NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

The Working Paper Series was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series.

U. S. Department of Education
Office of Educational Research and Improvement

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

STATISTICS FOR POLICYMAKERS or Everything You Wanted to Know About Statistics But Thought You Could Never Understand

Working Paper No. 97-21 June 1997

Contact: Susan W. Ahmed
Statistical Standards and Services Group
(202) 219-1781
e-mail: susan_ahmed@ed.gov

U. S. Department of Education
Office of Educational Research and Improvement

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Ramon C. Cortines
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

Statistical Standards and Services Group
Susan W. Ahmed
Chief Mathematical Statistician

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand*, Working Paper No. 97-21, by Susan Ahmed, Chief Statistician. Washington, D.C.: 1997.

June 1997

Foreword

Each year a large number of written documents are generated by NCES staff and individuals commissioned by NCES which provide preliminary analyses of survey results and address technical, methodological, and evaluation issues. Even though they are not formally published, these documents reflect a tremendous amount of unique expertise, knowledge, and experience.

The ***Working Paper Series*** was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series. Consequently, we encourage users of the series to consult the individual authors for citations.

To receive information about submitting manuscripts or obtaining copies of the series, please contact Ruth R. Harris at (202) 219-1831 or U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Ave., N.W., Room 400, Washington, D.C. 20208-5654.

Susan Ahmed
Chief Mathematical Statistician
Statistical Standards and
Services Group

Samuel S. Peng
Director
Methodology, Training, and
Service Program

This page intentionally left blank.

STATISTICS FOR POLICYMAKERS

or

Everything You Wanted to Know About Statistics But Thought You Could Never Understand

Prepared by:

Susan Ahmed
Chief Statistician
National Center for Education Statistics

June 1997

This page intentionally left blank.

Introduction

This working paper contains overheads used in a seminar developed by Susan Ahmed, NCES Chief Statistician. The seminar, titled "Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand," is designed to introduce some basic concepts of statistics to nonstatisticians. There are two main parts to the seminar. The first covers basic statistical concepts; the second covers some basic principles of research design and analysis.

Dr. Ahmed has presented the seminar to policymakers at the Department of Education, at an NCES Summer Data Conference, to newspaper reporters at the *Baltimore Sun*, to education writers at two Education Writers Association Annual Meetings, at the 1997 annual meeting of the National Commission of State Legislatures, and as the key note address at the 1997 meeting of state library data coordinators.

Essentials of Statistics and Analysis: An Overview

I. Essentials of Statistics

- A. Population, Sample, and Inference
- B. Standard Errors and Confidence Intervals
  - What are they and why are they important? How do you interpret them?
- C. Statistical Significance
  - What does it mean when a result is statistically significant?
  - What is the difference between statistical and substantive significance?
  - Can a result not be statistically significant and still be noteworthy? If a result is statistically significant, does it mean it's true?
- D. Correlation and Linear Regression
  - What are they? How do you interpret results based upon correlation or regression? Can you determine causality from cross-sectional data? From longitudinal data?
- E. Graphics
  - A discussion of how graphics can both mislead and enlighten the reader of statistical reports. Pitfalls in interpreting graphics.
  - The importance of skepticism.

II. Some Basic Principles of Research Design and Analysis

- A. Operationalizing Your Terms
- B. Selections Bias
- C. Need for Control Group
- D. Nonresponse Bias
- E. Confounding
- F. Validity
- G. Reliability
- H. Generalizing/External Validity

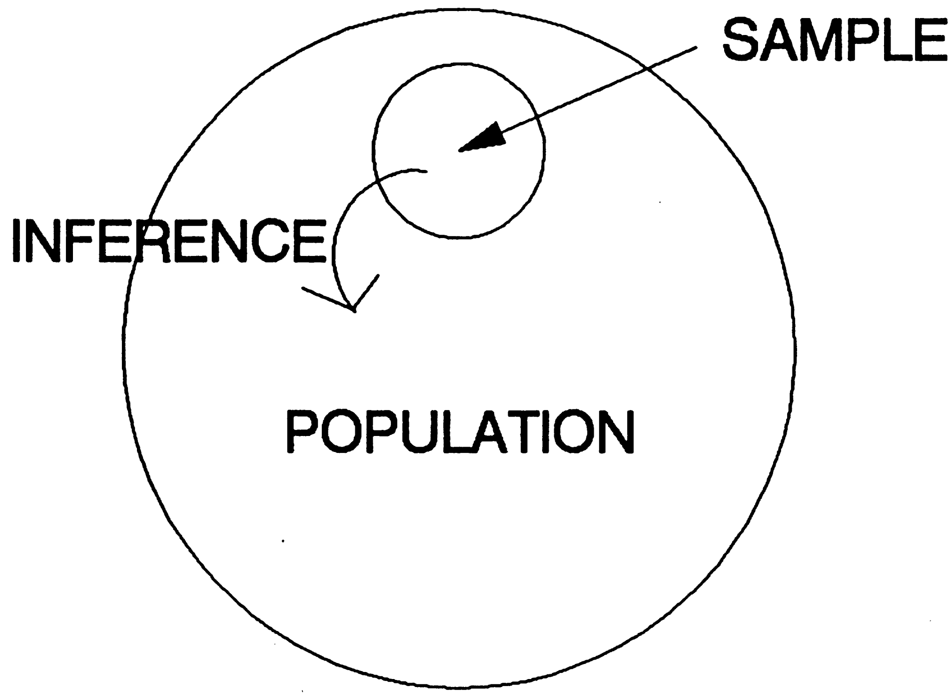
This page intentionally left blank.

I. ESSENTIALS OF STATISTICS

This page intentionally left blank.

EVALUATION OF CLAIMS MADE ABOUT DATA:
- WHEN TO BELIEVE THEM
- WHEN TO BE SKEPTICAL
- WHEN TO IGNORE THEM

POPULATION AND SAMPLE



POPULATION: the set of units about which we wish to draw an inference

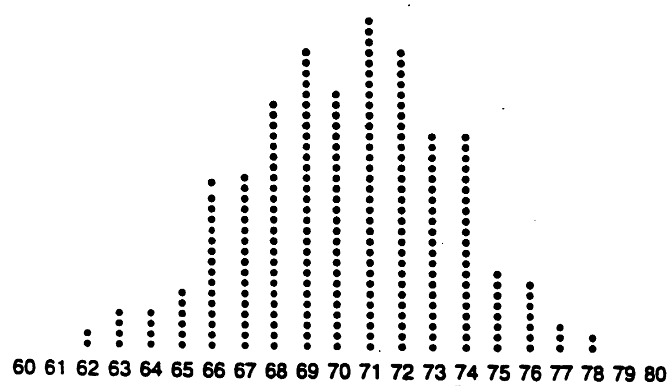
SAMPLE: a subset of a population

INFERENCE: a conclusion drawn about a population based on information from a sample

FREQUENCY DISTRIBUTION

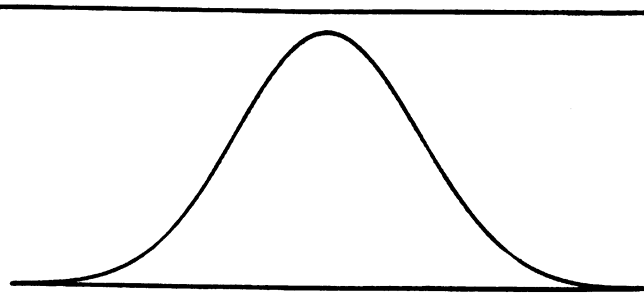
EXAMPLE: CLASS OF H.S. BOYS LINED UP FROM SHORTEST TO TALLEST

The raw material of a frequency distribution



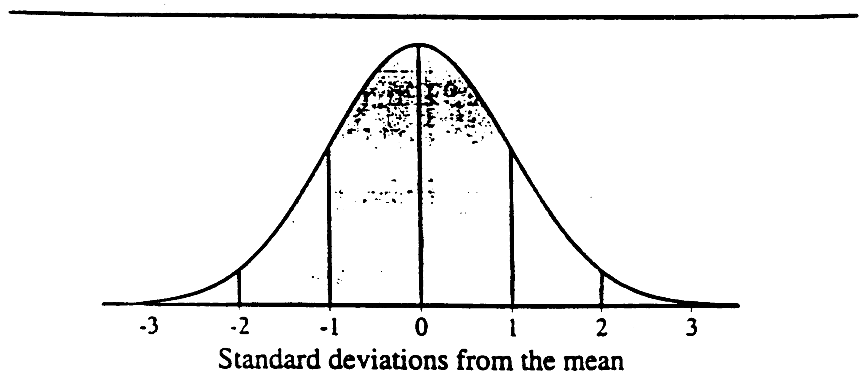
BELL CURVE/NORMAL DISTRIBUTION

A perfect bell curve



STANDARD DEVIATION: A MEASURE OF VARIABILITY. ALMOST LIKE AN AVERAGE DISTANCE FROM THE MEAN. ACTUALLY SQRT OF AVERAGE SQUARED DISTANCE FROM THE MEAN.

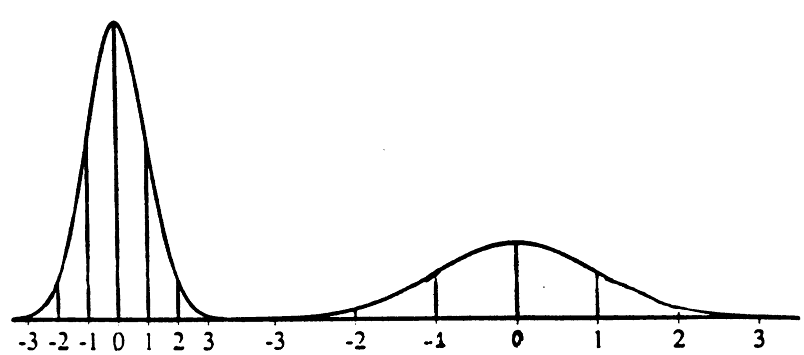
A bell curve cut into standard deviations



68% OF THE POPULATION LIES WITHIN +/- 1 STD DEV
95% OF THE POPULATION LIES WITHIN +/- 2 STD DEVS
99% OF THE POPULATION LIES WITHIN +/- 3 STD DEVS

COMPARING STANDARD DEVIATIONS:

Standard deviations cut off the same portions of the population for any normal distribution



E.G. HEIGHTS OF WOMEN GYMNASTS AND HEIGHTS OF BASKETBALL PLAYERS:

MEAN(WG) = 61" SD=2"
MEAN(BP) = 78" SD=4"

WHICH OF THE FOLLOWING IS MORE UNUSUAL?

A 66" WG OR A 84" BP?

WG = (66-61)/2 = 2.5 (2.5 SDs ABOVE MEAN)
BP = (84-78)/4 = 1.5 (1.5 SDs ABOVE MEAN)

THE WG IS MORE UNUSUAL THAN THE BP.

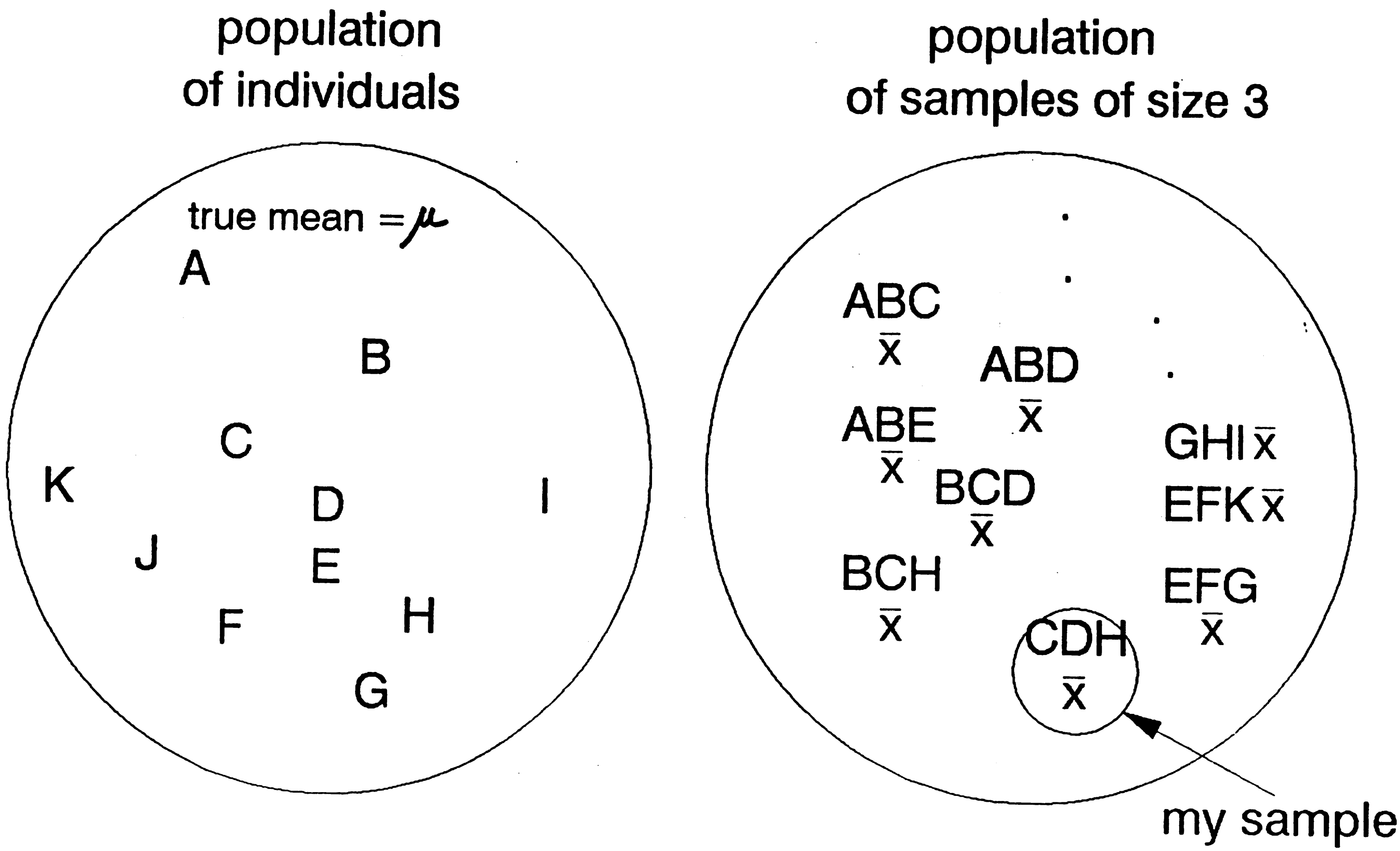
POTENTIAL CLAIMS

1. THE ONE YEAR ATTRITION RATE AMONG VOC ED TEACHERS IN PRIVATE SCHOOLS IN 1990-91 WAS 44%. THE RATE FOR ALL PRIVATE SCHOOL TEACHERS WAS 12%.

