A Test of Graphicacy in Children

Howard Wainer
Bureau of Social Science Research

A test of graphicacy was developed, administered to third- through fifth- grade schoolchildren, and scored using the Rasch model with Gustafsson's conditional maximum likelihood estimation method. After removing children with scores at or below chance, the model fit well. It was found that of the four types of displays used (tables, line charts, bar charts, pie charts), the line chart was inferior to the others, which were all equal. There was some interaction between the kind of question asked and the display technique. Third-grade children were much poorer at reading graphs than fourth- or fifth-grade children, but the differences between these latter two groups were modest.

In 1786, with the publication of Playfair's *Commercial and Political Atlas*, statistical graphics were born. Reactions to Playfair's inventions were mixed, but it could not be denied that this new mode of presentation allowed the understanding of data quickly, with very little pretraining. Playfair suggested that the busy reader would do well to read the preliminary remarks about the charts and, in so doing, would have easy access to the rest of the information in the atlas. The ease with which Playfair's graphs were understood by widely different linguistic groups suggests that the ability to read graphs, graphicacy, is separate from literacy. That young children seem to be able to read graphs with almost no instruction supports the view that graphicacy is a more "basic skill." For example, graphs are often used to teach mathematical concepts rather than vice versa. A child can tell from a pie chart that one-third is larger than one-fourth long before he/she learns to make that distinction from the fractions themselves.

The author was interested in the extent to which children learn to use graphic displays and at what age this learning is more or less complete. In addition, the broad nature of the areas of application led to other questions: What kinds of questions can graphical displays answer efficaciously? What sorts of displays "work" best? For which sorts of questions? Do all children become "graphicate"?

There is a long, albeit disorganized, history of research that has attempted to answer these questions. An excellent review of empirical research in graphics has been prepared by MacDonald-Ross (1978), and the interested reader is referred to that source. The current study is an attempt to concatenate knowledge gained from earlier work with modern methodology in an attempt to develop a tool that can be used to further investigate graphicacy.

The questions that graphs can answer have been classified into three types by Bertin (1973):

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 3 Summer 1980 pp. 331–340

1. *Elementary.* The extraction of exact information, e.g., how much rain fell in February in San Francisco?
2. *Intermediate.* The detection of trends, e.g., during what season did the amount of rainfall decrease each month?
3. *Comprehensive.* The comparison of whole structures, e.g., which season has the most rain?

Method

Format of the Test

The format of the test is simple: Eight questions are asked. These questions are shown in Figure 1 and reflect the three question types just described. Of course, the choices differ, depending upon the data set displayed. There were four data sets used, derived from four different cities (Chicago, Boston, San Francisco, and Washington, DC). In addition, there were four display forms used (line chart, bar chart, pie chart, and table). The data from four different cities were used to eliminate the effects of differential data complexity, which, though of possible importance, were not of immediate interest in this study. The display form associated with each city was varied in a regular pattern (see Table 1) to yield four forms of the test. Thus, each form consisted of 32 items, with eight questions associated with each display type. Across the four forms, each display was paired with each city exactly once. The order of the cities was the same on all forms, so that the presentation order of the displays was balanced.

A sample of each display type is shown in Figure 2. These are generally familiar forms, with the possible exception of the "pie chart." This is not the traditional pie chart, in which the area associated with each segment of the pie varies as a function of the data; the radius of each segment remains constant so that the angle associated with that segment is the key variable. The display technique used in this study is a modification of the "Nightingale Petals" (named for its inventor, Florence Nightingale) and is useful

Figure 1
The Eight Questions Used with San Francisco Choices

All the questions on this page are about the rain in San Francisco, California. All the information needed to answer the questions is in the chart to the left.

