spacing the states in a way that is reflective of their performance. We omitted any indication of statistical significance because earlier surveys of the population for which this display was intended strongly indicated that such notation yielded more confusion than enlightenment. This suspicion was confirmed in the current survey.

Average Reading Proficiency 1994

Grade 12

By Parents' Education Level		By Race/Ethnicity	
Graduated College	298	White	
	296		
	294		
	292		
	290		
Some Education after HS			
Total US	288	Total US	
	286		
	284		
	282	Asian, Pacific Islander	
	280		
	278		
Graduated HS		American Indian	
	276		
	274		
	272		
	270	Hispanic	
	268		
	266	Black	
Did Not Finish HS			
	266		
	264		
	262		
	260		
	258		
	256		
	254		
	252		
	250		
I Don't Know	248		

Grade 8

	274	Asian
	272	
	270	
	268	White
	266	
	264	
	262	
Total US	260	Total US
	258	Pacific Islander
	256	
	254	
	252	
	250	American Indian
	248	
	246	
	244	
	242	Hispanic
	240	
	238	
	236	Black
	234	
	232	
	230	
	228	
	226	
	224	Graduated College
	222	Some Education after HS
	220	
	218	
	216	
	214	
Total US	214	Total US
	212	
	210	
	208	Graduated HS
	206	I Don't Know
	204	
	202	American Indian
	200	
	198	
	196	
	194	
	192	Hispanic
	190	
	188	Black
	186	

Grade 4

Grade	US Mean
12	288
8	260
4	214

DISPLAY 2-R. The alternative we proposed replaces the tables with three back-to-back stem and leaf plots. The scales representing the separate grades are aligned to add context to the within grade differences observed.

The fourth display

The fourth display chosen (see display 4-O) is a graphic representation of the 1994 state data shown previously in displays 3-O and 3-R. The design is in the form of Cartesian Rectangles. This display reorders the data that was previously shown alphabetically in display 3-O, thus allowing comparisons to be done more

TABLE 6 Average Grade 4 Reading Proficiency NAEP Trial State Assessments in Reading Public Schools Only THE NATION'S REPORT CARD 1992 1994 Reading Assessment		
	1992 Average Proficiency	1994 Average Proficiency
Nation	215	212
Region		
Northeast	220	212
Southeast	211	208
Central	218	218
West	212	212
State		
Alabama	207	208
Arizona	209	206
Arkansas	211	209
California	202	197<
Colorado	217	213
Connecticut	222	222
Delaware	213	206<<
Florida	208	205
Georgia	212	207
Hawaii	203	201
Indiana	221	220
Iowa	225	223
Kentucky	213	212
Louisiana	204	197<<
Maine	227	228
Maryland	211	210
Massachusetts	226	223
Minnesota	221	218
Mississippi	199	202
Missouri	220	217
Montana†	—	222
Nebraska†	221	220
New Hampshire†	228	223<
New Jersey	223	219
New Mexico	211	205<
New York	215	212
North Carolina	212	214
North Dakota	226	225
Pennsylvania†	221	215<
Rhode Island†	217	220
South Carolina	210	203<<
Tennessee†	212	213
Texas	213	212
Utah	220	217
Virginia	221	213<<
Washington	—	213
West Virginia	216	213
Wisconsin†	224	224
Wyoming	223	221
Other Jurisdictions		
DoDEA	—	218
Guam	182	181

Differences between two groups may be partially explained by other factors not included in this table.
<< The value for 1994 was significantly lower than the value for 1992 at or about the 95 percent certainty level.
These notations indicate statistical significance from a multiple comparison procedure based on 38 jurisdictions participating in both 1994 and 1992. If looking at only one state, < indicates the value for 1994 was significantly lower than the value for 1992 at or about the 95 percent certainty level. Statistically significant differences between 1994 and 1992 for the state comparison samples for the nation and regions are not indicated.
† Did not satisfy one of the guidelines for school sample participation rates (see Appendix A).
— Jurisdiction did not participate in 1992 Trial State Assessment
DoDEA Department of Defense Education activity Overseas Schools
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessments

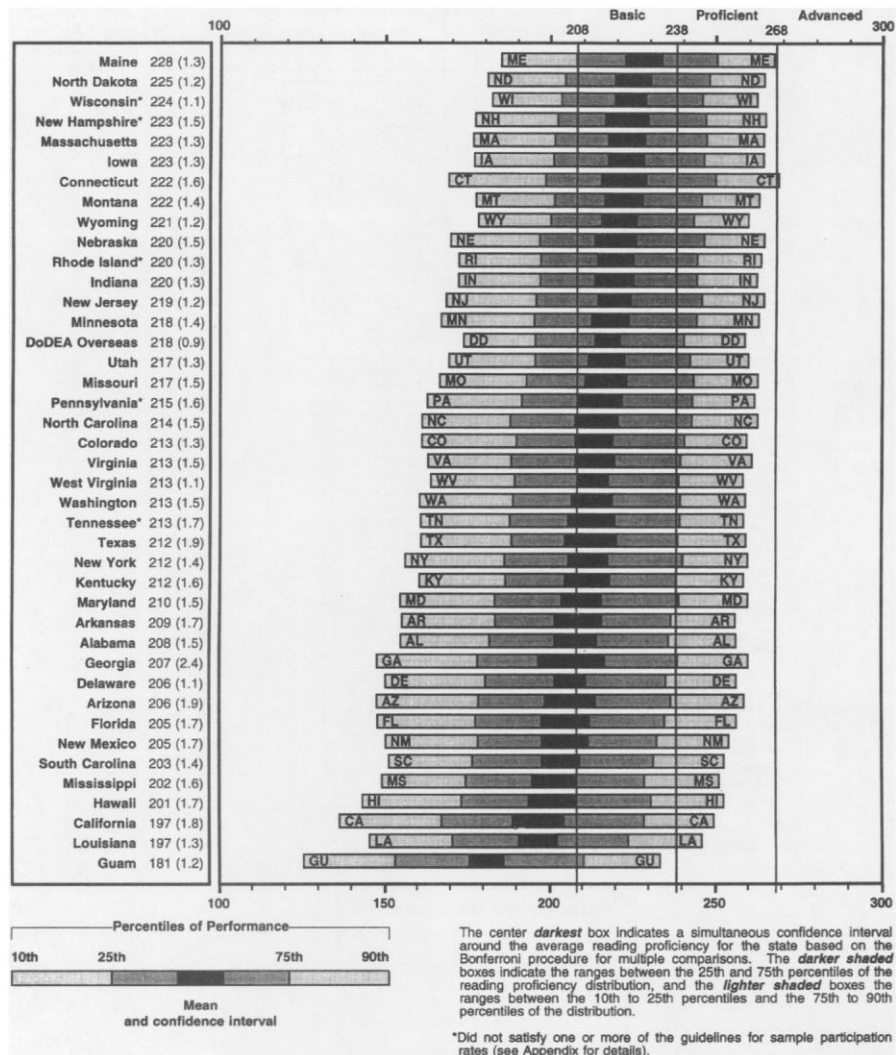
DISPLAY 3-O. A table showing average reading proficiency alphabetically by state and region, for 1992 and 1994. This was Table 6 in Williams et al., 1995.