2. BLACK EIGHTH GRADERS AND WHITE EIGHTH GRADERS DIFFER IN MATH ACHIEVEMENT SCORES.

3. THERE IS A NEGATIVE ASSOCIATION BETWEEN TV WATCHING AND ACHIEVEMENT SCORES.

SAMPLING DISTRIBUTION/SAMPLING VARIABILITY



QUESTIONS TO BE ASKED BEFORE ACCEPTING AN ESTIMATE OR A CLAIM

1. SINCE THIS ESTIMATE IS BASED ON ONE SINGLE SAMPLE AMONG MANY THAT MIGHT HAVE BEEN DRAWN, AND KNOWING THAT DIFFERENT SAMPLES WOULD MOST LIKELY PRODUCE DIFFERENT ESTIMATES, HOW COMFORTABLE CAN I FEEL WITH THIS RESULT?

↓↓↓

HOW MUCH WOULD ESTIMATES FROM DIFFERENT SAMPLES VARY?

(STANDARD ERROR)

HOW CERTAIN CAN I BE ABOUT THIS ESTIMATE? WHAT IS THE MARGIN OF ERROR? HOW FAR OFF COULD I BE?

(CONFIDENCE INTERVALS)

2. IN MAKING A STATEMENT COMPARING TWO GROUPS OR ABOUT THE ASSOCIATION BETWEEN TWO VARIABLES, DOES THE EVIDENCE PROVIDED BY THE DATA SUPPORT THE STATEMENT?

↓↓↓

HOW DO WE PROVE OR DISPROVE A HYPOTHESIS REGARDING GROUP DIFFERENCES OR ASSOCIATIONS?

(HYPOTHESIS TESTING)

COULD THE DIFFERENCE OR THE ASSOCIATION WE ARE SEEING BE DUE TO CHANCE?

(STATISTICALLY SIGNIFICANT)

3. HOW CAN WE DISPLAY OUR RESULTS HONESTLY?

(MISLEADING GRAPHS)

QUESTION 1

SINCE THIS ESTIMATE IS BASED ON ONE SINGLE SAMPLE AMONG MANY THAT MIGHT HAVE BEEN DRAWN, AND KNOWING THAT DIFFERENT SAMPLES WOULD MOST LIKELY PRODUCE DIFFERENT ESTIMATES, HOW COMFORTABLE CAN I FEEL WITH THIS RESULT?

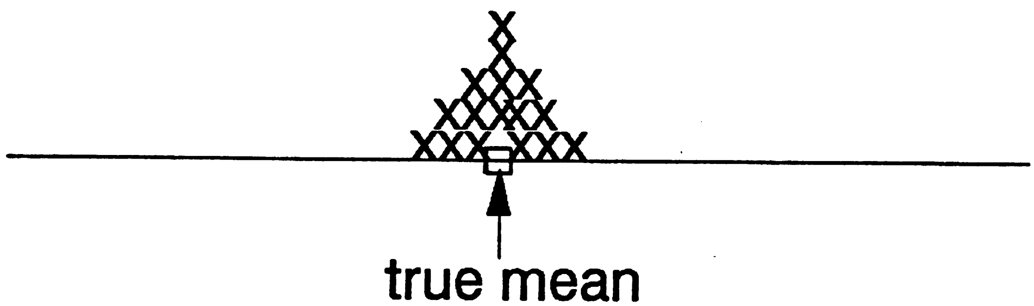
↓↓↓

QUESTION 1A

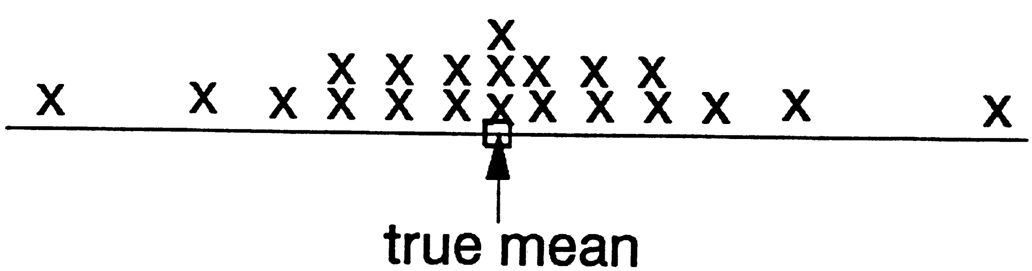
HOW MUCH WOULD ESTIMATES FROM DIFFERENT SAMPLES VARY?

(STANDARD ERROR)

STANDARD ERROR: MEASURE OF THE VARIABILITY OF A STATISTIC



small std error: all sample means are tightly grouped around true mean.



large standard error: sample means are widely spread around true mean.

95% of all sample means will lie within 2 std errors of the true mean

WHAT AFFECTS THE SIZE OF THE STANDARD ERROR?

The standard error is affected by

(1) the amount of variability of the measurement in the population
(2) the sample size

less variability → smaller std error
larger sample size → smaller std error

A1.--Standard errors for attrition rates from the teaching profession, by main field of assignment: 1987–88 to 1988–89 and 1990–91 to 1991–92 (table 1)

	Public		Private	
	1987–88	1990–91	1987–88	1990–91
Total	0.30	0.36	0.85	0.80
Kindergarten	0.69	1.56	2.65	2.74
General elementary	0.64	0.61	1.23	1.28
Art/music	0.79	1.44	4.38	3.26
Bilingual/ESL	3.11	2.04	--	--
Business	2.27	3.64	24.45	7.65
English/language arts	1.76	1.09	3.38	3.12
Health	0.81	0.85	2.99	4.37
Home economics	2.35	1.08	19.44	--
Industrial arts	1.27	0.87	--	--
Math	0.74	1.29	2.64	2.89
Reading	1.25	1.22	3.13	13.49
Social studies	1.73	1.22	2.86	3.66
Science total	1.21	1.96	2.25	2.08
Biology	0.94	1.17	5.05	3.55
Chemistry/physics	2.06	2.38	4.12	3.28
General science/earth science	2.09	3.71	3.75	3.05
Special education total	1.23	0.93	9.21	3.95
Mentally retarded	4.24	1.72	15.84	--
Learning disabled	0.65	0.92	10.34	2.57
Other special education	2.51	1.26	18.13	6.91
Vocational education	2.47	1.67	0.00	30.80
Foreign languages	++	0.44	++	3.69
All others*	0.78	1.01	3.64	3.03

--Too few cases for a reliable estimate.
*Includes computer science, remedial education, religion, gifted, prekindergarten, and all others (and foreign languages in 1987–88).
++Foreign languages in 1987–88 was included in the "All others" category.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Teacher Followup Survey, 1988–89 and 1991–92.

QUESTION 1B

HOW CERTAIN CAN I BE ABOUT THIS ESTIMATE? WHAT IS THE MARGIN OF ERROR? HOW FAR OFF COULD I BE?

(CONFIDENCE INTERVALS)

INTERPRETATION OF A CONFIDENCE INTERVAL

EXAMPLE: IN THE CONDITION OF EDUCATION, INDICATOR 13 PRESENTS THE FOLLOWING DATA FOR NAEP MATH SCORES FOR EIGHTH GRADERS:

BLACKS: MEAN=249 SE=2.3

A 95% CONFIDENCE INTERVAL IS AN INTERVAL CONSTRUCTED IN SUCH A WAY THAT YOU CAN BE 95% CONFIDENT THAT THE VALUE FOR THE WHOLE POPULATION FALLS IN THE INTERVAL.

A 95% CONFIDENCE INTERVAL WOULD BE CALCULATED AS FOLLOWS:

estimate +/- 1.96(se)

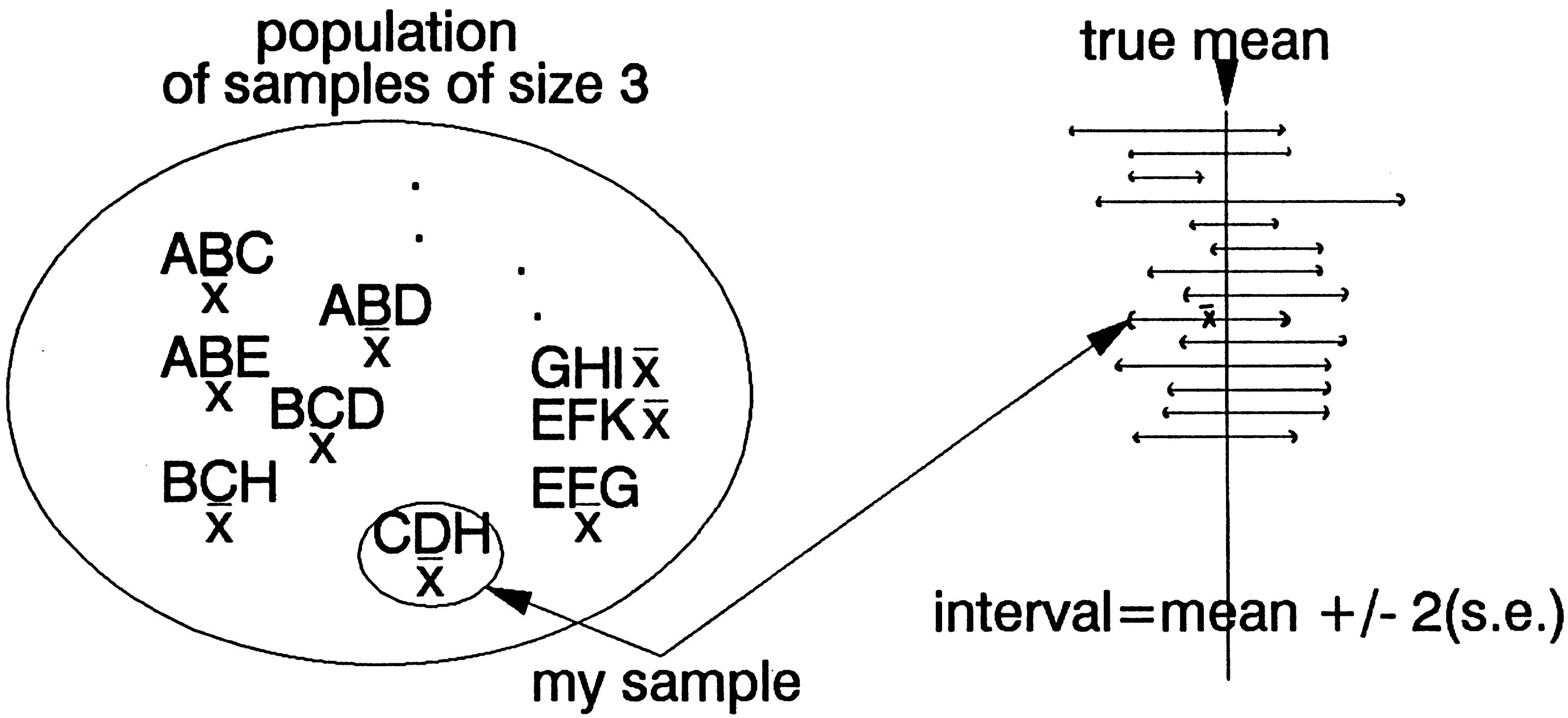
249 +/- (1.96)(2.3) = 249 +/- 4.5 ← margin of error

= (244.5, 253.5)

INTERPRETATION:

WE ARE 95% CONFIDENT THAT THE INTERVAL (244.5, 253.5) INCLUDES THE TRUE AVERAGE NAEP SCORE FOR ALL BLACK EIGHTH GRADERS.

WHAT DOES "95% CONFIDENT" MEAN?



WE ARE 95% CONFIDENT THAT OUR INTERVAL COVERS THE TRUE MEAN
because if we were to draw all possible samples of the same size and construct confidence intervals for each, then 95% of the intervals would include the true mean.

General Information

The information presented in this report was obtained from many sources, including federal and state agencies, private research organizations, and professional associations. The data were collected using many research methods including surveys of a universe (such as all school districts) or of a sample, compilations of administrative records, and statistical projections. Users of *The Condition of Education* should take particular care when comparing data from different sources. Differences in procedures, timing, phrasing of questions, interviewer training, and so forth mean that the results are not strictly comparable. Following the general discussion of data accuracy below, descriptions of the information sources and data collection methods are presented, grouped by sponsoring organization. More extensive documentation of procedures used in one survey than in another does not imply more problems with the data, only that more information is available.

Unless otherwise noted, all comparisons cited in the text were tested for significance using t-tests and are significant at the .05 level. However, when multiple comparisons are cited, a Bonferroni adjustment to the significance level was made. When other tests were used, they are described in a note on the indicator page or in the supplemental note for the indicator.

The accuracy of any statistic is determined by the joint effects of "sampling" and "nonsampling" errors. Estimates based on a sample will differ somewhat from the figures that would have been obtained if a complete census had been taken using the same survey instruments, instructions, and procedures. In addition to such sampling errors, all surveys, both universe and sample, are subject to design, reporting, and processing errors and errors due to nonresponse. To the extent possible, these nonsampling errors are kept to a minimum by methods built into the survey procedures. In general, however, the effects of nonsampling errors are more difficult to gauge than those produced by sampling variability.

The estimated standard error of a statistic is a measure of the variation due to sampling and can be used to examine the precision obtained in a particular sample. The sample estimate and an estimate of its standard error permit the construction of interval estimates with prescribed confidence that the interval includes the average result of all possible samples. If all possible samples were selected, each of these surveyed under essentially the same conditions, and an estimate and its standard error were calculated from each sample, then approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average value from all possible samples; 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average value of all possible samples; and 99 percent of all intervals from 2.5 standard errors below the estimate to 2.5 standard errors above the estimate would include the average value of all possible samples. These intervals are called 90 percent, 95 percent, and 99 percent confidence intervals, respectively.

To illustrate this further, consider the text table for indicator 1 and table 1-2 for estimates of standard errors from Census Current Population Surveys. For the 1991 estimate of the percentage of 3-year-olds enrolled in school (28.2 percent), supplemental table 1-2 shows a standard error of 1.2. Therefore, we can construct a 95 percent confidence interval from 30.6 to 25.8 ($28.2 \pm 2 \times 1.2$). If this procedure were followed for every possible sample, about 95 percent of the intervals would include the average for all possible samples.

Standard errors can help assess how valid a comparison between two estimates might be. The standard error of a difference between two sample estimates is approximately equal to the square root of the sum of the squared standard errors of the estimates. The standard error (se) of the difference between sample estimate "a"

CONF INTERVAL FOR PRIVATE SCHOOL TEACHERS

$12.3 \pm 1.96(0.80) = (10.7, 13.9)$

(12.3 ± 1.6)

Table 1.--Attrition rates from the teaching profession, by main field of assignment: 1987-88 to 1988-89 and 1990-91 to 1991-92

	Public		Private	
	1987-88	1990-91	1987-88	1990-91
Total	5.6	5.1	12.7	12.3
Kindergarten	3.1	4.0 [1]	10.5	11.9
General elementary	5.6	5.3	11.9	10.4
Art/music	4.2	5.9	17.7	13.0
Bilingual/ESL	8.2 [1]	4.5 [1]	--	--
Business	5.9 [1]	7.7 [1]	21.1 [2]	10.7 [2]
English/language arts	8.5	5.1	18.7	13.9
Health	3.8	3.3	6.3 [1]	15.6
Home economics	6.6 [1]	4.2	31.7 [2]	--
Industrial arts	3.7 [1]	2.7 [1]	--	--
Math	4.9	5.2	10.8	10.9
Reading	5.1	3.4 [1]	6.7 [1]	31.8 [1]
Social studies	5.1 [1]	6.7	8.4 [1]	10.8 [1]
Science total	5.4	6.1 [1]	9.2	7.3
Biology	3.2	3.7 [1]	8.5 [2]	6.6 [2]
Chemistry/physics	4.1 [1]	4.4 [2]	7.0 [2]	7.7 [1]
General science/earth science	7.1	8.0 [1]	10.9 [1]	7.5 [1]
Special education total	7.3	4.9	13.7 [2]	9.4 [1]
Mentally retarded	12.6 [1]	3.7 [1]	6.4 [2]	--
Learning disabled	4.3	3.2	7.6 [2]	3.4 [2]
Other special education	8.4 [1]	5.8	23.7 [2]	13.5 [2]
Vocational education	6.7 [1]	5.6 [1]	0.0	44.1 [2]
Foreign languages	++	2.3	++	14.1
All others [3]	5.2	4.8	18.2	19.0

--Too few cases for a reliable estimate.
++Foreign languages in 1987-88 was included in the "All others" category.
[1]Coefficient of variation between 30% and 50%.
[2]Coefficient of variation greater than 50%.
[3]Includes computer science, remedial education, religion, gifted, prekindergarten, and all others (and foreign languages in 1987-88).
NOTE: The attrition rate is the percentage of teachers who left the teaching profession between school years 1987-88 to 1988-89 and 1990-91 to 1991-92 (percent "leavers").