SF1. How much does it rain in March?
a. 30 mm. b. 40 mm. c. 50 mm. d. 60 mm.
SF2. Which month has 25 mm. of rain?
a. March b. April c. October d. November
SF3. How many months have less than 40 mm. of rain?
a. 5 b. 6 c. 7 d. 8
SF4. Which season has the most rain?
a. Winter b. Spring c. Summer d. Fall
SF5. In which season does each month have less rain than the month before?
a. Winter b. Spring c. Summer d. Fall
SF6. Which season has more rain, the summer or the spring?
a. More in the summer b. More in the spring
SF7. How many months have more rain than the average month?
a. 4 b. 5 c. 6 d. 7
SF8. In San Francisco, as the weather gets warmer, there is generally more rain.
a. True b. False

Table 1
The Design of the Four Forms of the Graphicacy Test

	City			
	Washington	Boston	Chicago	San Francisco
Pie Chart	A	D	B	C
Table	B	C	A	D
Bar Chart	C	B	D	A
Line Graph	D	A	C	B

Note: Entry in table is the form on which the designated city-display pair appear.

for showing cyclic data. In this display the angles are constant and the radii vary as a function of the square root of the data.

The Test Scoring Model

There are a wide variety of inferences that the author of the present study would like to be able to make about the results of this test. To be able to do this, the measurement model must have certain properties, among them, interval properties of the ability estimates of the children taking the test (their graphicacy), interval properties of the item parameters, sample-free item calibration, and item independent person measurement.

As is well known (Suppes & Zinnes, 1963), the measurement properties arise from the model. A particular measurement model having the properties desired is used on a data set. If the model fits, the properties hold; if not, another model must be used. It was decided to use the most popular and simplest latent trait model—the two-parameter logistic or Rasch (1960) model. This model posits a logistic response function of two parameters: a person parameter (ability) and an item parameter (difficulty). It assumes that the slopes of all the item characteristic curves (ICCs) are equal. The interested reader is referred to Wright's work for a full description of the Rasch model (Wright, 1977; Wright & Panchapakesan, 1969; Wright & Stone, 1979).

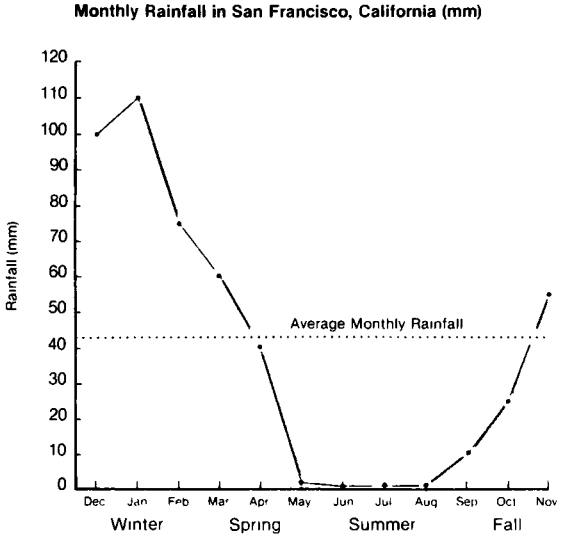
The estimation scheme used to determine these parameters is of some importance. It has long been known that the optimal method of estimation—for a variety of reasons, mostly for the desirable asymptotic properties of the estimators—is the method of conditioned maximum likelihood (Anderson, 1972). It is equally well known that this method was computationally impractical for tests of even modest length. Recent computational breakthroughs by Fischer (1974) and Gustafsson (1977) have made the use of this method practical for test lengths of 80 or more. The results from the present study were obtained with this new method.[1]