Table 6. Alternate

Average Grade 4 Reading Proficiency NAEP Trial State Assessments for Public Schools States Ranked by 1994 performance		
	1992 Average Proficiency	1994 Average Proficiency
Nation	215	212
Region		
Central	218	218
Northeast	220	212
West	212	212
Southeast	211	208
State		
Maine	227	228
North Dakota	226	225
Wisconsin	224	224
New Hampshire	228	223
Massachusetts	226	223
Iowa	225	223
Montana	—	222
Connecticut	222	222
Wyoming	223	221
Indiana	221	220
Nebraska	221	220
Rhode Island	217	220
New Jersey	223	219
Minnesota	221	218
Missouri	220	217
Utah	220	217
Pennsylvania	221	215
North Carolina	212	214
Washington	—	213
Virginia	221	213
Colorado	217	213
West Virginia	216	213
Tennessee	212	213
New York	215	212
Kentucky	213	212
Texas	213	212
Maryland	211	210
Arkansas	211	209
Alabama	207	208
Georgia	212	207
Delaware	213	206
Arizona	209	206
New Mexico	211	205
Florida	208	205
South Carolina	210	203
Mississippi	199	202
Hawaii	203	201
Louisiana	204	197
California	202	197
Other Jurisdictions		
DoDEA	—	218
Guam	182	181

DISPLAY 3-R. *The alternative we proposed reorders the states by the 1994 average proficiency. Physical gaps are inserted where there is a matching gap in the data series.*

Figure C.1 Distribution of Overall Reading Proficiency Organized by Average Proficiency for the 1994 Trial State Reading Assessment, Grade 4, Public Schools Only

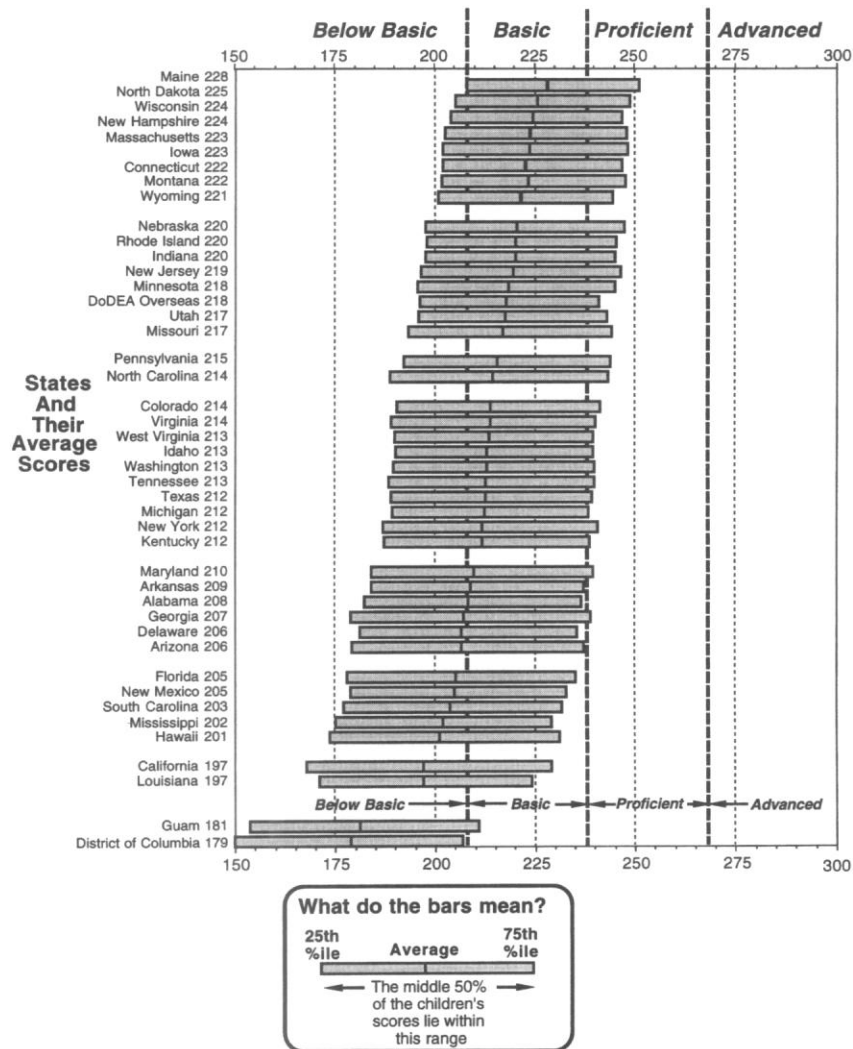


DISPLAY 4-O. A graph showing the average reading proficiency by state for grade 4 in 1994. This is ordered by average proficiency (as Display 3-R). It also shows the variability of performance within each state on a background that depicts the four performance categories. This was Figure C.1 in Williams et al., 1995.

easily. Previous research showed that display 4-O was particularly ineffective at communicating information to educational policymakers (Hambleton & Slater, 1996) and it was for that reason that this display was included in our research.

The redesign (display 4-R) simplifies some aspects of the data by only showing the 25th and 75 percentiles of the state distributions. It also places vertical gaps in

Reading Proficiency for the 1994 Trial State Assessment Grade 4, Public Schools



DISPLAY 4-R. *The alternative we proposed simplifies the depiction of variability by only showing the distribution between the 25th and 75th percentiles. Physical gaps are inserted where there is a matching gap in the data series and there are a few other cosmetic changes.*

TABLE 10		Reading Achievement Levels by Parents' Education Level									
1992 Assessment						1994 Assessment					
		Percentage of Students					Percentage of Students				
	Percentage of Students	At or Above Advanced	At or Above Proficient	At or Above Basic	Below Basic	Percentage of Students	At or Above Advanced	At or Above Proficient	At or Above Basic	Below Basic	
Grade 4											
Total		6	29	62	38		7	30	60	40	
Parent's Education Level											
Graduated College	39	10	39	71	29	42	11	39	70	30	
Some Education after H.S.	9	8	33	69	31	8	9	37	70	30	
Graduated High School	12	3	22	58	42	13	4	22	54	46	
Did Not Finish High School	4	1	12	39	61	4	1	9	32	68	
I Don't Know	36	3	21	55	45	34	4	22	52	48	
Grade 8											
Total		3	29	69	31		3	30	70	30	
Parent's Education Level											
Graduated College	41	5	40	80	20	43	5	40	79	21	
Some Education after H.S.	19	3	32	76	24	20	3	33	77	23	
Graduated High School	24	1	19	61	39	21	1	20	62	38	
Did Not Finish High School	8	1	13	51	49	7	0	10	46	54	
I Don't Know	8	0	12	45	55	9	0	12	48	52	
Grade 12											
Total		4	40	80	20		4	36<	75<	25>	
Parent's Education Level											
Graduated College	41	6	52	87	13	43	7	48	84<	16>	
Some Education after H.S.	27	3	41	83	17	25	3	36	78<	22>	
Graduated High School	22	2	28	72	28	21	2	24	66<	34>	
Did Not Finish High School	8	0	21	63	37	7	1	15	53<	47>	
I Don't Know	2	0	10	44	56	3	0	6	32	68	

Differences between two groups may be partially explained by other factors not included in this table.