SOURCE: U.S. Department of Education, National Center for Education Statistics, Teacher Followup Survey, 1988-89 and 1991-92.

CONF INTERVAL FOR VOC ED:

$44.1 \pm 1.96(30.8) = 44.1 \pm 60.4$

$= (-16.3, 104.5)$

A1.--Standard errors for attrition rates from the teaching profession, by main field of assignment: 1987-88 to 1988-89 and 1990-91 to 1991-92 (table 1)

	Public		Private	
	1987-88	1990-91	1987-88	1990-91
Total	0.30	0.36	0.85	0.80
Kindergarten	0.69	1.56	2.65	2.74
General elementary	0.64	0.61	1.23	1.28
Art/music	0.79	1.44	4.38	3.26
Bilingual/ESL	3.11	2.04	--	--
Business	2.27	3.64	24.45	7.65
English/language arts	1.76	1.09	3.38	3.12
Health	0.81	0.85	2.99	4.37
Home economics	2.35	1.08	19.44	--
Industrial arts	1.27	0.87	--	--
Math	0.74	1.29	2.64	2.89
Reading	1.25	1.22	3.13	13.49
Social studies	1.73	1.22	2.86	3.66
Science total	1.21	1.96	2.25	2.08
Biology	0.94	1.17	5.05	3.55
Chemistry/physics	2.06	2.38	4.12	3.28
General science/earth science	2.09	3.71	3.75	3.05
Special education total	1.23	0.93	9.21	3.95
Mentally retarded	4.24	1.72	15.84	--
Learning disabled	0.65	0.92	10.34	2.57
Other special education	2.51	1.26	18.13	6.91
Vocational education	2.47	1.67	0.00	30.80
Foreign languages	++	0.44	++	3.69
All others*	0.78	1.01	3.64	3.03

--Too few cases for a reliable estimate.
*Includes computer science, remedial education, religion, gifted, prekindergarten, and all others (and foreign languages in 1987-88).
++Foreign languages in 1987-88 was included in the "All others" category.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Teacher Followup Survey, 1988-89 and 1991-92.

ır press releases when
newscasters have the
ıly the sample results.
ırveys have brought a
spapers or broadcast-
1980 and found that
cian, an advertiser, or
ɔoll without complete

ng error and which are
· answers.

the data.
to mail in a coupon

a sample survey. Label
rror and explain your

sampling frame.
three calls.
treet to interview.

data you look at. The
d States says that there
data list all crimes that
; and then reported by
The other main source
ey, a sample of 10,000
Bureau of Justice Statis-
een victims of a crime.
141,000 ± 6000 rapes in
ost side by side in the

many more rapes than

be attached to the FBI

to nonsampling errors.
f nonsampling error for
rime Survey?

report of a Gallup poll,
uary 1, 1990. At the end

90's, in Poll: A Good Life Amid Old Ills

MICHAEL R. KAGAY

As Americans look to the year 2000, most of them anticipate a better life for themselves, but at the same time they foresee a worsening of many of the nation's social and economic problems, according to a new Gallup Poll.

Seventy-seven percent of the 1,234 adults polled said they expected the overall quality of their own life to be better by 2000. Similarly, 77 percent anticipated that their family life would be better in 10 years' time. Seventy-four percent said their financial situation would be better. Eighty-two percent of employed adults also predicted their job situation would improve in 10 years.

Somewhat smaller majorities of Americans also anticipated that by 2000 people would be spending more time on leisure and recreation (68 percent), and more time with their families (58 percent). A minority said people would be spending more time on jobs (38 percent) or household chores (13 percent).

The poll, conducted by telephone Nov. 16–19 had a margin of sampling error of plus or minus four percentage points.

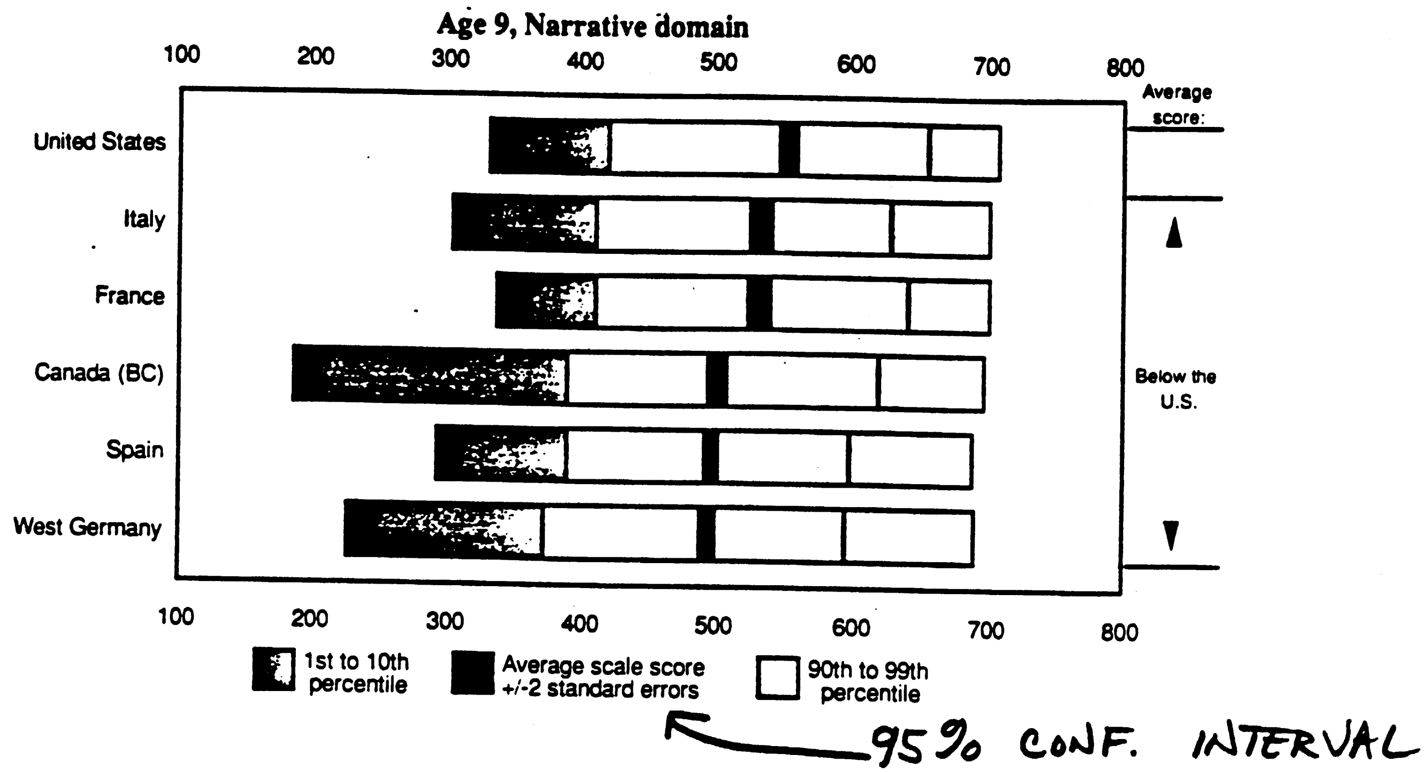
The participants' optimism about their own lives was accompanied by a more pessimistic outlook on many current social and economic problems. Large majorities expected by 2000 to see increases in the rate of inflation (74 percent), the crime rate (71 percent), poverty (67 percent), homelessness (62 percent), and environmental pollution (62 percent).

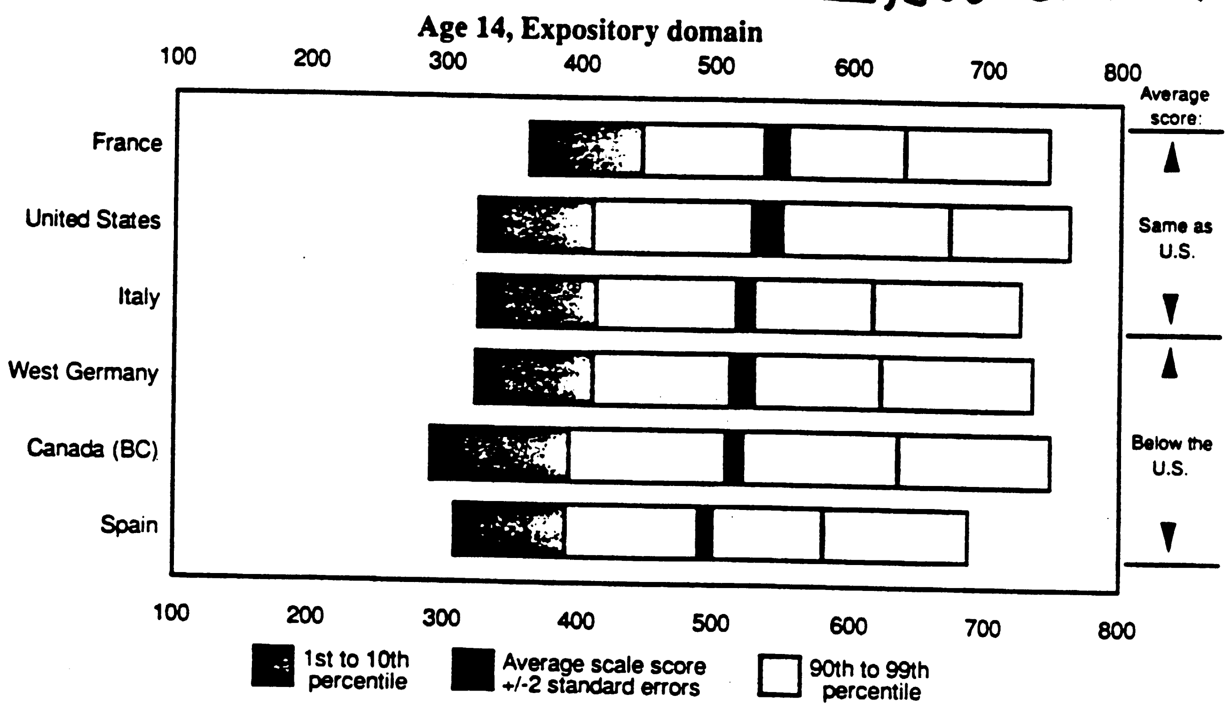
of Section 5 are some questions that should be answered by a careful account of a sample survey. Which of these questions does this newspaper report answer, and which not? Give the answers whenever the article contains them.

1.44. Market research is sometimes based on samples chosen from telephone directories and contacted by telephone. The sampling frame therefore omits households having unlisted numbers and those without phones.
   (a) What groups of people do you think will be underrepresented by such a sampling procedure?
   (b) How can households with unlisted numbers be included in the sample?
   (c) Can you think of any way to include in the sample households without telephones?

1.45. We have seen that the method of collecting the data can influence the accuracy of sample results. The following methods have been used to collect data on television viewing in a sample household:

Distribution of scale scores on reading literacy assessment, by age and country: School year 1991–92





NOTE: The vertical lines at ability score 500 marks the average score for each age group for all participating countries. The standard deviation is 100.

SOURCE: International Association for the Evaluation of Educational Achievement, Study of Reading Literacy, *How in the World Do Students Read?*, 1992.

QUESTION 2

IN MAKING A STATEMENT COMPARING TWO GROUPS OR ABOUT THE ASSOCIATION BETWEEN TWO VARIABLES, DOES THE EVIDENCE PROVIDED BY THE DATA SUPPORT THE STATEMENT?

↓↓↓

HOW DO WE PROVE OR DISPROVE A HYPOTHESIS REGARDING GROUP DIFFERENCES OR ASSOCIATIONS?

(HYPOTHESIS TESTING)

COULD THE DIFFERENCE OR THE ASSOCIATION WE ARE SEEING BE DUE TO CHANCE?

(STATISTICALLY SIGNIFICANT)

HYPOTHESIS TESTING

EXAMPLE

NULL HYPOTHESIS

H_0: THERE IS NO DIFFERENCE IN AVERAGE MATH ACHIEVEMENT SCORES OF BLACK AND WHITE EIGHTH GRADERS.

ALTERNATIVE HYPOTHESIS

H_A: THERE IS A DIFFERENCE IN THE AVERAGE MATH ACHIEVEMENT SCORES OF BLACK AND WHITE EIGHTH GRADERS.

TEST OF A HYPOTHESIS

AN INVESTIGATION OF THE CREDIBILITY OF A NULL HYPOTHESIS.

We collect some data on a sample and wish to see if these data are consistent with the null hypothesis.

EXAMPLE:

WHITES: MEAN=276
BLACKS: MEAN=249

Condition Indicator 13

HYPOTHESIS TESTING

EXAMPLE:

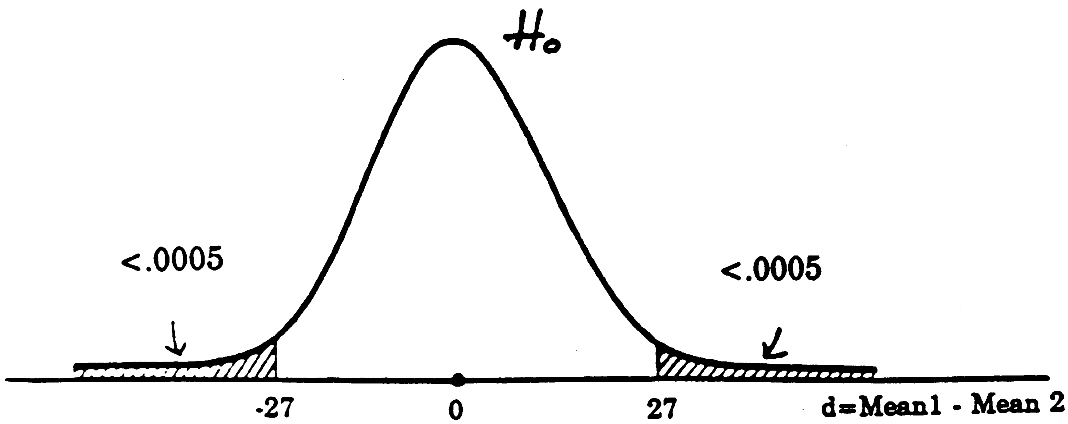
INDICATOR 13 : 1990 NAEP DATA
WHITES: MEAN=276 SE=1.1
BLACKS: MEAN=249 SE=2.3

Observed difference = 276-249 = 27

Are these data consistent with the null hypothesis?