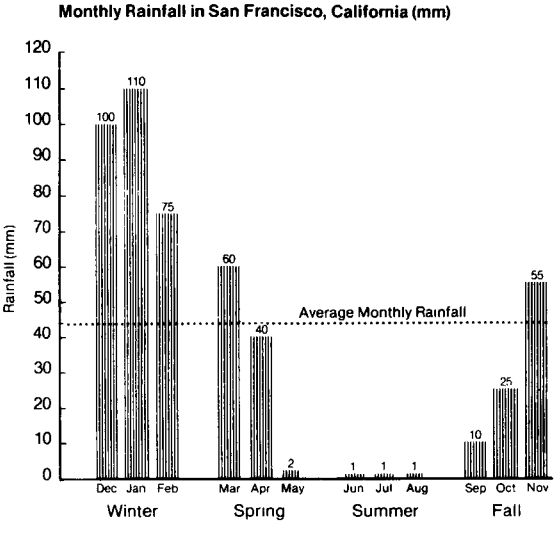
Procedure

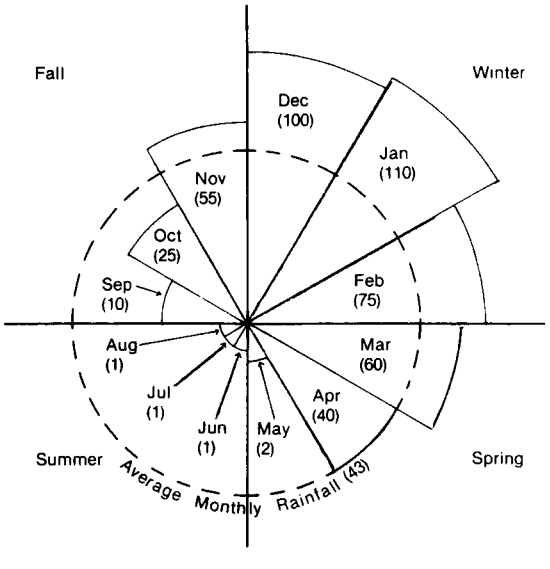
The test was administered to 360 children divided equally into three school grades (third, fourth, and fifth grade) and into the four forms specified earlier. This yielded 12 data sets with 30 children in each. The schools were in suburban Fairfax County, Virginia, and had the kinds of socioeconomic mixtures of children common to schools in such a heavily upper-middle class area. After fitting each form of the test to the Rasch model separately, it was found that the fit (as determined by Martin-Löf's, 1973, chi-square test) was generally good for most items but was overall only marginal on some forms due to inflation caused by a large contribution to the chi-square from some of the more difficult

[1]Using Gustafsson's program (PML2) on an Amdahl V/6 computer, the latent trait analyses performed (32 items and 75 subjects) ran in less than 1 second and cost less than $1.

Figure 2
Displays Used in the Graphicacy Test







Monthly Rainfall in San Francisco, California (mm)

Season	Months	Rainfall (mm.)
Winter	December	100
	January	110
	February	75
Spring	March	60
	April	40
	May	2
Summer	June	1
	July	1
	August	1
Fall	September	10
	October	25
	November	55
Average Monthly Rainfall		43

items. This result is characteristic of tests in which guessing occurs. All children whose raw scores were at or below chance were omitted. This resulted in trimming about one-third of the third graders but affected the fourth and fifth graders only slightly, as is evident from the data shown in Table 2.

The censored data were then fit with the Rasch model and rather good fits were obtained, with reliabilities generally over .80 (Kuder-Richardson 20). It must be remembered that all subsequent discussion deals with a censored data set, which means, practically, that statements about the graphicacy of third graders are overestimates, whereas for fourth and fifth graders the graphicacy estimates are probably about correct.

This method (censoring subjects whose scores were below chance) was used instead of trying to model this behavior because of the substantial evidence indicating that such models are not yet practical. The most likely alternative for such work is the three-parameter logistic model (Lord & Novick, 1968), which Lord (1968, p. 1015) says, "usually does not converge properly." In subsequent studies Lord (1975) and, more recently, Ree (1979) show clearly that none of the available algorithms for estimating either the lower asymptote or the slope of the ICC work satisfactorily. Thus, the author selected the method of censoring subjects whose scores were below chance, which has been strongly recommended for small-sample applications (Lord, 1979; Wright & Stone, 1979).

Table 2
Summary Statistics for the Graphicacy Test, After Deletion of Children with Chance Scores or Below

Test Form	Raw Score Mean			N			Sum	Rasch Fit			K-R
	3rd	4th	5th	3rd	4th	5th		χ^2	df	p	20
A	19	20	22	19	27	29	75	700	713	.6	.9
B	17	22	23	18	28	30	76	513	527	.6	.8
C	20	24	24	20	30	30	80	580	558	.2	.8
D	18	22	22	19	25	27	71	522	496	.2	.8
Means	19	22	23	Sums 76	110	116	302	2315	2244	.4	
Graphicacy Mean*	-.3	1.0	1.2		120	120	360				