< The value for the 1994 assessment was significantly lower (<: higher) than the value for 1992 at about the 95 percent confidence level.

The percentages of students in the subgroups may not total 100 percent due to rounding.

The standard errors for the 1994 (a) Advanced Level, parents' education level percentages range from 0.3 to 1.8, (b) Proficient Level, parents' education level percentages range from 1.2 to 2.6, and (c) Basic Level, parents' education level percentages range from 0.7 to 3.9.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1992 and 1994 Reading Assessment.

DISPLAY 5-O. A table showing the cumulative proportion of children in each of the four performance levels by grade and parents' education level, for 1992 and 1994. This was Table 10 in Williams et al., 1995.

the display commensurate with the gaps in the states' performance to ease the task of clustering states of similar performance. There are also a few other cosmetic changes in labeling and emphasis.

The fifth display

The fifth display chosen (see display 5-O) is also a table showing the proportion of students at each of the four levels of reading proficiency among 4th, 8th and 12th graders, as a function of parent's education level for 1992 and 1994. The principal difference seen in the redesign (display 5-R) is a replacement of the cumulative proportion in some categories by the proportion that were actually in that category. This important change was made because we have found that for the typical population of readers of NAEP reports, the use of cumulants caused considerable confusion (Hambleton & Slater, 1996). The redesign was chosen to ease this difficulty.

Table 10 Alternate

Reading Achievement Levels by Parents' Educational Level

		1992 Assessment Percentage of Students					1994 Assessment Percentage of Students				
		Percentage of Students	Below Basic	Basic	Proficient	Advanced	Percentage of Students	Below Basic	Basic	Proficient	Advanced
Grade 4											
Total			38	33	23	6		40	30	23	7
Parents' Education Level											
Graduated College		39	29	32	29	10	42	30	31	28	11
Some Education After HS		9	31	36	25	8	8	30	33	28	9
Graduated High School		12	42	36	19	3	13	46	32	18	4
Did Not Finish High School		4	61	27	11	1	4	68	23	8	1
I Don't Know		36	45	34	18	3	34	48	30	18	4
Grade 8											
Total			31	40	26	3		30	40	27	3
Parents' Education Level											
Graduated College		41	20	40	35	5	43	21	39	35	5
Some Education After HS		19	24	44	29	3	20	23	44	30	3
Graduated High School		24	39	42	18	1	21	38	42	19	1
Did Not Finish High School		8	49	38	12	1	7	54	36	10	0
I Don't Know		8	55	33	12	0	9	52	36	12	0
Grade 12											
Total			20	40	36	4		25	39	32	4
Parents' Education Level											
Graduated College		41	13	35	46	6	43	16	36	41	7
Some Education After HS		27	17	42	38	3	25	22	42	33	3
Graduated High School		22	28	44	26	2	21	34	42	22	2
Did Not Finish High School		8	37	42	21	0	7	47	38	14	1
I Don't Know		2	56	34	10	0	3	68	26	6	0

DISPLAY 5-R. The alternative we proposed has many cosmetic differences but the principal difference is that percentages rather than cumulants are displayed.

Statistical significance

The redesigns omit much “clutter” associated with making judgments about statistical significance; standard errors, Bonferroni bounds, differential shading and appended *’s are all missing. Why? We did not elide these important accoutrement of statistical inference lightly. But after careful consideration we found that: (a) most of the policymakers that we viewed as the audience of these documents did not make use of this aspect of the displays at all, and when they did (b) they were more likely to misuse them than use them correctly, and moreover, (c) removing them made the principal messages of the display clearer. Last, if we may be permitted an *obiter dictum*, the arbitrary establishment of 95% bounds as the *sine qua non* of two units being different is too often misleading. If the difference between the performance of two states is only significant at the 0.30 level (far from the minimally acceptable level), it does not mean that we should accept the performance of those two states as indistinguishable. There is evidence that one did better than the other; weak evidence to be sure, but evidence nevertheless. If we had to place a bet as to which state would do better on retesting there would be no hesitation. So, in the end, we felt that there was more to be gained in terms of proper usage than was to be lost by omitting such material. This conclusion might be different if the displays were for a different audience with a different purpose.

The Questions Asked

The questions asked of educational policymakers were developed to mirror three levels of information that can be carried in a display. Also, the questions were the ones we believed that policymakers or the public might ask when presented with the original displays. Before going further in describing the questions asked, a brief description of the theory that informed the development of the questions will be presented.

Rudiments of a theory of graphicacy

Fundamental to the measurement of graphicacy is the broader issue of what kinds of questions graphs can be used to answer. Wainer’s (1992c) revisions of Bertin’s (1973) three levels of questions are:

- (i) Elementary level questions involve data extraction, for example, “What was the mean reading proficiency of 4th graders in New York in 1992?”
- (ii) Intermediate level questions involve trends seen in parts of the data, for example, “Between 1992 and 1994 how has the performance of New York’s 4th graders changed?”
- (iii) Overall level questions involve an understanding of the deep structure of the data being presented in their totality, usually comparing trends and seeing groupings, for example, “Which groups of states shows the most dramatic increase in performance?” or “Which states show the same pattern of growth?”

These kinds of questions are often used in combination; for example in the detection of outliers—unusual data points—we need a sense of what is usual (e.g., a trend—level 2) and then we look for points that do not conform to this trend (level 1).

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.²

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 be reduced to firstness.³

Peirce traced 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 a noun, secondness as being 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 graphicacy?

Reading a graph at the intermediate level is clearly different than at the elementary level; a concept of trend requires the notion of connectivity. If the horizontal axis were not two years but instead two countries ordered alphabetically, the idea of an increasing trend would be meaningless. Comparing trends among different grades likewise requires an additional notion of connectivity, but this time across the dependent variable (for example, reading proficiency). This connectedness is characterized by a common vertical axis.