How likely is it that we would get such a large difference if in fact the two population means were the same?

Is this difference real or due to chance?



The chances of getting such a large difference if the true means were the same is less than .001. This is the "p value".

p value: the probability of getting an outcome at least as extreme as what we actually got if H0 were true

If p is small, the evidence against the null hypothesis is strong.

HOW SMALL IS "SMALL"? THIS IS DECIDED BY THE SIGNIFICANCE LEVEL: α.

α = CHANCE YOU ARE WILLING TO TAKE YOU WILL REJECT THE NULL HYPOTHESIS WHEN IT IS REALLY TRUE.

IF p IS SMALLER THAN α, THEN WE SAY THE DIFFERENCE IS <u>STATISTICALLY SIGNIFICANT AT THE α% LEVEL.</u>

EXAMPLE: $p<.001$, $\alpha=.05$, $P<\alpha$.
REJECT THE NULL HYPOTHESIS.
CONCLUDE MEANS FOR BLACKS AND WHITES ARE DIFFERENT.

STEPS IN HYPOTHESIS TESTING

(1) SET UP THE NULL AND ALTERNATIVE HYPOTHESES.

The test is designed to assess the strength of the evidence against ho. Ha is a statement of the alternative we will accept if the evidence against h0 is sufficiently strong.

(2) CHOOSE THE SIGNIFICANCE LEVEL α.

This states the chance you are willing to take that you will reject the null hypothesis when it is really true. It is an indication of how much evidence against H0 will be decisive.

(3) FIND THE P VALUE FOR THE OBSERVED DATA.

This is the probability of getting a difference at least as extreme as what we got if the null hypothesis were true, i.e., the probability that the test statistic would weigh against H0 at least as strongly as it does for these data if H0 were in fact true.

(4) IF THE p VALUE IS LESS THAN α, REJECT THE NULL HYPOTHESIS. THE RESULT IS SAID TO BE STATISTICALLY SIGNIFICANT AT LEVEL α.

HYPOTHESIS TESTING

		TRUTH	
		Ho True (Not Different)	Ho False (Different)
CONCLUSION OF STATISTICAL TEST	Reject Ho (Different)	Incorrect Type I Error (Alpha)	Correct Power (1 - Beta)
	Do Not Reject Ho (Not Different)	Correct (1 - Alpha)	Incorrect Type II Error (Beta)

HAVING CARRIED OUT THE STATISTICAL TEST,

THE STATISTICIAN WILL TELL YOU THE RESULTS BY SAYING THAT THE RESULTS ARE OR ARE NOT

"STATISTICALLY SIGNIFICANT".

IF WE FIND:

(1) THAT THE DIFFERENCE IS <u>STATISTICALLY SIGNIFICANT,</u>

THIS MEANS THAT

- **the null hypothesis was rejected**
- **the data are not consistent with H0**
- **chance is not likely to have caused the difference we observed**

AND THUS OUR CONCLUSION ABOUT THE POPULATION IS THAT

- **"blacks and whites differ in avg math achievement".**

IF WE FIND:

(2) THAT THE DIFFERENCE IS <u>NOT STATISTICALLY SIGNIFICANT,</u>

THIS MEANS THAT

- **the null hypothesis was not rejected**
- **the data are not inconsistent with H0**
- **chance may have caused the difference we observed**

AND THUS OUR CONCLUSION ABOUT THE POPULATION IS THAT

- **"we do not have enough evidence to conclude that blacks and whites differ in avg math achievement".**

WHEN WE SAY THAT CHANCE IS NOT LIKELY TO HAVE CAUSED THE DIFFERENCE WE ARE SEEING, WHAT DOES "NOT LIKELY" MEAN? HOW UNLIKELY IS IT?

DETERMINED BY THE SIGNIFICANCE LEVEL α.

α IS THE PROBABILITY YOU WILL REJECT THE NULL HYPOTHESIS WHEN IT IS TRUE, I.E., THE CHANCE THAT YOU WILL CONCLUDE THE GROUPS ARE DIFFERENT WHEN THEY ARE NOT.

WHAT DOES A STATISTICAL TEST TELL YOU?

THE ONLY THING A STATISTICAL TEST TELLS YOU IS WHETHER CHANCE OR SAMPLING VARIABILITY IS LIKELY TO HAVE PRODUCED THE RESULTS YOU HAVE OBSERVED.

A STATISTICALLY SIGNIFICANT DIFFERENCE IS A DIFFERENCE WHICH IS TOO LARGE TO HAVE OCCURRED BY CHANCE ALONE.

STATISTICAL SIGNIFICANCE
vs
SUBSTANTIVE SIGNIFICANCE

HYPOTHESIS TESTING

ALL DIFFERENCES CITED IN NCES REPORTS HAVE BEEN SUBJECTED TO HYPOTHESIS TESTS AND ARE STATISTICALLY SIGNIFICANT UNLESS OTHERWISE NOTED.

Trends in the mathematics proficiency of 9-, 13-, and 17-year-olds

- **Overall, at ages 9 and 13, average mathematics proficiency improved somewhat between 1973 and 1990, but scores for 17-year-olds showed no improvement over the same period.**
- **Since 1973, white, black, and Hispanic 9-year-olds have shown improvement in average mathematics proficiency (10, 18, and 12 scale points, respectively). Most of this improvement occurred between 1982 and 1990.**
- **In 1990 large gaps existed between the mathematics proficiency of whites and their black and Hispanic peers. However, for blacks the gaps were narrower than they had been in 1973.**

Proficiency in mathematics is an important outcome of education. In an increasingly technological world, the mathematics skills of the nation's workers may be a crucial component of economic competitiveness. In addition, knowledge of mathematics is critical for success in science, computing, and a number of other related fields of study.

- **In 1990, large variability in average mathematics proficiency scores across states was found. A difference of 35 scale points existed between average eighth-grade students' performance in the highest and lowest scoring states (supplemental table 13-5).**

Average mathematics proficiency (scale score), by age and race/ethnicity: 1973–1990

Year	Age 9				Age 13				Age 17			
	All races	White	Black	Hispanic	All races	White	Black	Hispanic	All races	White	Black	Hispanic
1973	[1]219	[1]225	[1]190	[1]202	[1]266	274	[1]228	[1]239	304	310	[1]270	277
1978	[1]219	[1]224	[1]192	[1]203	[1]264	[1]272	[1]230	[1]238	[1]300	[2]306	[1]268	276
1982	[1]219	[1]224	[1]195	[1]204	269	274	[1,2]240	[2]252	[1,2]299	[1,2]304	[1]272	277
1986	[1]222	[1]227	[2]202	205	269	274	[2]249	[2]254	302	308	[1,2]279	283
1990	[2]230	[2]235	[2]208	[2]214	[2]270	276	[2]249	[2]255	305	310	[2]289	284

Average mathematics proficiency (scale score), by age and sex: 1973–1990

Year	Age 9		Age 13		Age 17	
	Male	Female	Male	Female	Male	Female
1973	[1]218	[1]220	[1]265	267	309	301
1978	[1]217	[1]220	[1]264	[1]265	[2]304	[1]297
1982	[1]217	[1]221	269	268	[1,2]302	[1,2]296
1986	[1,2]222	[1]222	[2]270	268	305	299
1990	[2]229	[2]230	[2]271	270	306	303

[1] Statistically significant difference from 1990.
[2] Statistically significant difference from 1973.
Note: Mathematics Proficiency Scale has a range from 0 to 500
Level 150: Simple arithmetic facts
Level 200: Beginning skills and understandings
Level 250: Numerical operations and beginning problem solving
Level 300: Moderately complex procedures and reasoning
Level 350: Multi-step problem solving and algebra

SOURCE: National Assessment of Educational Progress, *Trends in Academic Progress: Achievement of American Students in Science, 1969-70 to 1990, Mathematics, 1973 to 1990, Reading, 1971 to 1990, and Writing, 1984 to 1990*, 1991.

Table 7.1

Teachers' Reports on Amount of Time Spent Each Week Instructing and Helping Students with Writing, Grade 8, 1988 and 1992

In this class, about how much time do you spend each week on instructing and helping students with their writing?	*30 Minutes or Less*		*60 Minutes*		*90 Minutes*		*120 Minutes or more*	
	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency
Nation								
1992	15(1.6)	259(2.1)	40(2.0)	264(1.5)	22(2.0)	264(2.8)	23(2.3)	265(2.1)
1988	30(2.5)	--	42(2.2)	--	17(1.5)	--	11(0.6)	--
High Ability								
1992	9(3.1)	271(6.6)	36(4.7)	284(4.2)	29(4.8)	282(4.7)	26(4.2)	282(3.0)
Average Ability								
1992	15(2.4)	266(3.0)	45(3.1)	266(2.6)	20(2.6)	263(3.5)	20(2.4)	269(2.5)
Low Ability								
1992	21(3.5)	242(3.9)	36(3.4)	248(3.4)	21(3.5)	245(3.5)	23(4.8)	246(2.9)
Mixed Ability								
1992	14(3.2)	262(4.0)	38(4.1)	265(2.5)	23(3.8)	266(4.0)	26(5.0)	264(3.3)

The standard errors of the estimated percentages and proficiencies appear in parentheses. It can be said with 95 percent confidence for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference (see Appendix for details). --The 1992 Item Response Theory (IRT) scaling methods were not available in 1988 to calculate average writing proficiencies. Percentages may not total 100 percent due to rounding error.

SOURCE: National Assessment of Educational Progress (NAEP), 1988 and 1992 Writing Assessments.

Average writing proficiency did not differ significantly by amount of writing instruction. Also, teachers' reports on attention to writing instruction were relatively uniform across students in classes of different ability levels, though students in high-ability classes apparently spent more time on writing instruction than those in low-ability classes. For example, although this difference was not statistically significant, 91 versus 80 percent received an hour or more of instruction per week.

ASSOCIATION AND CORRELATION

ASSOCIATION:

The occurrence together of two or more characteristics or events more often than would be expected by chance.

CORRELATION:

A measure of the strength of association that assumes a linear relationship between the variables. The correlation coefficient, r, is a number between -1 and 1.

ısure of the
ınd y. The
ıoment r. It
−1 and +1.
ıits such as
ɔresents no
ıes to either
· it approxi-
:elationship
an inverse

(13.1)

(13.2)

ıber because
x.and y, is

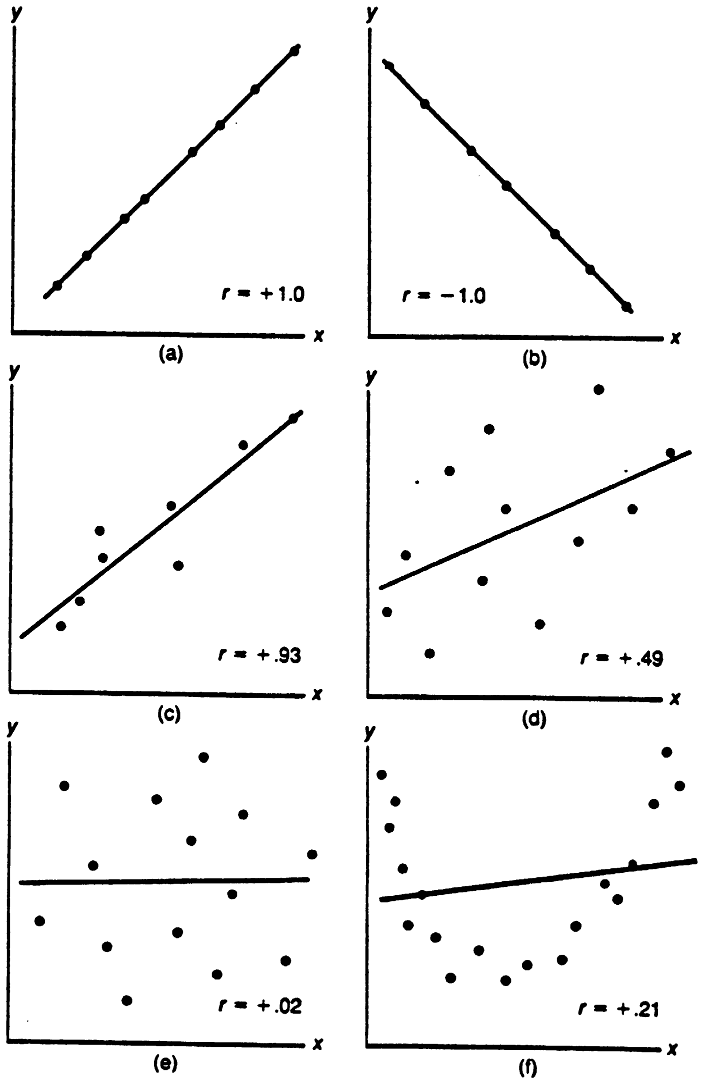
(13.3)

ıow they are
.0, a perfect
ıt line. It is
Figure 13.3b
ˌain fall on a

ɔbservations;
examples of
oint average,
lation would
e in drinking

remains high
ʿigure 13.3c).

Figure 13.3 Examples of Various Values of r



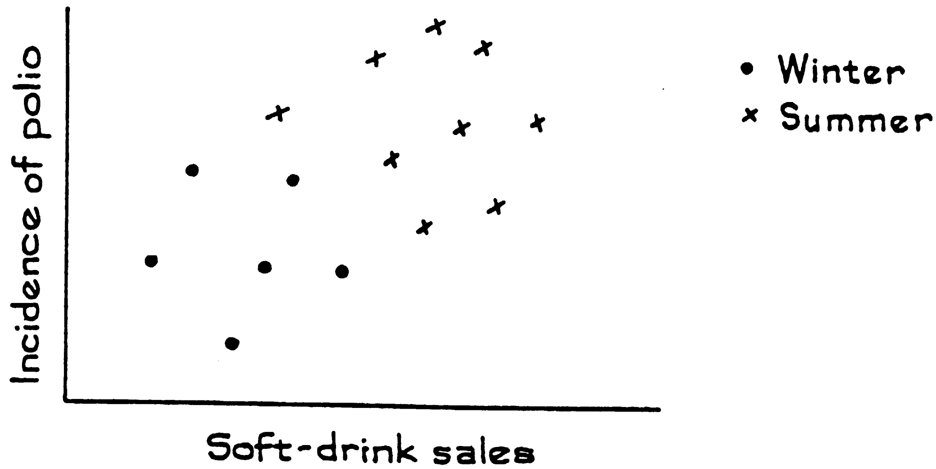
The coefficient becomes smaller and smaller as the distribution of points clusters less closely around the line (Figure 13.3d), and it becomes virtually zero (no correlation between the variables) when the distribution approximates a circle (Figure 13.3e). Figure 13.3f illustrates one drawback of the correlation coefficient: it is ineffective for measuring a relationship that is not linear. In this case we observe a neat curvilinear relationship whose

CORRELATION $\neq$ CAUSATION

EXAMPLES:

1. **POLIO AND SOFT DRINKS**
2. **STORKS AND BABIES**

Figure 1. A misleading correlation. Soft-drink sales are correlated with the incidence of polio.