*In LOGITS

Results

The students were randomly assigned to the various test forms. This supports the assumption that the mean ability on all forms is equal. This assumption was used to equate the scales for the four forms. The analysis design was conceived as a three-way Cities × Questions × Displays design. When this was done, the various effects in the design were calculated. Interest was in the question effects (the difficulty of the various kinds of

questions) and the display effects (e.g., which display techniques were the easiest to read? which were the hardest?). Interactions were also sought between question and display (e.g., are some displays better for some questions than others?). City effects, City × Question interactions, and City × Display interactions were viewed initially as contaminents and it was hoped that they did not appear. Of course, the general similarity of the four data sets in terms of their complexity pushed in the direction of no city effects or interactions.

An analysis of variance was performed. The only significant effects were questions and displays ($p<.01$). There was also a small Question × Display interaction, which subsequent contrasts showed to be significant.

Shown in Table 3 are the difficulties of the eight items and their average standard errors. Note that the easiest items were Elementary type questions, and the most difficult were of the Intermediate or Comprehensive types. The test is about three logits wide. Table 3 also shows the effects associated with the various kinds of displays. The similarity of three display types—the pie chart, table, and bar chart—and the greater difficulty of the line chart supports the display mode selected in this paper. The results obtained from the analyses of uncensored data show the same general structure, but with the sort of regression effect toward zero that is expected with the addition of a random contaminant.

The Displays × Questions interaction effects are also shown (Table 4), which indicate the difficulty of extracting Elementary information (Questions 1, 2, and 3) with a line graph, and the relative ease in seeing trends and making comparisons (Questions 5, 6, 7, and 8). This property of line graphs will be used in displaying the trends in graphicacy.

The line chart in Figure 3 shows graphicacy as a function of raw score, as well as the associated 95% confidence interval of these graphicacy estimates.

After arriving at the appropriate transformation to change raw score to graphicacy, *all* children's results were transformed into this logit metric. Thus, those children who were omitted in the segment of the study involved in item scal-

Table 3
Mean Difficulties for Various Aspects of the Graphicacy Test, in Logits for Both Censored and Uncensored Samples

	Censored	Uncensored	Standard Errors
Question			
1	-1.5	-1.1	.3
2	-1.6	-1.2	.3
3	0.6	0.5	.2
4	-0.2	-0.2	.2
5	1.8	1.6	.3
6	-0.9	-1.0	.3
7	1.6	1.4	.3
8	0.2	0.0	.2
Display			
Pie Chart	-0.4	-0.3	
Table	-0.4	-0.3	
Line Graph	1.2	0.9	
Bar Chart	-0.4	-0.3	

Table 4
Interaction Effects for Displays and Questions for Censored Data

	Question							
	1	2	3	4	5	6	7	8
Pie Chart	-.4	-.5	-.1	0	.3	.2	.3	.3
Table	-.6	-.3	-.2	.2	.3	.4	.2	0
Line Graph	.9	.7	.6	-.1	-.7	-.4	-.5	-.5
Bar Chart	.1	.2	-.3	0	.1	-.1	0	.2

ing because of low scores—and, hence, had a high likelihood of guessing—were reinserted into the analysis. The reason for this was that interest was in assessing the graphicacy of the children grade by grade. If these low-scoring children were omitted, the estimate of the third grade would be biased upward. By including them, the estimate of third graders' graphicacy is still too high (due to their spuriously high scores, even at this low level), but it is less biased than had they been omitted. Thus, the sample size for the following displays and analyses is 360 (120 per grade).