Experience has shown that questions associated with graphics tend to be questions of the first kind, although often they are compounded through the use of nongraphical complexity. For example, one might ask several elementary level questions and ask that numerical operations be performed on them. We tried to avoid such nongraphical complexity. Instead, we tried to make up questions associated with each level. For those interested, we have included, in Appendix A, all of the actual questions asked. Here, we shall discuss only a sampling of six of those questions.

Display 3:

Question 5—A level 3 question requiring comparisons and groupings
Identify the top three states in reading performance in 1994.

Display 4:

Question 2—A level 1 question requiring the extraction of a pair of numbers
What is the range of scores of the middle 50% of the grade 4 students in New York?

Question 6—A level 3 question requiring comparisons and groupings
Maine is the best performing state. In which states would Maine's 25th percentile student be about average? Name Three states.

Question 7—A level 3 question requiring comparisons and groupings
In how many states was the average student performing in the Below Basic category?

Display 5:

Question 2—A level 1 question requiring the extraction of a single number
What percentage of the students performed at the Basic level?

Question 5—A level 1 question requiring the extraction of a single number
In 1994, what percentage of 8th grade students with parents who did not finish high school were in the Basic performance category?

These six questions were not chosen at random. These were the six questions in which there was the greatest differences in performance between those interviewed in the study who were presented the original display and those who saw the redesign. A quick perusal of the questions and their associated displays will confirm that these represent a reasonable sample of questions that the data in these displays were gathered to be able to answer. In the next subsection we describe the details of the interview protocol used, and in the following section we present the results.

The Interview Protocol

In June and July of 1997, pilot testing was carried out to assess the suitability of the interview protocol and seven alternative graphical displays of NAEP findings. Our goal was to improve both, prior to the actual experimental study. A group of ten individuals comprised primarily of University of Massachusetts School of Education faculty and doctoral level graduate students was interviewed. The ten were selected because of their interest in educational reports such as NAEP executive summaries. In addition, five doctoral students who served as interviewers were asked to critique the interview protocol and displays, and their comments were used in the revision process.

The materials developed for the study included two forms of data displays (denoted, A and B), the interview protocol, and answer sheets for recording interview responses. For the pilot testing, Forms A and B each consisted of seven displays. Form A contained four displays that were originals and three that were alternatives. Form B contained four that were alternatives and three that were originals. Original displays were photocopied directly from the 1994 NAEP reading executive summary report. Alternative displays, designed by the first author with critical reviews from the UMass team of researchers, were major revisions of the originals. The displays were set up such that the originals in Form A corresponded to the alternatives in Form B and vice versa. During pilot testing the third draft of the interview protocol was used. It consisted of twelve background questions and an average of twelve questions per display.

Interviews were conducted on an individual basis by a trained interviewer. Each participant was randomly assigned to either Form A or Form B. After obtaining background information, the panelist was presented with the first display and

questioned about information contained in the display. The interview was complete when all seven displays had been discussed. Because response time was an important variable for comparing the displays, the amount of time the participant took to answer each question was measured with a handheld stopwatch operated by the interviewer.

As a result of pilot testing, changes were made to both the interview protocol and the displays. One major finding from the pilot test was that the interviewers were taking longer than expected. Our expectation was that an interview lasting longer than 60 minutes would complicate the problem of finding suitable participants (recall, our main interest was with educational policymakers) for the main study, and possibly lead to fatigue effects toward the end of the interviews. The times for completing forms ranged from 65 to 95 minutes. We found that fatigue set in after about an hour. To accomplish a 50-minute interview two displays were dropped from the study.

Some questions were revised and many questions were removed from the interview protocol. Questions which did not relate directly to the relative strengths and weaknesses of original and revised displays were dropped. Questions which were awkward to administer or which were almost always answered correctly were also dropped. The number of questions asked per display was reduced from 12 to no more than 7.

There were three additional noteworthy changes made. First, due to its cumbersome nature, the separate answer sheet was dropped. The interview protocol was revised so that answers and response times could be recorded directly on the interview protocol. Second, the timing of responses was reduced from all items to only a few. This change was intended to make the interview less pressured. Finally, copies of the questions were printed out for the panelists so they could read along with the interviewers. It was hoped that this would reduce confusion.

The final version of the interview protocol was administered by six trained interviewers to, mainly, educational policymakers in six state departments of education. About 20% of the interviews were conducted with teachers and graduate students majoring in educational leadership and policy. A total of 102 interviews was conducted with each person receiving either Form A or Form B, assigned at random.

The Results

Performance Results

The truth is rarely pure and never simple. This is especially true in any sort of experiment involving humans. Thus we were surprised and delighted to discover the remarkably consistent results from the interviews.

Table 1 (based on 102 interviews) provides the percent of correct answers to each question presented in the interviews. The data are reported separately for the original and revised versions of the displays. On display 1, there were essentially no performance differences, with respondents doing very well in answering the questions regardless of the version of the display they were given. Similar findings were observed with display 2 though there may have been a slight improvement in

Table 1. Summary of performance on all items

Mean Percent Correct										
Display 1										
		Question								
		1	2	3	4				Mean	
	Redesign	1.00	0.91	.094	0.98				0.96	
	Original	0.96	0.70	0.94	1.00				0.90	
	Difference	0.04	0.20	0.00	0.00				0.05	
Display 2										
		Question								
		2	3	5.1	5.2	5.3	5.4	5.5	5.6	
	Redesign	0.94	0.96	0.94	0.94	1.00	1.00	0.94	0.96	0.96
	Original	0.98	0.94	0.94	0.91	0.96	0.96	0.85	0.91	0.93
	Difference	-0.04	0.02	0.00	0.03	0.04	0.04	0.09	0.05	0.03
Display 3										
		Question								
		2	4	5	6					
	Redesign	0.92	0.98	1.00	0.89				0.95	
	Original	0.90	0.98	0.64	0.88				0.85	
	Difference	0.02	0.00	0.36	0.02				0.10	
Display 4										
		Question								
		1	2	3	4	5	6	7		
	Redesign	0.82	0.74	0.39	0.39	0.53	0.63	0.80	0.61	
	Original	0.58	0.40	0.58	0.51	0.42	0.32	0.18	0.43	
	Difference	0.24	0.34	-0.19	-0.12	0.11	0.31	0.62	0.18	
Display 5										
		Question								
		1	2	3	4	5	6	7		
	Redesign	0.80	0.96	0.88	0.96	0.90	0.94	0.92	0.92	
	Original	1.00	0.20	0.92	0.96	0.39	0.94	0.96	0.77	
	Difference	-0.10	0.76	-0.04	0.00	0.51	0.00	-0.04	0.15	

performance with the revised display (the mean performance was higher on six of the eight questions by an average of 4.0%).

Sizable performance differences were observed with the other three displays. With display 3, the revised version resulted in improved performance on three of the four questions, and on one of the questions (question 5), the improvement was 36% (64% vs. 100%).