> Correlation measures association. But association is not the same as causation.

Part I explained the difference between observational studies and controlled experiments. The same kind of distinction is useful here. In a laboratory experiment, the investigator usually varies the independent variable on his own initiative, and watches the effect on the dependent variable. For example, Robert Hooke (England, 1653–1703) was able to determine the relationship between the length of a spring and the load placed on it. He just hung weights of different sizes on the end of a spring, and watched what happened. When the load was increased, the spring got longer. When the load was reduced, the spring got shorter. In this experiment, weight was the independent variable: Hooke could vary that at will. Length was the dependent variable. Hooke did not choose its value, but watched how it responded to weight. Since the weight was under the direct control of the experimenter, there is no question here about what was causing what. The weight caused the spring to get longer.

Table 4.8.--Percentage of eighth graders who cite various probabilities for graduating from high school, by selected background characteristics

Student Characteristics	Probability of Completing High School			
	Very Sure Will Graduate	Will Probably Graduate	Probably Will Not Graduate	Very Sure Will Not Graduate
TOTAL	82.5	15.7	1.1	0.7
RACE				
Asian and Pacific Islanders	77.6	21.1	0.8	0.4
Hispanic	70.6	25.6	2.1	1.4
Black	81.5	16.6	1.2	0.7
White	85.0	13.6	0.9	0.6
American Indian and Native Alaskan	72.1	22.8	3.0	2.1
PARENTS' EDUCATION				
Did Not Finish High School	68.5	25.8	3.4	2.4
High School Graduate	80.3	17.7	1.2	0.9
High School Plus Some College	83.0	15.7	0.8	0.5
College Graduate	88.7	10.6	0.4	0.3
Graduate Degree	91.3	8.1	0.4	0.1
SES QUARTILE				
Lowest Quartile	71.8	24.0	2.5	1.7
25-49%	82.0	16.3	1.0	0.7
50-74%	85.1	14.1	0.6	0.3
Highest Quartile	91.1	8.4	0.3	0.2
FAMILY INCOME				
Less than $15,000	73.9	22.6	1.9	1.5
$15,000 - $50,000	83.8	14.8	0.8	0.6
Over $50,000	90.7	8.7	0.4	0.2
OLDER SIBLINGS WHO HAVE DROPPED OUT BEFORE GRADUATING				
None	84.7	14.0	0.8	0.5
One	71.9	23.9	2.2	2.1
Two	73.4	19.9	3.8	3.0
Three	69.6	27.2	3.2	0.0
Four	60.6	31.9	3.7	3.7
Five	68.1	25.8	3.8	2.3
Six or more	71.7	26.2	2.1	0.0
EVER REPEATED A GRADE				
Yes	71.2	24.4	2.6	1.7
No	86.4	12.6	0.6	0.4
DAYS OF SCHOOL MISSED IN PAST FOUR WEEKS				
None	86.2	13.0	0.5	0.4
1 or 2 days	84.8	14.2	0.7	0.4
3 or 4 days	77.4	19.6	2.0	1.1
5 to 10 days	74.7	21.3	2.6	1.5
More than 10 days	62.8	27.3	4.6	5.3
TIMES LATE FOR SCHOOL IN PAST FOUR WEEKS				
None	86.1	12.9	0.7	0.4
1 or 2 days	80.6	17.5	1.2	0.8
3 or 4 days	75.1	21.5	2.2	1.3
5 to 10 days	73.6	21.3	3.2	1.6
More than 10 days	64.1	27.3	3.1	5.4

SOURCE: U.S. Department of Education, National Center for Education Statistics, "National Education Longitudinal Study of 1988: Base Year Student Survey."

CORRELATION VS CAUSATION

CORRELATION: ARE TWO VARIABLES ASSOCIATED?

CAUSATION: WILL A CHANGE IN THE PREDICTOR ACTUALLY CHANGE THE OUTCOME?

CORRELATION DOES NOT IMPLY CAUSALITY!

ESTABLISHING A CAUSAL LINK:

1. SHOW THAT A CHANGE IN THE PREDICTOR PRODUCES A CHANGE IN THE OUTCOME.

2. SHOW THAT THERE IS NO PLAUSIBLE ALTERNATIVE EXPLANATION.

3. HAVE AN IDEA ABOUT WHAT MECHANISM IS AT WORK.

4. REPLICATE THE STUDY IN DIFFERENT POPULATIONS AT DIFFERENT TIMES.

5. STRENGTH OF ASSOCIATION.

6. DID THE PREDICTOR PRECEDE THE OUTCOME?

EXPLANATIONS WHEN AN ASSOCIATION BETWEEN TV WATCHING AND PERFORMANCE IS OBSERVED

Explanation	Type of Association	Basis for Association	What's really going on in the population?
Chance	Spurious	Random error	TV watching and performance are not related
Bias	Spurious	Systematic error	TV watching and performance are not related
Effect–cause	Real	Cart before the horse	Poor performance is a cause of excess TV watching
Effect–effect	Real	Confounding	Poor performance and excessive TV watching are both caused by a third extrinsic factor.
Cause–effect	Real	Cause and effect	Excess TV watching is a cause of poor performance

REGRESSION

- A STATISTICAL TECHNIQUE THAT IS USEFUL FOR STUDYING THE LINEAR ASSOCIATION BETWEEN A DEPENDENT VARIABLE AND ONE OR MORE INDEPENDENT VARIABLES.
- REGRESSION CAN
  - MEASURE THE DEGREE OF ASSOCIATION
  - MEASURE THE STATISTICAL SIGNIFICANCE OF AN ASSOCIATION
  - MEASURE THE EXTENT TO WHICH THE ASSOCIATION EXPLAINS THE VARIATION IN THE OUTCOME
  - SERVE AS A BASIS FOR PREDICTION
  - ASSESS THE RELATIVE IMPORTANCE OF SEVERAL PREDICTORS
  - ASSESS THE EFFECT OF ONE PREDICTOR, CONTROLLING FOR OTHERS

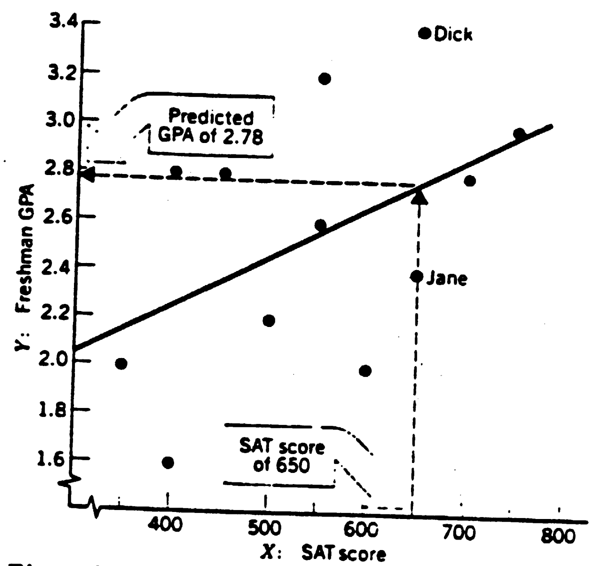


Figure 8.1 The prediction of freshman GPA at Alpha College from SAT scores.

of predictive error. In the next three sections we will examine what is meant by the line of "best fit" and learn how to use the formula for that line in making predictions. Then, we will learn how to attach a margin of error to our predictions. Finally, we will discover that our newly acquired knowledge provides another basis for interpreting the correlation coefficient.

8.2 THE LINE OF BEST FIT

It is all very well to speak of finding the straight line of best fit to the data, but how is one to know when the "best fit" has been achieved? Indeed, "best fit" could be defined in several ways. Let's look at the way that applies when we use Pearson r as the index of association and when our purpose is prediction.

We will let Y represent the actual score value of the variable to be predicted and Y' represent its corresponding predicted value (X will continue to represent the predictor variable). Then, an error of prediction is the discrepancy between the actual and predicted values:

$$\text{error} = (Y - Y')$$

Crime in the schools

> Research on effective schools has identified a safe and orderly environment as a prerequisite for promoting student academic success. Lack of school safety can reduce school effectiveness, inhibit student learning, and place students who are already at risk for school failure for other reasons in further jeopardy. In recent years, educators and policymakers have voiced growing concern about possible increases in the incidences of school-related criminal behavior.

- **Between 1976 and 1991, blacks were both more likely to be theatened with and more likely to be injured with a weapon in school than whites. In 1991, for example, about 1 in 10 black and about 1 in 19 white high school seniors reported being injured with a weapon at school. However, there were few other differences in the in-school victimization rates of black and white high school seniors over this period.**
- **For blacks, in most crime categories, there was little increase in the victimization rate between 1976 and 1991, except for something being stolen. In most crime categories whites did experience some increase in victimization.**
- **In 1991, of those high school seniors reporting being victimized, the most frequently reported type of victimization was having had something stolen (approximately 4 in 10). The least frequently reported type of victimization was having been injured with a weapon (nearly 1 in 19). About 1 in 4 reported that their property had been deliberately damaged or that they were threatened without a weapon.**

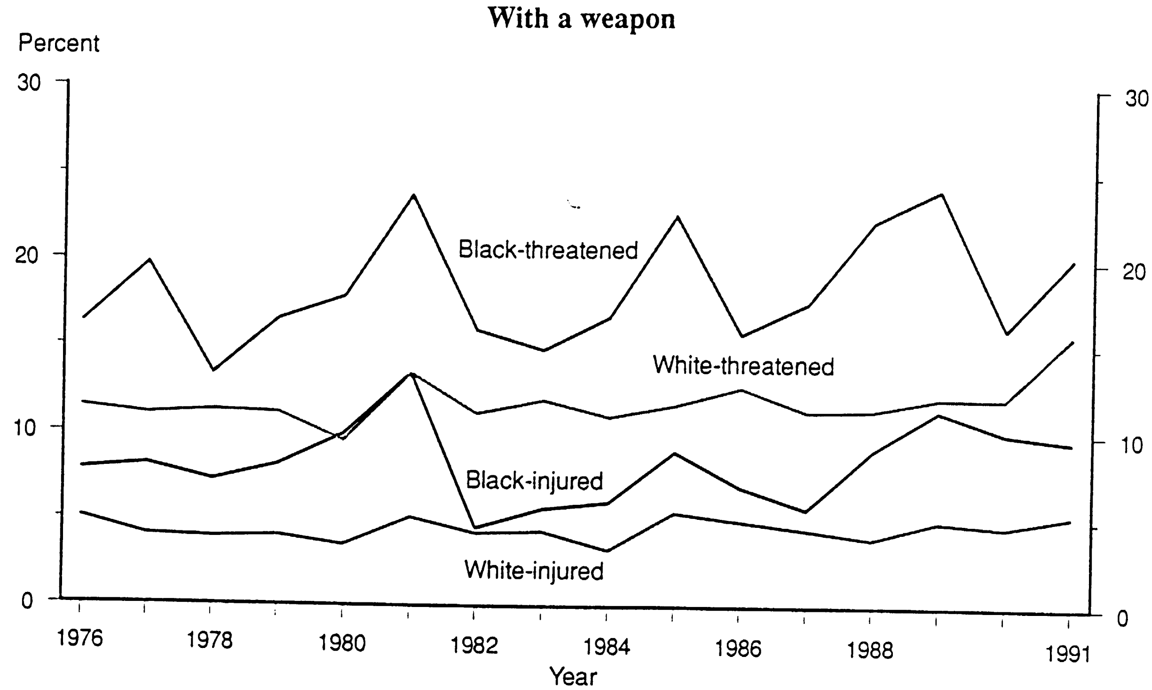
Percentage of high school seniors reporting being victimized in school, by type of victimization, and by race/ethnicity: 1976–1991

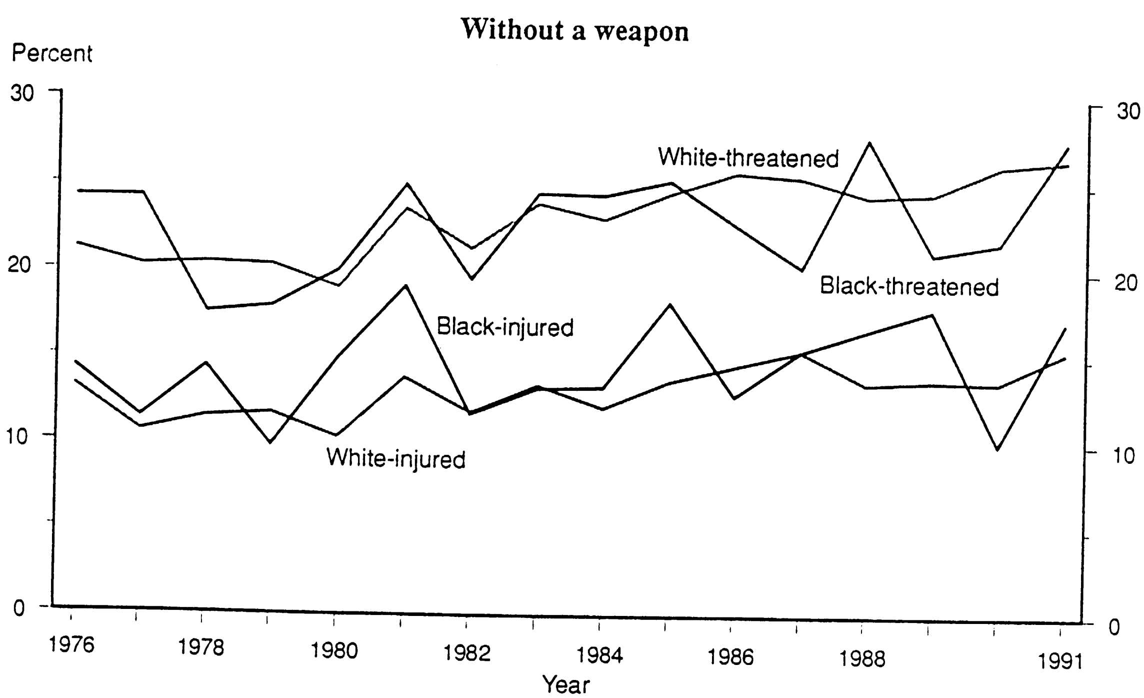
	Something stolen from you[1]		Property deliberately damaged		Injured you with a weapon[2]		Threatened you with a weapon[2]		Injured you without a weapon[2]		Threatened you without a weapon[2]	
Year	White	Black	White	Black	White	Black	White	Black	White	Black	White	Black
1976	38.9	35.9	25.1	30.1	5.0	7.8	11.4	16.3	13.2	14.3	21.2	24.2
1977	40.4	32.8	24.3	21.0	4.0	8.1	11.0	19.7	10.6	11.4	20.2	24.2
1978	38.8	32.4	25.7	21.2	3.9	7.2	11.2	13.3	11.5	14.4	20.4	17.5
1979	34.6	27.2	24.5	20.8	4.0	8.1	11.1	16.5	11.7	9.8	20.3	17.9
1980	34.3	33.1	25.3	21.9	3.5	9.9	9.5	17.8	10.3	14.9	19.0	20.0
1981	40.1	39.2	30.4	29.8	5.1	13.4	13.4	23.7	13.8	19.1	23.6	25.0
1982	37.9	42.0	25.6	25.4	4.2	4.5	11.1	15.9	11.8	11.7	21.3	19.5
1983	39.4	39.2	25.0	23.1	4.3	5.6	11.9	14.8	13.4	13.2	23.9	24.5
1984	38.4	35.3	24.3	21.8	3.2	6.0	10.9	16.7	12.1	13.3	23.0	24.4
1985	39.3	35.2	26.6	28.0	5.4	8.9	11.6	22.6	13.6	18.2	24.5	25.2
1986	41.1	36.3	25.7	24.5	4.9	6.9	12.6	15.7	14.5	12.8	25.7	22.7
1987	42.1	39.4	27.0	25.0	4.4	5.6	11.2	17.5	15.4	15.4	25.4	20.2
1988	41.4	46.6	27.4	25.8	3.9	9.0	11.3	22.2	13.5	16.6	24.3	27.7
1989	39.4	46.4	26.0	28.9	4.9	11.3	12.0	24.1	13.7	17.8	24.5	21.0
1990	41.6	42.2	28.9	26.1	4.6	10.0	12.0	16.0	13.6	10.0	26.1	21.7
1991	41.4	44.3	28.4	24.6	5.3	9.6	15.7	20.2	15.4	17.1	26.5	27.5

[1] The response category "something stolen from you" is comprised of two separate questions: 1) "Has something of yours (worth under $50) been stolen?", and 2) "Has something of yours (worth over $50) been stolen?" The responses to both questions have been collapsed in this category.
[2] The weapons category includes: knife, gun or club. The question was: "Has someone injured you with (or without) a weapon (like a Knife, Gun or Club)?"
NOTE: A regression analysis was used to determine trends over time between the races. Therefore, individual year differences between the races might be statistically different, while the trend over time is not.