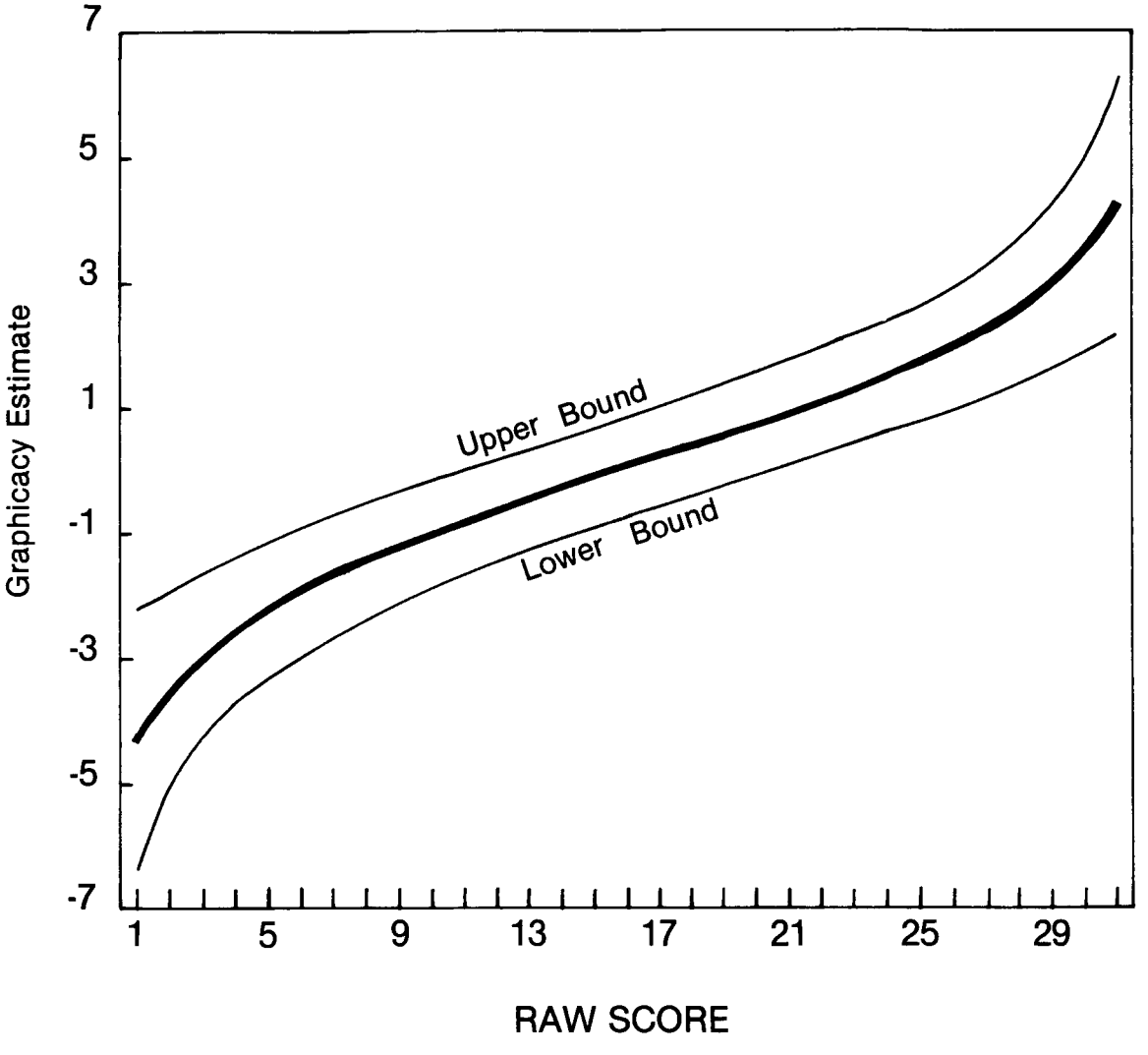
A Box-and-Whisker plot (Tukey, 1977) of the distributions of graphicacy for the three grades is shown in Figure 4. The *'s represent the extreme scores; the top and bottom of each box represent the 75th and 25th percentile respectively; and the middle bar represents the median. The box encloses the middle 50% of the children for each grade. As is evident, the class graphicacy increases substantially from third to fourth grade, and less so from fourth to fifth. An analysis of variance confirmed this observation, showing that the differences between grades was statistically significant (F=58, p<.0001). Two orthogonal contrasts performed subsequently indicate that the difference between third-grade performance and the average of fourth and fifth-grade performance is significant (p<.001) and that the difference between fourth and fifth grade is only marginal (.1< p <.05).