With display 4, the revision resulted in considerably higher performance levels on four of the seven questions (over a 25% improvement), and on one of these four

Table 2. Median Response times, in seconds, for test items, averaged within display.

	Display				
	1	2	3	4	5
Redesign	15	5	11	32	4
Original	14	5	18	37	9
Difference	1	0	-7	-5	-5

questions, the improvement was over 60%. On one question, there was a reversal: the original display produced better results—about 20% (question 3).

With display 5, the revision resulted in a highly significant change in performance on two of the seven questions. On question 2, the increase in performance was 76.1% (from 20.0% with the original display to 96.1% with the revised display). On question 5, the increase was 51.4% (from 38.8% with the original display to 90.2% with the revised display).

When considering the full set of 30 common questions to the presentation of original and revised displays—on only two questions did the redesigned displays do worse by more than 10% (yield correct responses less frequently) than the original displays and on those, the decrease in performance was modest (19.8% and 13.1%). With nine of the questions, the redesigns did substantially better (an average increase of 36.6%). Putting aside the questions on the first two displays where performance differences were not found, with the remaining three displays, the average performance on the redesigned displays over 18 questions was 15.7% better than the average performance on the original displays.

On 11 of the 30 questions response times were recorded. This information provides a rough measure of the ease with which respondents could extract information from the displays. A summary of the times are presented in Table 2. It is clear that respondents took about the same time or less time to answer questions with the redesigned displays, and those displays that resulted in the greatest improvement in accuracy (revised displays 3, 4, and 5) showed the largest decrease in time required to respond. For four of the 11 questions timed, respondents were faster by more than 10 seconds, on the average, in answering the questions when working with the revised displays.

Five additional questions were included in the interviews to assess the understandability of the original displays. From display 1, question 5A, we learned that only 50% of the respondents were able to understand the information about significance testing that was reported. From display 2, question 4B, we learned that only 50% of the respondents understood the table well enough to make a simple calculation. In display 3, respondent answers to questions 1A and 3A revealed substantial numbers of respondents understood the footnote about data interpretations but only 60% of them understood the meaning of the symbol addressing significant differences. Finally, only 15% of the respondents identified the standard error statistics reported in display 4. All five of the questions we asked about features of the original displays were reasonable and performance, with the exception of one question, was below 60%. Clearly, these displays were not communicating meaningful information in all instances.

Preference Results

At the end of questioning on each display, respondents were asked to comment on the display they had been working with and, at the end of the interview, as time permitted, respondents were presented with the alternative version of at least one of the displays they had seen and asked to compare the original and the redesigned version. A brief summary of these comments follows. A more extensive listing of respondent comments on the displays is available from the second author.

For the first set of displays (1-O and 1-R), respondents preferred the original version by a substantial margin (22 vs. 9). Apparently, bars are preferred to lines, although it must be mentioned that display 1-O appeared in color, and color may have contributed to the positive reaction from respondents. The preference for the original display was more apparent among those who worked with the original display than those who worked with the revision (16 to 3, vs. 6 to 6). Some respondents did indicate that they preferred to see the statistical test information on the display and therefore they preferred the original display.

For the second set of displays (2-O and 2-R), respondents had a slight preference for the revision (by a margin of 14 to 11). The preference was much stronger among those who worked with the revision (by a margin of 9 to 3). It seemed that the revision, at first appearance, looked complicated, but those who had the opportunity to work with it liked the format very much. A few did not like the revision because they were reluctant to view a direct linkage of race with parents' educational levels.

For the third set of displays (3-O and 3-R), there was a strong preference for the revised display (by a margin of 18 to 3). But ranking the states was not the only reason for preferring the revised display. Respondents did not like the small print, clutter, and footnote in the original display. A summary of the comments collected suggested that respondents liked the bigger print, additional white space, lack of a footnote, and the fact that the states were ranked in the revised display. Clearly, display 3-R provided more information and presented that information more clearly than the original.

For the fourth set of displays (4-O and 4-R), some unusual results were obtained. There was a slight preference for the revised display (by a vote of 12 to 11) but, if the respondents answered the questions using the revised display, they were more likely to prefer the original display (the vote was 8 to 1). Conversely, those using the original display preferred the revised display when they saw it (by a vote of 11 to 3). One conclusion might be that respondents disliked the revised display less than they disliked the original one!

Features considered to be improved in the revised display included (a) the top scale and labels for basic, proficient, and advanced categories, and (b) the explanation of the bars in the footnote. The footnotes in the original display were not described in glowing terms by many respondents. Two features that respondents tended to prefer in the original display were (a) the larger percentile range of the bars (10th to the 90th percentile), and (b) the state abbreviation labels in the bars.

For the fifth set of displays (5-O to 5-R), preferences were clear. The revised display was overwhelmingly preferred (by a vote of 19 to 0) and was thought to be

clearer and easier to understand. Respondents liked the ordering of categories from low to high, that the percentages added up to 100%, and they liked the fact that no footnotes were used. Getting rid of the cumulants was a revision all respondents preferred.

Discussion

In this study we have described a methodology that can be used to aid in the evolution of more effective data displays. The key element in our method is to assess how well the prospective audience can use the information in the displays. Although we asked the participants in our survey how well they liked each display, we were more interested in how accurately they were able to extract the sorts of information from the display that they typically would want. It is important and instructive to note that preference for a display and performance on it did not always go together. Hence we feel it is important that empirical testing of displays should not solely rely on user preference. Display choice is not a beauty contest.⁴

One important finding of the study is that respondents generally performed well on the set of questions used in the interviews. For 21 of the 30 questions that were common to the original and revised displays, performance levels exceeded 80% correct. Clearly, there was considerable merit and clarity to the original displays. On the other hand, the original display of the detailed presentation of state NAEP results (see display 4) caused many problems, though it might also be said that the revised display may have been oversimplified and could contribute to some misinterpretations about the ranking of states. This point highlights the trade-offs that often must be made between simplicity of presentation to enhance understandability and technical accuracy to minimize improper inferences.

There is one major disappointment uncovered in our experiment. Specifically, it is the apparent incomprehensibility of results presented as cumulants (display 5-O). From a statistical point of view, it seems sensible to present performance in ordered categories as cumulants, but empirically we found that the segment of the users of NAEP spanned by our sample broadly misunderstood these results. Thus we must conclude that such results either ought to be presented unaccumulated (as we did in the redesign, see display 5-R) or a serious attempt at educating the audience must be initiated. This latter possibility might be accomplished through the use of a sequence of displays that culminates with something that uses cumulants. As part of such a teaching sequence one might include some examples of inferences that can be made easily from cumulants that are much less easily done from the disaggregated display. But, from our results, it was clear that continuing with this sort of display as it is yields a misperception in what seems to be an unacceptably large proportion of the intended audience.