SOURCE: University of Michigan, Survey Research Center, Institute for Social Research, *Monitoring the Future*, unpublished tabulations..

Percentage of high school seniors reporting being victimized in school, by race/ethnicity: 1976–1991





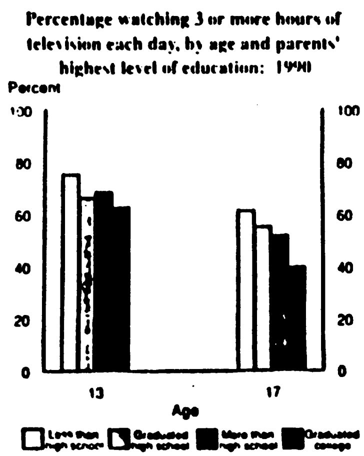
SOURCE: University of Michigan, Survey Research Center, Institute for Social Research, *Monitoring the Future*, unpublished tabulations.

QUESTION 3

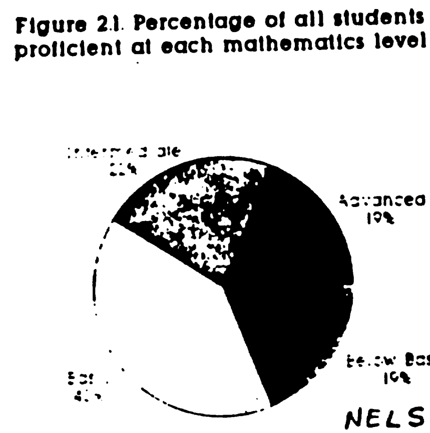
HOW CAN WE DISPLAY OUR RESULTS HONESTLY?

(MISLEADING GRAPHS)

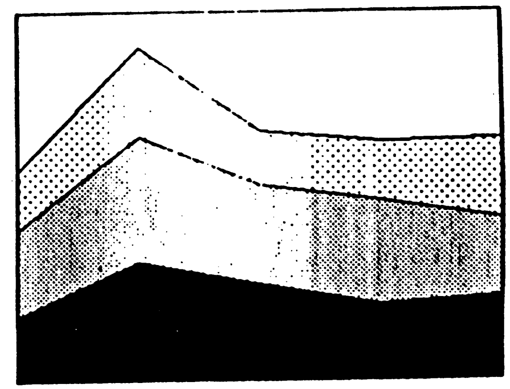
Types of Graphs



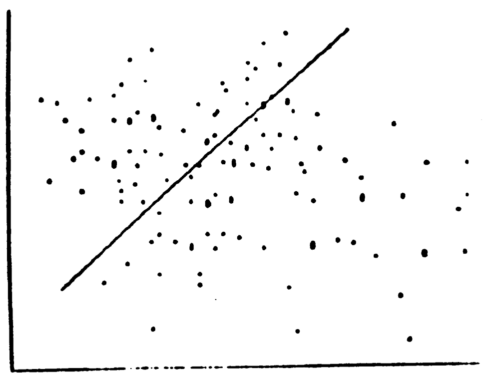
Bar Graph



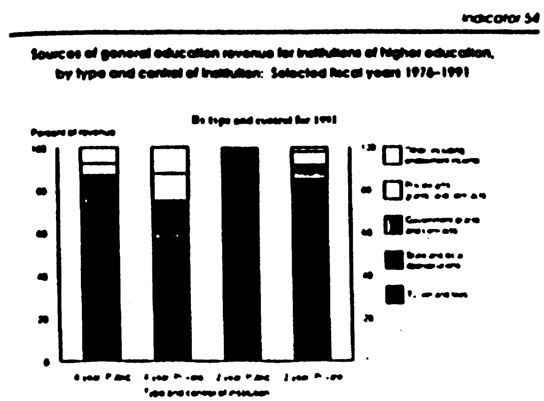
Pie Chart



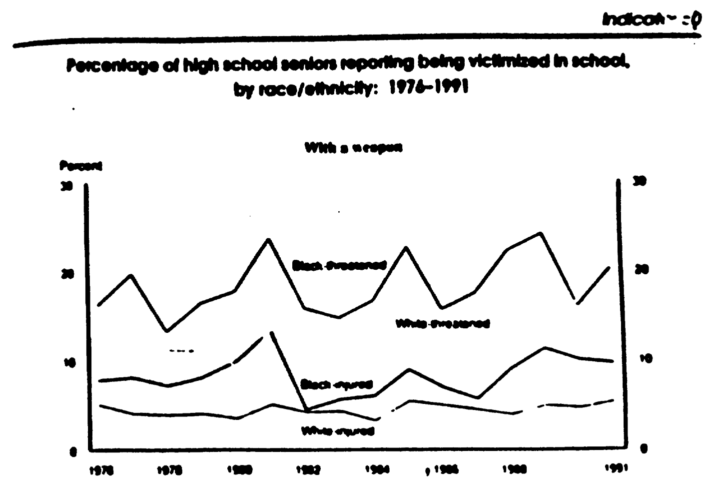
Area Graph



Scatter Plot



Stacked Bar Graph



Line Graph

51

MISLEADING GRAPHS

1. FLEXIBLE GRID

2. IRREGULAR SPACING ON GRID.

3. AXIS DOESN'T START AT 0.
   NEED SCALE BREAK.

4. VISUAL AREA AND NUMERICAL MEASURE.

5. IGNORING THE VISUAL IMAGE.

6. DOUBLE AXES.

7. PERSPECTIVE

8. CHANGE SCALES IN MID-AXIS

9. EMPHASIZE TRIVIAL, IGNORE IMPORTANT

1. FLEXIBLE GRID

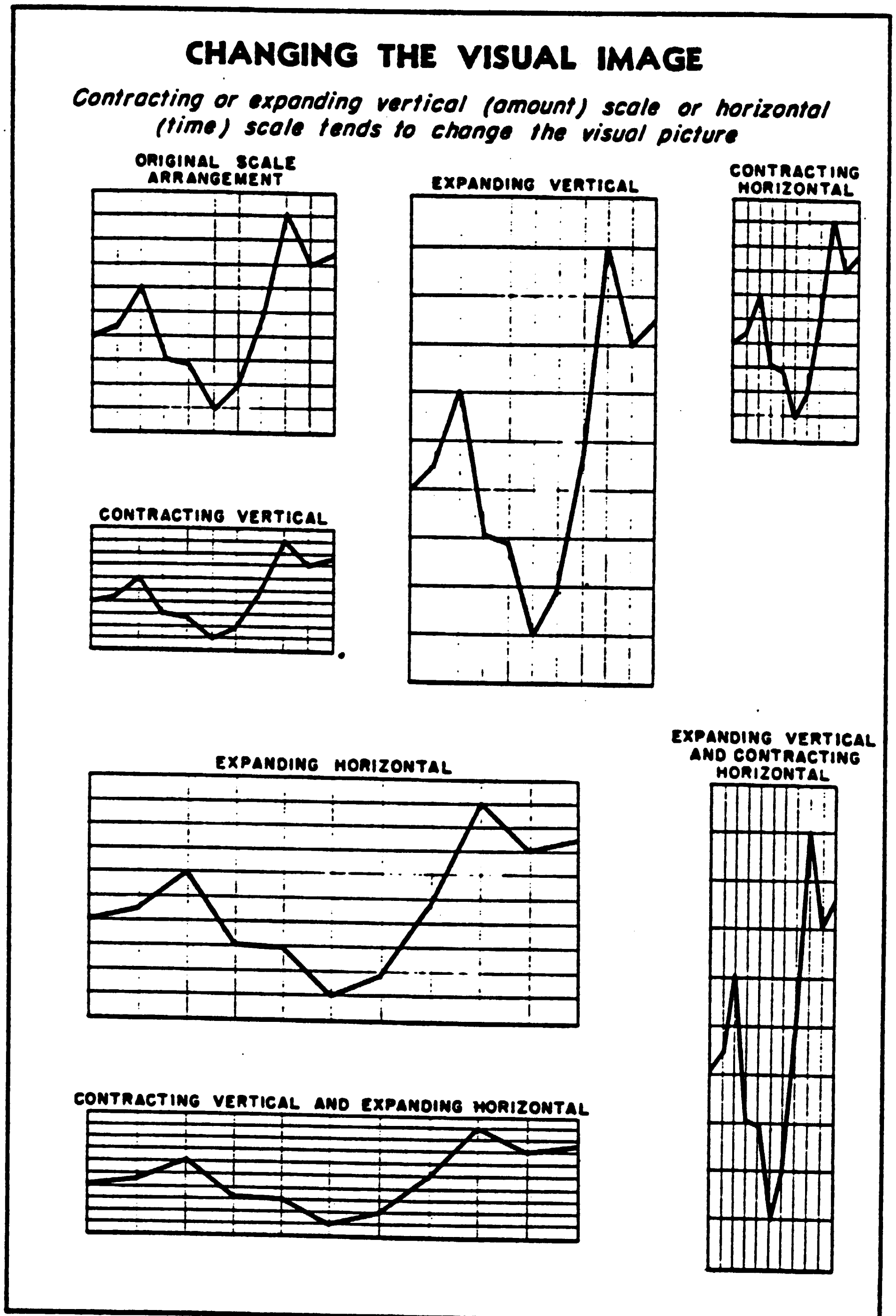


FIG. 3-1 Contracting and expanding the grid.

2. IRREGULAR SPACING ON GRID.

SKIPPING THE GRID

A familiar layout in reports and advertisements is seen in Fig. 3-2A. In order to dramatize the story, a little fudging is done with the time scale. It is not noticeable at a casual glance that the time sequence is not uniform. It seems to be a neat, clean-cut, see-how-we've-grown story. Even the dates lettered at right angles to the base line make the irregular date plotting less noticeable.

Chart *B* in Fig. 3-2 shows what the trend looks like when laid out with the correct grid spacing for each year. Amount plottings for the given years are the same. Spread out this way is not as dramatic, but is the true story.

Chart *G* in Fig. 3-12 makes no allowance for the missing years.

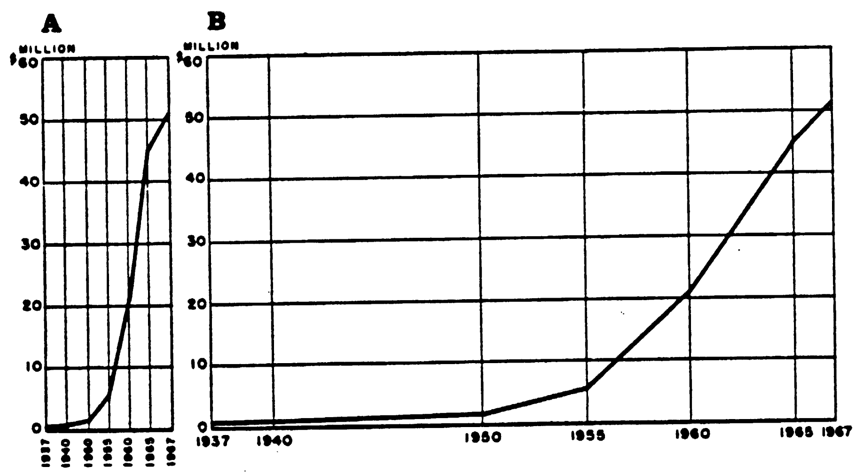


FIG. 3-2 Spacing an irregular time sequence.

3. AXIS DOESN'T START AT 0. NEED SCALE BREAK.

The broken amount scale is commonly used to enlarge on a story. Watch out for it, as it is bound to exaggerate; differences appear greater, and trends seem steeper.

It is essential that charts with an arithmetic scale begin at the zero base line in order to show the true variation in movements. Compare the visual differences between charts *A* and *B* in Fig. 3-3.

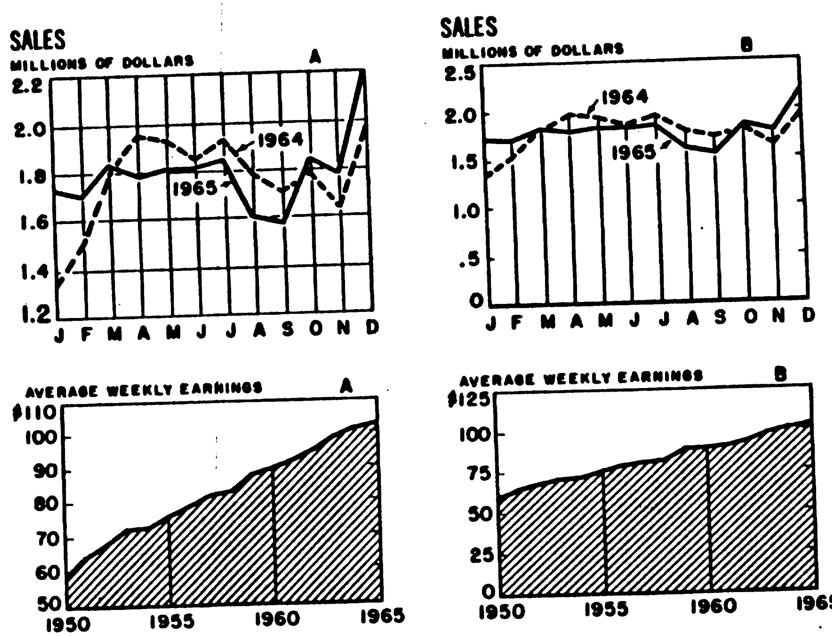


FIG. 3-3 The broken amount scale.

An even greater distortion of the true relationship of amounts occurs when columns or component surface layouts break their scales (see Fig. 3-4).