Discussion

The primary findings of this exploratory study into the measurement of graphicacy are the following:

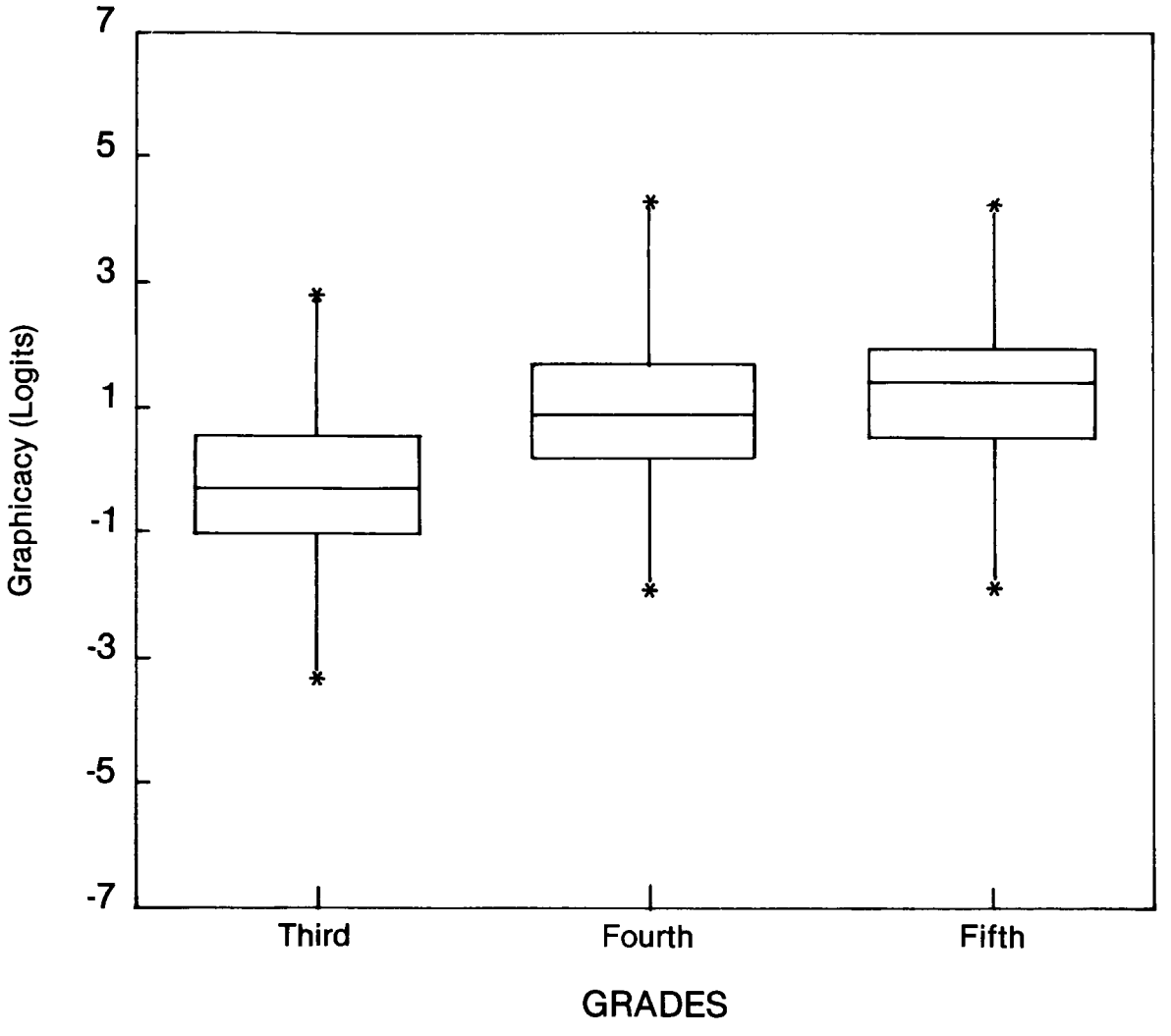
1. Elementary questions are more easily answered than Intermediate or Comprehensive questions.
2. Line charts appear to be more difficult than the three other forms, at least *as they are here represented.*
3. The interaction between display types and question types indicates that line charts are more difficult for answering Elementary questions, but relatively easier for answering Intermediate and Comprehensive questions.
4. The children's graphicacy, as measured by this test, seems to improve substantially from third to fourth grade, but little change was seen from fourth to fifth. This, combined with the generally high performance at or above fourth grade, indicates that there is little room for further improvement. The test was constructed to mirror accurately the complexity of graphics that a literate adult would be expected to deal with in a day-to-day existence. Thus, it would appear that these children's performance is high enough to consider that they have, on the average, reached a minimally acceptable level of adult graphicacy by fourth grade.