One explanation for the apparent paradoxical behavior of the subjects of our survey (not preferring a display that empirically is shown to allow more accurate answers to questions) that can be drawn from the comments of those interviewed deals with the issue of convention. Conventional displays carry with them the force of experience. For this reason it is hard for a new display format to beat an

incumbent. Some of the revisions we used were original, or at least had a number of original elements. Thus the disquieting effects of something new manifested themselves in the preference data we report, while, at the same time, the clearer depiction allowed them to answer the questions more often correctly. If this inference is correct, the solution is just to use the new format for a while until it becomes conventional. A little introductory discussion probably wouldn't hurt either.

Hambleton and Slater (1996) and our findings in this study highlight clearly the difficulties many educational policymakers have in using NAEP reports (at least those reports introduced up through 1994). One might speculate that these difficulties extend to many other displays of educational data as well. What seems called for is a greater commitment of time and resources to the design of data displays that can fully inform the intended audiences. These designs ought to be based on a theory of data display (Bertin, 1983; Cleveland, 1994; Tufte, 1983, 1990; Wainer, 1997c), and there needs to be recognition that, like test items, displays ought to be pilot-tested prior to their use. A pilot-test of displays with intended audiences, in which performance as well as preference is measured, can identify display shortcomings that can be corrected before those displays receive wider distribution.

Notes

² 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.

³ This paragraph is a pretty close paraphrasing of a description by Martin Gardner (1978, p. 23).

⁴ Question 7 associated with Display 4 provides a clear illustration of this point; 41 of the 51 who viewed the redesign were able to answer this correctly, whereas only 15 of the 51 who received the original design could do so. Yet even such a profound difference in performance did not yield a concomitant difference in preference.

Appendix

INTERVIEW PROTOCOL

Interviewer: _____ Date: _____

Interviewee: _____

Location: _____

Opening Remarks

Begin the interview with some introductory remarks, like those below:

My name is [xxxxxxxxx] and I am working with Professor Ronald Hambleton at the University of Massachusetts and Howard Wainer from the Educational Testing Service on a project for the United States Department of Education.

Our project is intended to determine the extent to which educational policy makers and educators understand the contents of reports being produced by the US Department of Education. The main question concerns whether or not these reports communicate national, regional, and state test results from the National Assessment of Educational Progress (NAEP) in an understandable way. [NAEP is a national testing program run by the Federal Government through ETS and Westat. National assessments are conducted every two years with several subjects included in the testing program each administration. Reading and mathematics are the most frequently assessed subjects. Only students in grades 4, 8, and 12 are tested. NAEP is intended to provide accurate information about the status of achievement on important outcomes of schooling, and to provide a basis for monitoring change over time.]

The US Department of Education is concerned that these reports describing the results may not be completely understandable. Problems could be due to the lack of knowledge and experience of the persons reading the reports or due to faults in reporting, or both problems could be present. The results of our interviews should be informative for the Government because the interviews will address the success of various displays of data in communicating understandable information.

In summary, be sure to address in your opening remarks:

1. Who we are and who is funding the project.
2. The purposes of our study. [To determine the success of various displays of data in communicating clear and understandable information about the NAEP test results. But note too that the findings will have implications for all types of reporting of educational test results.]

3. The reason the study is important for American education. [Little reason in continuing to spend millions of dollars and not be sure that the information is clear and understandable. And this point applies to reporting any educational test results. Why spend money for the testing program if the reports of the results are not clear and understandable?]
4. Thank participants for their valuable time and interest.

Now you can be specific about the tasks ahead:

Let interviewees know that their tasks are to look through several displays in the 1994 report of the grade 4, 8, and 12 national and state test results in reading, and to answer some questions. They will also be given the opportunity to comment on the displays they are looking at.

Let interviewees know that for some questions their response times will be collected but that they should not let the timing of their answers influence their approach to reading the displays and answering the questions.

Mention too that they will be given the set of questions to help them follow along with the interview.

Finally, if you are unsure of how to score an interviewee's answer, record the answer in as much detail as possible, and we will sort the scoring out at a later time in Amherst.

Background Information

Interviewer:

We need to obtain some background information from you because we need to be able to clearly describe participants in the study. We will not attach your name to individual answers in reporting of our results. We are interested only in a summary of the group information we collect but we would like to have names and addresses in case a follow-up to the interview is necessary.

Background Questions

1. Name: _____
2. Address: _____

3. Race (circle one): Black White Hispanic Asian Other

4. Sex (circle one): Male Female
5. Job Description: _____

6. Number of years working as a professional? _____
7. What is your level of interest in national and state student achievement results? (circle one)
- Answer: High Medium Low
8. What is your experience and/or knowledge about educational tests and statistics (e.g., college courses? other training?)? (circle one)
- Answer: None one course more than
 one course
9. Do you have knowledge of the National Assessment of Educational Progress, sometimes referred to as NAEP? (circle one)
- Answer: Yes No Unsure
10. Are you interested in receiving a copy of our final research report when it becomes available? (circle one)
- Answer: Yes No
- Then say: We are now ready to move to the first display.

[Pass over to interviewees Form A or B on a random basis.]

Mark the form which is being administered:

(circle one) Form A Form B

[Most questions are common to both Forms. Questions marked A or B should only be used with that form. The interviewees will only see the questions which are specific to their form on their copy of the questions. As a result, question numbers will not be consecutive on their copies.]

Let interviewees know that they can mark up their displays and question sheets. They will be their materials to keep.]

Display 1 (Figure 3, page 8: Average Reading Proficiency by Grade and by Region - NAEP 1992 and 1994", Alternative 2)

Please turn to Display 1 on page 1. Take a minute or two to read this display and then we will begin the questioning. (When the interviewee is ready, pass over the interviewee questions, and direct the person to page 1 for the questions on display 1.)

1. In 1994, which region of the country, had the lowest average reading proficiency at all three grade levels?
(Answer: Southeast.)

TIME: Min/sec to Min/sec

Correct	Incorrect	Comment :
---------	-----------	-----------

2. At the national level, what was the general trend in results between 1992 and 1994? (Answer: down in grades 4 and 12, about the same in grade 8.)

Correct	Partially	Incorrect
	Correct	(missed all three grades)
	(answered 1 or 2 grades correctly)	

Comment :

3. Rank the regions from best to worst in terms of average reading proficiency for grade 12 in 1994? (Answer: 12th grade: Central, Northeast and West are even, and then the Southeast.)