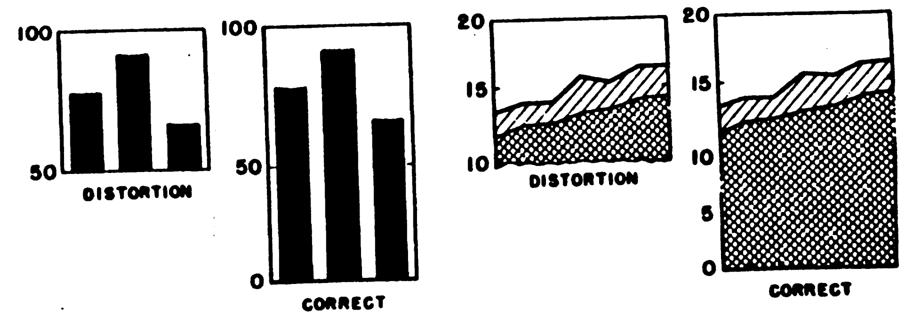


FIG. 3-4 Distortions when breaking the grid.

The draftsman may have indicated that the grid was broken by using several methods, but that is risky. The distorted impression is the one remembered, not the broken scale (see Fig. 3-5).

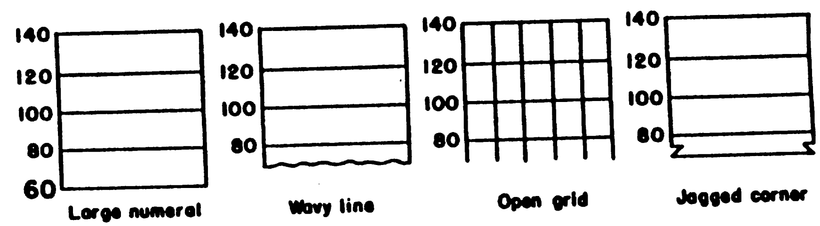


FIG. 3-5 Look for signs of a broken scale.

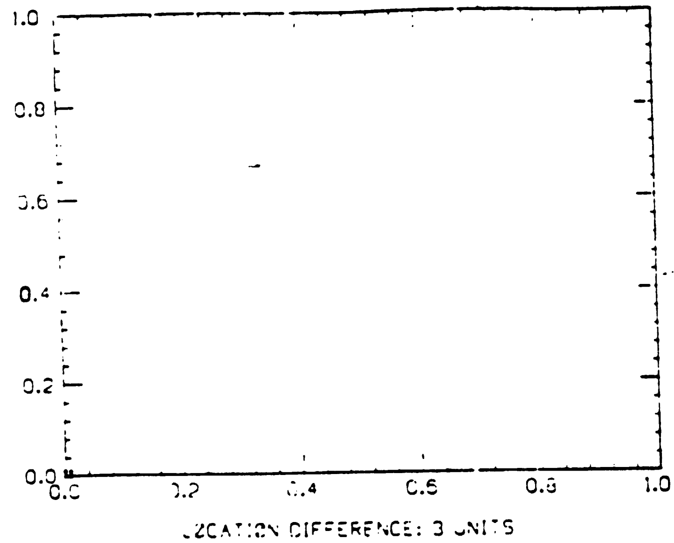


Figure 2. A low density graph (from Friedman and Rafsky 1981 [ddi = .5]).

the worse it is. Tufte (1983) has devised a scheme for measuring the amount of information in displays, called the data density index (ddi), which is "the number of numbers plotted per square inch." This easily calculated index is often surprisingly informative. In popular and technical media we have found a range from .1 to 362. This provides us with the first rule of bad data display.

Rule 1—Show as Few Data as Possible (Minimize the Data Density)

What does a data graphic with a ddi of .3 look like? Shown in Figure 1 is a graphic from the book *Social Indicators III (SI3)*, originally done in four colors (original size 7" by 9") that contains 18 numbers (18/63 = .3). The median data graph in SI3 has a data density of .6 numbers/in²: this one is not an unusual choice. Shown in Figure 2 is a plot from the article by Friedman and Rafsky (1981) with a ddi of .5 (it shows 4 numbers in 8

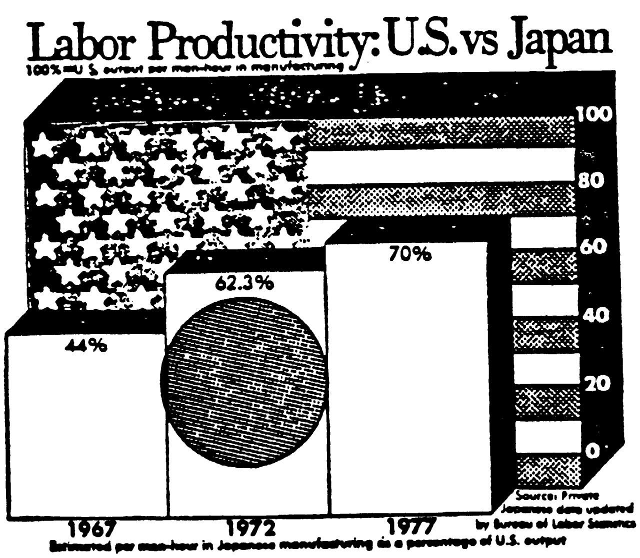


Figure 3. A low density graph (© 1978, The Washington Post) with chart-junk to fill in the space (ddi = .2).

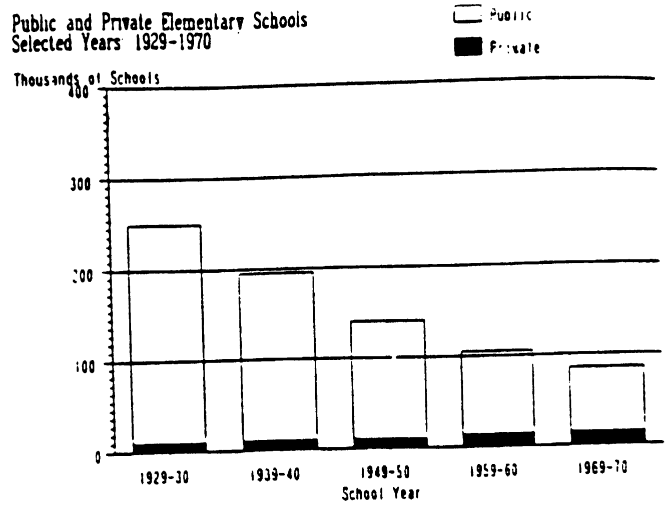


Figure 4. Hiding the data in the scale (from SI3).

in²). This is unusual for JASA, where the median data graph has a ddi of 27. In defense of the producers of this plot, the point of the graph is to show that a method of analysis suggested by a critic of their paper was not fruitful. I suspect that prose would have worked pretty well also.

Although arguments can be made that high data density does not imply that a graphic will be good, nor one with low density bad, it does reflect on the efficiency of the transmission of information. Obviously, if we hold clarity and accuracy constant, more information is bet-

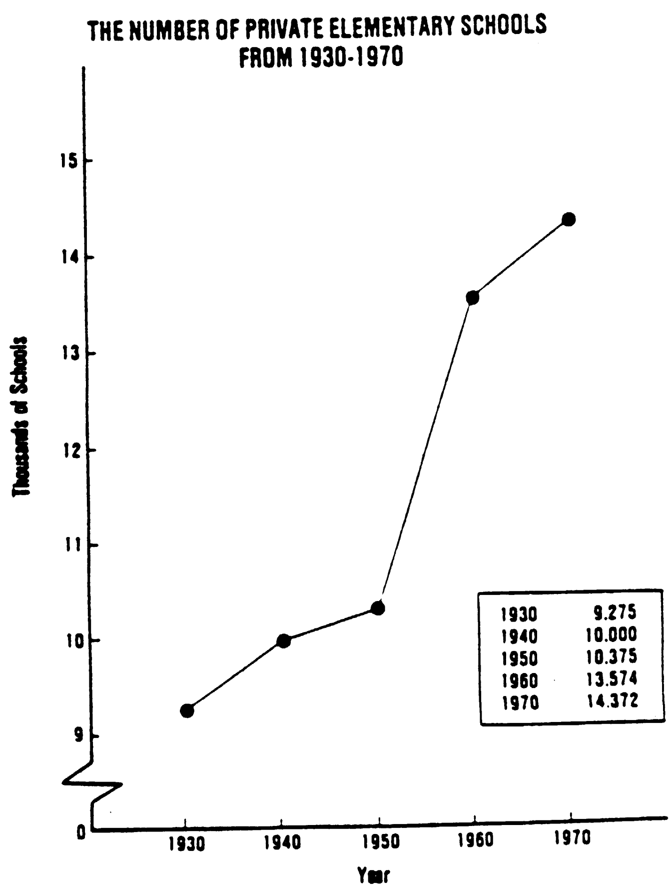
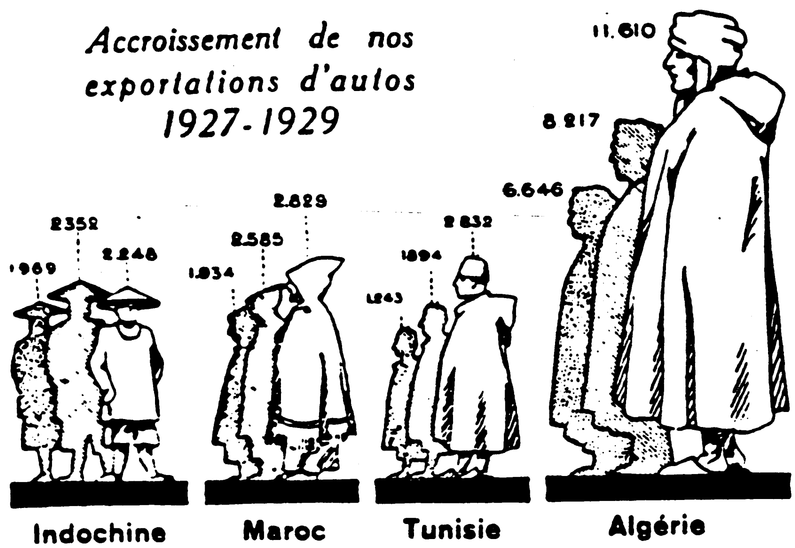


Figure 5. Expanding the scale and showing the data in Figure 4 (from SI3).

4. VISUAL AREA AND NUMERICAL MEASURE.

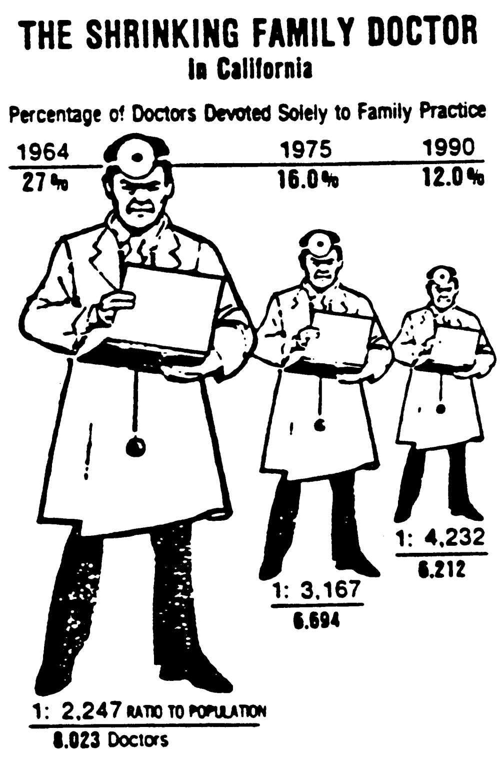
Visual Area and Numerical Measure

Another way to confuse data variation with design variation is to use areas to show one-dimensional data:



R. Satet, *Les Graphiques* (Paris, 1932), p. 12.

And here is the incredible shrinking doctor, with a Lie Factor of 2.8, not counting the additional exaggeration from the overlaid perspective and the incorrect horizontal spacing of the data:



Los Angeles Times, August 5, 1979, p. 3.

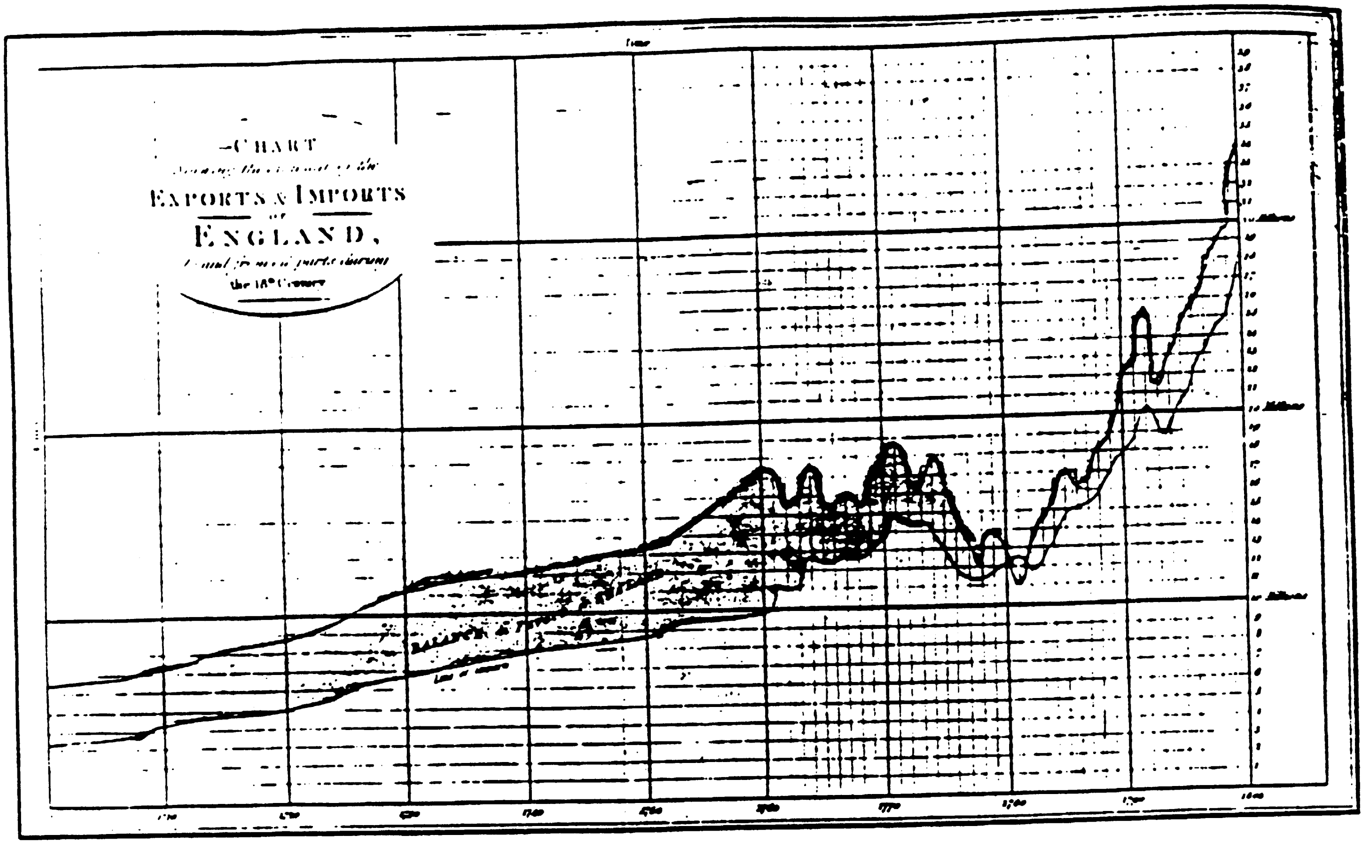


Figure 8. A plot on the same topic done well two centuries earlier (from Playfair 1786).