Figure 3
Graphicacy as a Function of Raw Score (with 95% Confidence Interval)



This study has reinforced some earlier findings (Wainer, Groves, & Lono, 1978), which showed that use of a table as a communication medium works rather well for data of only modest complexity. Bertin (1978), one of the staunchest proponents of graphic communication, has commented that for small data sets anything will work and that a rigorous test of a presentation scheme must use very complicated data. However, for most common data used in everyday communication, a tabular format (when constructed carefully) seems to work as well as anything else. This supports Ehrenberg's (1977) recommendations about the use of tables when the data are simple. The evidence gathered in this study are insufficient to discuss the efficacy

Figure 4
The Distribution of Graphicacy across Three Grades



of these competitive display techniques when the data are large and complex. It is surprising to find that even very young children seem to be able to read a variety of different kinds of charts with little difficulty and to answer questions quite accurately. There seems to be a real difference in graphicacy between the third graders and the older children. This may be a "stage" or, more probably, may merely reflect the differences in reading ability at this age. Regulations associated with the use of subjects made it impossible to obtain reading scores on these children to test this hypothesis. There are many questions unanswered, and a great deal of research on graphicacy remains to be performed. However, the graphicacy test has strong psychometric properties, and can be a useful tool in subsequent investigations.

References

Andersen, E. B. The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society Series B,* 1972, *34,* 42-54.

Bertin, J. *La Semiologie Graphique.* The Hague, Netherlands: Mouton-Gautier, 1973.

Bertin, J. Personal communication, October 1978.

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

Fischer, G. H. *Einfurung in die theorie psychologischer tests. Grundlagen und anwendungen.* Bern: Huber, 1974.

Gustafsson, J. E. *The Rasch model for dichotomous items: Theory, applications, and a computer program.* Goeteberg, Sweden: University of Goteborg, The Institute of Education, 1977.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement,* 1968, *28,* 989-1020.

Lord, F. M. *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (ETS RB-75-33). Princeton, NJ: Educational Testing Service, 1975.

Lord, F. M. Small *N* justifies Rasch methods. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

MacDonald-Ross, M. *Research in graphic communication* (IET Monograph No. 7). Milton Keynes, United Kingdom: The Open University, Institute of Educational Technology, 1978.

Martin-Löf, P. *Statistika modeller. Anteckningar fran seminarier lasaret 1969-1970 utarbetade av Rolf Sunberg* (2:a uppl.) Institut for forskakringsmatematik och matematisk statistik vid Stockholms Universitet, 1973.

Playfair, W. *The commercial and political atlas* (3rd ed.) London: Stockdale, 1801. (Originally published, 1786)

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Danmarks: Paedagogiske Institut, 1960.

Ree, M. J. Estimating item characteristic curves. *Applied Psychological Measurement,* 1979, *3,* 371-385.

Suppes, P., & Zinnes, J. L. Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1). New York: Wiley, 1963.

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

Wainer, H., Groves, C., & Lono, M. *Some experiments in graphical comprehension.* Paper presented at the annual meeting of the American Statistical Association, San Diego, 1978.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement,* 1977, *14,* 97-115.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 1969, *29,* 23-48.

Wright, B. D., & Stone, M. H. *Best test design.* Chicago: MESA Press, 1979.

Acknowledgment

This research was supported by a grant from the National Science Foundation (SOC 76-17768). I thank Fairfax County Public Schools for their participation and, in particular, R. W. Webb, A J. Ramsey, and W. Zepka. Susan Wise and Andrea Weckstein aided in the development and administration of the graphicacy test, and Douglas Neal prepared the graphs.

Author's Address

Send requests for reprints or further information to Howard Wainer, Bureau of Social Science Research, 1990 "M" Street, N.W., Washington, DC 20036.