TIME: Min/sec _____ to Min/sec _____

Correct	Incorrect	Comment :
---------	-----------	-----------

4. Which region showed the greatest decline in performance for 4th graders from 1992 to 1994? (Answer: Northeast.)

Correct	Incorrect	Comment :
---------	-----------	-----------

- 5.A Can you conclude, from looking at the display, that fourth grade students from the Central region in 1994 are performing significantly better than fourth grade students from the Northeast region in 1994? (Answer: No. We want interviewees to focus on the significance testing which was done. A significance test on the comparison of interest was not reported. Therefore the conclusion is not justified.)

Correct Incorrect Comment:

6. What is your overall impression of the presentation of information in this display? In other words, is there anything about the display that you like, and/or anything you don't like?

Comments:

Display 2 (Tables 2 and 4, page 10 and 11: "Average Reading Proficiency by Race/Ethnicity", "Average Reading Proficiency by Parents' Education Level", respectively; alternative 3)

Now turn to Display 2 on page 2. Take a few moments to read this display and then we will continue on. [When they have finished, have them turn to page 2 in the question booklet.]

1. Does this display on results reported by Race/Ethnicity contain any information that is surprising or interesting to you?

(Answer: Will the reader notice that:

- a. the average reading proficiencies are higher for Asian and White students than for Blacks and Hispanics at all three grade levels, and higher than American Indians in grades 4 and 8?
- b. at grade 12, White students outperformed Asian students?
- c. no racial group showed an improvement from 1992 to 1994)?

- d. students of parents who graduated from college are the best performers at each grade level?
- e. students of parents who did not finish high school are the worst performers at each grade level?
- f. American Indians outperformed Blacks and Hispanics at each grade level?

Scoring: Keep notes on the observations interviewees make--several possibilities are coded above.

2. What is the average reading proficiency of eighth grade Asian students in 1994? (Answer: 273 or 274)

TIME: Min/sec _____ to Min/sec _____

Correct Incorrect Comment:

3. How different is this average score in reading proficiency for eighth grade Asian students, from the national average for all eighth grade students in 1994? (Answer: it is 13 or 14 points above the average.)

Correct Incorrect Comment:

- 4.B What is the average reading proficiency of 12th grade White students in 1992? (Answer: 298, take the 1994 average of 294, and add 4.)

Correct Incorrect Comment:

5. Now take a few minutes and answer question 5. When you have finished, let me have your answers.

Below are listed pairs of groups, the first of each pair appears in Column A and the second appears in Column B. Your task is to determine which group performed better on the 1994 reading assessment. In the space to the left of each pair, mark one of four choices:

- A if the group in Column A performed better
- B if the group in Column B performed better
- C if the two groups performed equally
- D if it cannot be determined which group performed better.

<u>Column A</u>	<u>Column B</u>
___ 1. 4th grade Whites	4th grade students with parents who are college graduates
___ 2. 4th grade Asians	8th grade Blacks
___ 3. 8th grade Asians	12 grade Blacks
___ 4. 12th grade Blacks	12th graders whose parents graduated high school
___ 5. 12th grade Hispanics	8th grade students with parents who completed college
___ 6. Blacks in Grade 12	Asians in Grade 8

(Scoring Key:

1. C
2. B
3. A
4. B
5. C
6. B)

6. What is your overall impression of the presentation of information in this display? In other words, is there anything about the display that you like, and/or anything you don't like?

Comment:

What, if anything, seems to be emphasized in this display?

Display 3 (Table 6, page 13: "Average Grade 4 Reading Proficiency NAEP Trial State Assessments in Reading Public Schools Only"; Table 6 alternative)

Please turn now to the next display on page 3. Take a few minutes to review this display and then we will discuss it.

- 1.A What is meant by the phrase found in the footnote at the bottom of the table, "Differences between two groups may be partially explained by other factors not included in this display?" (Answer: Many factors affect test results.)

Correct Incorrect Comment:

2. How different is New Mexico's average reading performance from its regional average (the West) in 1994? (Answer: New Mexico scored above the regional average in 1994 by 7 points: 212-205=7)

TIME: Min/sec _____ to Min/sec _____

Correct Incorrect Comment:

- 3.A In your own words, what do you think is meant by the "<< symbol" in the footnote? (Answer: This informs me that there is a significance difference between the results in 1992 and 1994 for a state--the difference almost certainly is not just due to chance or random differences.)

Correct Incorrect Comment:

4. Have national scores increased or decreased over the time reported in this display? (Answer: Generally there has been a decrease. Most of the means are down in 1994.)

Correct Incorrect Comment:

5. Identify the top three states in reading performance in 1994. (Answer: Maine, North Dakota, Wisconsin.)

TIME: Min/sec _____ to Min/sec _____

Correct Incorrect Comment:

6. What is the range of average 1994 reading scores among the states? You don't need to include the two "Other Jurisdictions" at the bottom of the display. (Answer: About 31 points.)

TIME: Min/sec _____ to Min/sec _____

Correct Incorrect Comment:

7. What is your overall impression of the presentation of information in this display? In other words, is there anything about the display that you like, and/or anything you don't like?

Comments:

Display 4 (Figure C1: "Distribution of Overall Reading Proficiency Organized by Average Proficiency for the 1994 Trial State Reading Assessment, Grade 4, Public Schools Only", alternative that explains the meaning of the interval between the 25th and 75th percentile)

Turn now to Display 4 on page 4. Study this display for a few minutes and then we will continue with the interview.

1. At the bottom of the graph is an explanation of the bar which is used to present the results for each state. Would you explain the meaning of the 25th and 75th percentiles, and the mean? (Answer: 25th percentile is the point on the score scale below which 25% of the students in a state scored. 75th percentile is the point on the score scale below which 75% of the students in a state scored. Mean or average provides an indication of where the "middle" student is in the state. If interviewee says "typical student" that is OK too in this context.)

Correct Incorrect Comment (please record any incorrect responses):

2. What is the range of scores of the middle 50 percent of the grade 4 students in New York? (Answer: about 55 points--185 to 240.)
- TIME: Min/sec_____ to Min/sec_____
- Correct Incorrect Comment:
3. What percentage of 4th grade students in Maine are performing Below Basic? (Answer: About 25%. You need to understand that the 25th percentile [low end of grey band] is on the 208 mark. A score of 208 is the lowest possible score for a student to be identified as Basic.)
- Correct Incorrect Comment:
4. Approximately what percentage of students in Texas are at or above Proficient? (Answer: about 25%.)
- Correct Incorrect Comment:
5. How you think the bands help you interpret differences among the states? (Answer: There are definitely differences across the states in average performance, but the bands show that there is also considerable variation in student performance within each state.)
- Correct Incorrect Comment:
6. Maine is the best performing state. In which states would Maine's 25th percentile student be about average? Name three states. (Answer: Arkansas, Maryland, or Alabama.)
- Correct Incorrect Comment:
7. In how many states was the average student performing in the Below Basic category? (Answer: about 10.)
- TIME: Min/sec_____ to Min/sec_____
- Correct Incorrect Comment:
- 8.B What are the numbers in parentheses ()? (Answer: standard errors)
- Correct Incorrect Comment:

9. What is your overall impression of the presentation of information in this display? In other words, is there anything about the display that you like, and/or anything you don't like?

Comments:

What confused you, if anything, about the display?

Display 5 (Table 10, page 21: "Reading Achievement Levels by Parents' Educational Level", Table 10 alternative.

Interviewer should say: Three performance standards on the NAEP score scale were established to distinguish Below Basic, Basic, Proficient, and Advanced students. In this fifth display, the results are reported using these four performance categories. The goal of these performance standards is to simplify the reporting of scores and to make the scores more meaningful. Take a few moments to read Display 5 on page 5 and then we will discuss it.

Look at the grade 4 reading results in 1994:

1. What percent of the students were Below Basic?
(Answer: 40)
- Correct Incorrect Comment:
2. What percent of the students performed at the Basic level? (Answer: 30)
- TIME: Min/sec _____ to Min/sec _____
- Correct Incorrect Comment:
3. What percent of the students scored Basic or above?
(Answer: 60)
- TIME: Min/sec _____ to Min/sec _____
- Correct Incorrect Comment:

4. What do you think is meant by the number 43 in the display? [Point to where this number is in the display--1994, grade 12 students reporting parents who graduated from college.]

(Answer: In 1994, 43% of the students tested in grade 12 reported having one or more parents who graduated from college.)

Correct Incorrect Comment: (what were the incorrect interpretations?)

5. In 1994, what percentage of 8th grade students with parents who did not finish high school were in the Basic performance category? (Answer: 36%)

Correct Incorrect Comment:

6. From looking at this display, how would you describe the change in performance in grade 12 reading achievement from 1992 to 1994? (Answer: There appears to be a small decline. For example, there is 5% more students performing Below Basic, and fewer students in the Basic and Proficient categories.)

TIME: Min/sec _____ to Min/sec _____

Correct Incorrect Comment:

7. From looking at the display, how would you describe the relationship between reading performance and parents' educational level for twelfth grade students in 1994? (Answer: There is a strong positive relationship between parents' educational level and reading proficiency.)

Correct Incorrect Comment:

8. What is your overall impression of the presentation of information in this display? In other words, is there anything about the display that you like, and/or anything you don't like?

Comments:

Concluding Section

At this point, pick one of the displays (perhaps one the person had had some difficulty with) and show the interviewee both displays of the same data. Give them time to look at both displays, and ask them to describe the things they like and don't like about each display. Record these remarks below. If time permits, repeat for a second display. This is an important part of the interview so try to allow time to get this task in.

Comments:

End by thanking the person for participating.

References

- Bertin, J. (1973). *Semiologie graphique* (2nd ed.). The Hague: Mouton-Gautier. (English translation done by William Berg & Howard Wainer and published as *Semiology of Graphics*, Madison, Wisconsin: University of Wisconsin Press, 1983.)
- Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.
- Gardner, M. (1978). On Charles Sanders Peirce: Philosopher and Gamesman. *Scientific American*, 240, 18-26.
- Hambleton, R. K., & Slater, S. C. (1995). Using performance standards to report national and state assessment data: Are the reports understandable and how can they be improved? In *Proceedings of the Joint Conference on Standard-Setting for Large-Scale Assessments* (pp. 325-343). Washington, DC: NCES.
- Hambleton, R. K., & Slater, S. C. (1996, April). *Are NAEP summary reports understandable to policymakers and educators?* Paper presented at the meeting of NCME, New York.
- Peirce, C. S. (1891). The architecture of theories. *The Monist*, 161-176.
- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.
- Wainer, H. (1980). A test of graphicacy in children. *Applied Psychological Measurement*, 4, 331-340.
- Wainer, H. (1984). How to display data badly. *The American Statistician*, 38, 137-147.
- Wainer, H. (1990a). Measuring graphicacy. *Chance*, 3, 52 & 58.
- Wainer, H. (1990b). Graphical visions from William Playfair to John Tukey. *Statistical Science*, 5, 340-346.
- Wainer, H. (1991a). Humble pie. *Chance*, 4(1), 52-53.
- Wainer, H. (1991b). Double Y-axis graphs. *Chance*, 4(2), 50-51.
- Wainer, H. (1991c). Integrating figures and text. *Chance*, 4(3), 58-60.
- Wainer, H. (1991d). Elegance, grace, impact and graphical displays. *Chance*, 4(4), 45-47.
- Wainer, H. (1992a). Sense-lining. *Chance*, 5(3-4), 71-72.

- Wainer, H. (1992b). Prologue—How the following article came to be. *Educational Researcher*, 21, 12-13.
- Wainer, H. (1992c). Understanding graphs and tables. *Educational Researcher*, 21, 14-23.
- Wainer, H. (1993a). Making readable overhead displays. *Chance*, 6(2), 46-49.
- Wainer, H. (1993b). Tabular presentation. *Chance*, 6(3), 52-56.
- Wainer, H. (1993c). Graphing multiple comparisons: Some comments on Tukey. *Journal of Computational and Graphical Statistics*, 2, 35-40.
- Wainer, H. (1993d). Graphical answers to scientific questions. *Chance*, 6(4), 48-50.
- Wainer, H. (1996a). Depicting error. *The American Statistician*, 50(2), 101-111.
- Wainer, H. (1996b). Using trilinear plots for NAEP data. *Journal of Educational Measurement*, 33, 41-55.
- Wainer, H. (1997a). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics*, 22, 1-30.
- Wainer, H. (1997b). Some multivariate displays in NAEP. *Psychological Methods*, 2, 34-63.
- Wainer, H. (1997c). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. New York: Copernicus Books.
- Williams, P. L., Reese, C. M., Campbell, J. R., Mazzeo, J., & Phillips, G. W. (1995). *NAEP 1994 reading: A first look*. Washington, DC: United States Department of Education.

Authors

- HOWARD WAINER is Principal Research Scientist, Educational Testing Service, Rosedale Road, Princeton, NJ 08541; hwainer@ets.org. *Degree*: PhD, Princeton University. *Specializations*: statistical graphics and psychometrics.
- RONALD K. HAMBLETON is Distinguished University Professor, School of Education, University of Massachusetts, Amherst, MA 01003-4140; rkh@educ.umass.edu. *Degrees*: BA, University of Waterloo, CANADA; MA, PhD, University of Toronto, CANADA. *Specializations*: item response theory, large-scale assessment issues and practices.
- KEVIN C. MEARA is Doctoral Candidate, Research and Evaluation Methods Program, University of Massachusetts, Amherst, MA 01003-4140; kmeara@educ.umass.edu. *Degree*: BA, Architecture and psychology, Washington University, St. Louis, MO. *Specialization*: educational measurement, standard-setting.