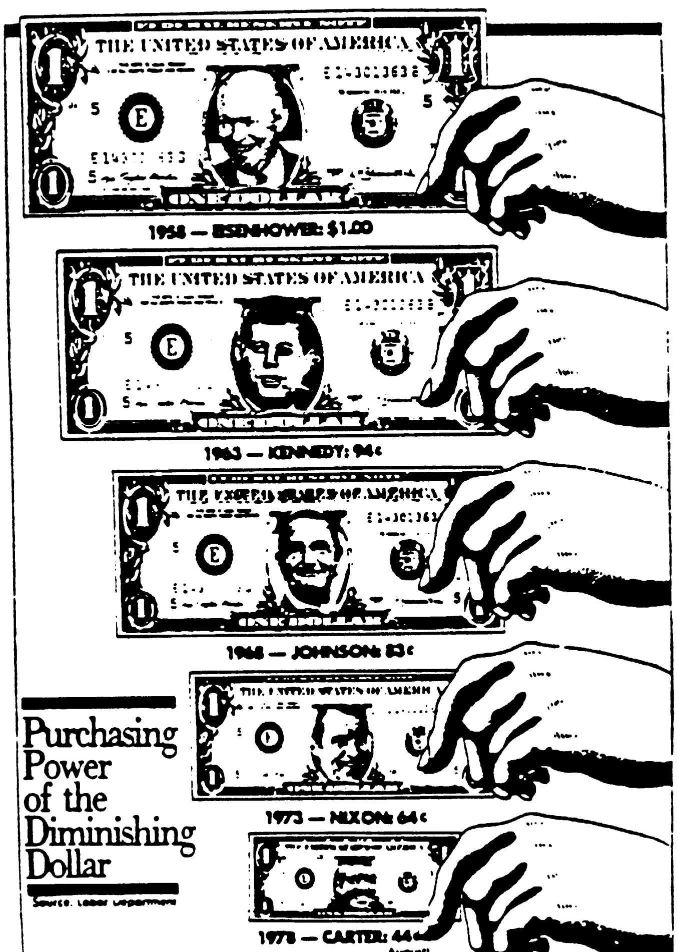


Figure 9. An example of how to goose up the effect by squaring the eyeball (© 1978, The Washington Post).

change in the value of the dollar from Eisenhower to Carter divided by the actual change. I read and measure thus:

$$\textit{Actual} \quad \frac{1.00 - .44}{.44} = 1.27 \qquad \textit{Measured} \quad \frac{22.00 - 2.06}{2.06} = 9.68$$

$$PD = 9.68/1.27 = 7.62$$

This distortion of over 700% is substantial but by no means a record.

A less distorted view of these data is provided in Figure 10. In addition, the spacing suggested by the

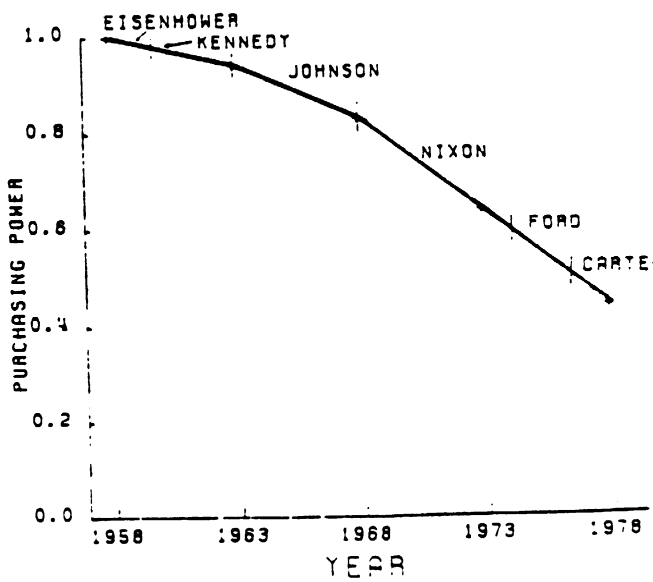
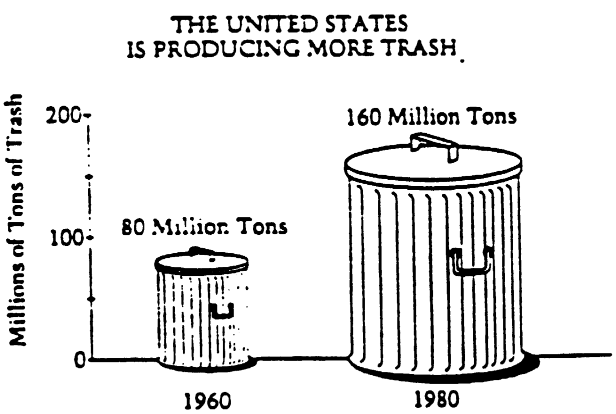


Figure 10. The data in Figure 9 as an unadorned line chart (from Wainer, 1980).

The "Trash Cans" question, which was in the Data Analysis, Probability, and Statistics content area, required eighth-grade students to examine a misleading pictograph and explain why the data display was misleading. To receive credit for a correct response, students needed to note that the 1980 can would hold more than twice the 1960 can or that both the width and height of the can had been doubled. (In particular, doubling the dimensions of the can would lead to an eightfold increase in the volume of the can, because doubling the radius [or diameter] results in a fourfold increase when the radius is squared in $v=\pi r^2 h$.) However, even though the general rather than the specific answer was scored correct, student performance at the national level was quite low, with 8 percent of the eighth-grade students providing an acceptable response.

The ability to read data from a graph, noting the correctness of the graph and the implied comparisons, is an important consumer skill. The ability to detect errors of the type presented in this question is an important outcome of the data analysis/quantitative literacy aspect of the school mathematics curriculum. While some students seem to have developed this critical skill, the results indicate that the vast majority have little conception of the effects that such visual representations can have on the possible interpretations of the data.

EXAMPLE 6: Data Analysis, Statistics, and Probability



Overall Percent Correct*
Grade 8 – 8 (0.8)

The pictograph shown above is misleading. Explain why.

Answer: Both the width and the height of the 1980 can have been doubled. Only the height should have been doubled.

(One of the possible answers)

*The standard errors of the estimated percentages appear in parentheses.

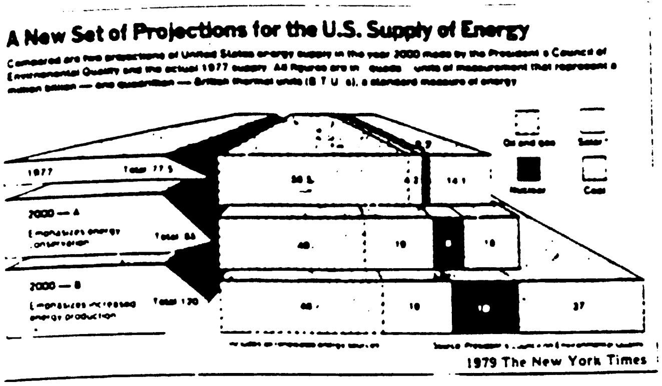
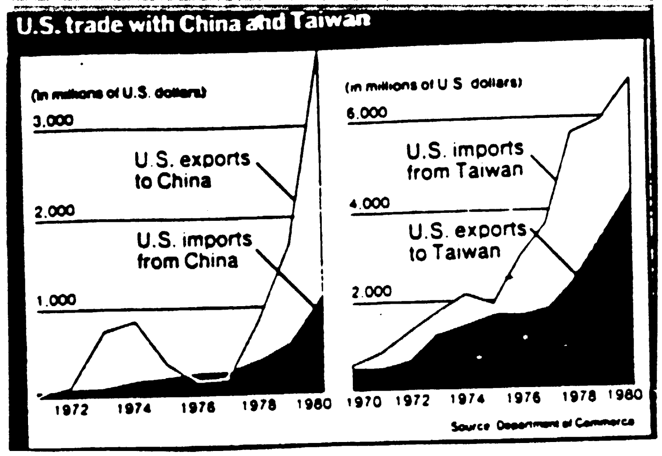


Figure 6. Ignoring the visual metaphor (© 1978, The New York Times).



5. IGNORING THE VISUAL IMAGE.

ter than less. One of the great ass niques is that they can convey large tion in a small space.

We note that when a graph cont mation the plot can look quite empty (Figure 2) and thus raise suspicions in the viewer that there is nothing to be communicated. A way to avoid these suspicions is to fill up the plot with nondata figurations—what Tufte has termed "chartjunk." Figure 3 shows a plot of the labor productivity of Japan relative to that of the United States. It contains one number for each of three years. Obviously, a graph of such sparse information would have a lot of blank space, so filling the space hides the paucity of information from the reader.

A convenient measure of the extent to which this practice is in use is Tufte's "data-ink ratio." This measure is the ratio of the amount of ink used in graphing the data to the total amount of ink in the graph. The closer to zero this ratio gets, the worse the graph. The notion of the data-ink ratio brings us to the second principle of bad data display.

Rule 2—Hide What Data You Do Show (Minimize the Data-Ink Ratio)

One can hide data in a variety of ways. One method that occurs with some regularity is hiding the data in the grid. The grid is useful for plotting the points, but only rarely afterwards. Thus to display data badly, use a fine grid and plot the points dimly (see Tufte 1983, pp. 94–95 for one repeated version of this).

A second way to hide the data is in the scale. This corresponds to blowing up the scale (i.e., looking at the data from far away) so that any variation in the data is obscured by the magnitude of the scale. One can justify this practice by appealing to "honesty requires that we start the scale at zero," or other sorts of sophistry.

In Figure 4 is a plot that (from SI3) effectively hides the growth of private schools in the scale. A redrawing of the number of private schools on a different scale conveys the growth that took place during the mid-1950's (Figure 5). The relationship between this rise and *Brown vs. Topeka School Board* becomes an immediate question.

To conclude this section, we have seen that we can display data badly either by not including them (Rule 1) exhausting the data through the data density; we can sometimes convince viewers that we have included the data through the incorporation of chartjunk. Hiding the data can be done either by using an overabundance of chartjunk or by cleverly choosing the scale so that the data disappear. A measure of the success we have achieved in hiding the data is through the data-ink ratio.

3. SHOWING DATA ACCURATELY

The essence of a graphic display is that a set of numbers having both magnitudes and an order are represented by an appropriate visual metaphor—the magnitude and order of the metaphorical representation match the numbers. We can display data badly by ignoring or distorting this concept.

Rule 3—Ignore the Visual Metaphor Altogether

If the data are ordered and if the visual metaphor has a natural order, a bad display will surely emerge if you shuffle the relationship. In Figure 6 note that the bar labeled 14.1 is longer than the bar labeled 18. Another method is to change the meaning of the metaphor in the middle of the plot. In Figure 7 the dark shading represents imports on one side and exports on the other. This is but one of the problems of this graph; more serious still is the change of scale. There is also a difference in the time scale, but that is minor. A common theme in Playfair's (1786) work was the difference between imports and exports. In Figure 8, a 200-year-old graph tells the story clearly. Two such plots would have illustrated the story surrounding this graph quite clearly.

Rule 4—Only Order Matters

One frequent trick is to use length as the visual metaphor when area is what is perceived. This was used quite effectively by *The Washington Post* in Figure 9. Note that this graph also has a low data density (.1), and its data-ink ratio is close to zero. We can also calculate Tufte's (1983) measure of perceptual distortion (PD) for this graph. The PD in this instance is the perceived

6. DOUBLE AXES.

ıat is, those using two or
<e units, are not suitable
hart is apt to be mislead-
eader is not familiar with

re used, they should start
ıould be taken to identify
mount scale. Both curve
s unit identification, that
In this chart, the vertical
er to "Population," while
al Income."
es, one should be certain
e directly opposite each
lude the plotted amount
ı include 69 millions, the
, whereas each scale unit
30. Such a process divides
into four equal units (see
n zero.
ımponents of a multiple-
r. The charts in Fig. 4-13
ıart in black and white or

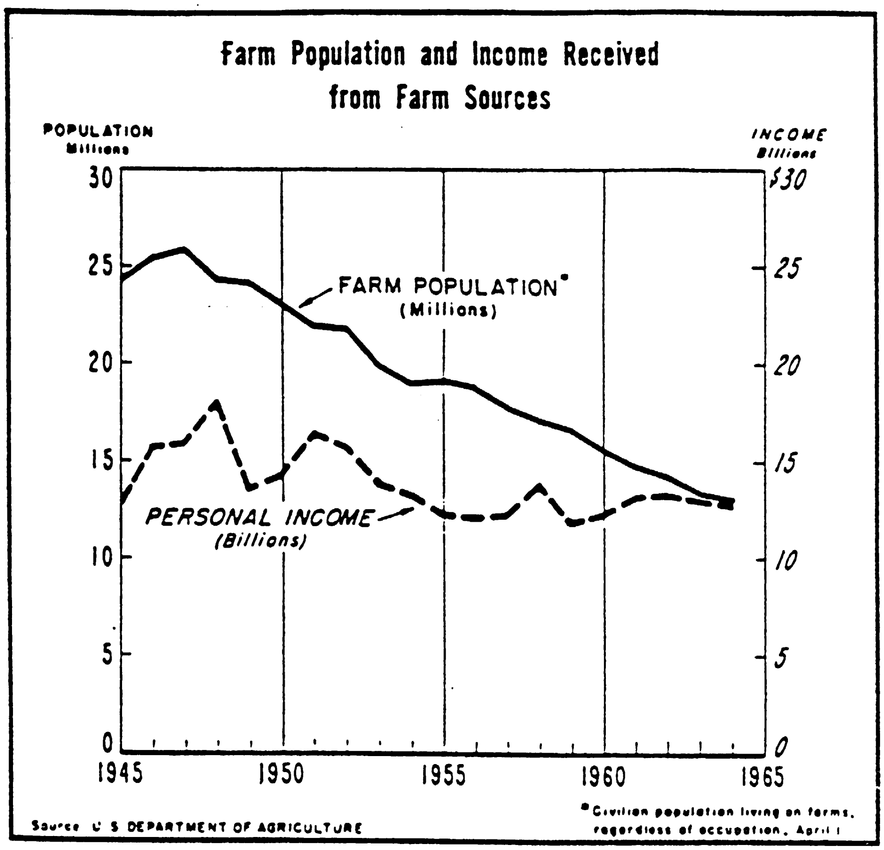


FIG. 4-12 Multiple-amount scales.

THE MULTISCALE COMPLEX

You will come across numerous charts using two or more scales purporting to prove a point. Beware of them. It is too easy to adjust the scales to make one trend visually appear greater in amount and more important than another trend.

Figure 3-9 shows that by changing the population scale in the chart in Fig. 4-12 the "Personal Income" trend assumes more importance.

Check to see that all scales begin from zero and that there is a scale unit relationship (see the discussion of multiple scales in Sec. 4).

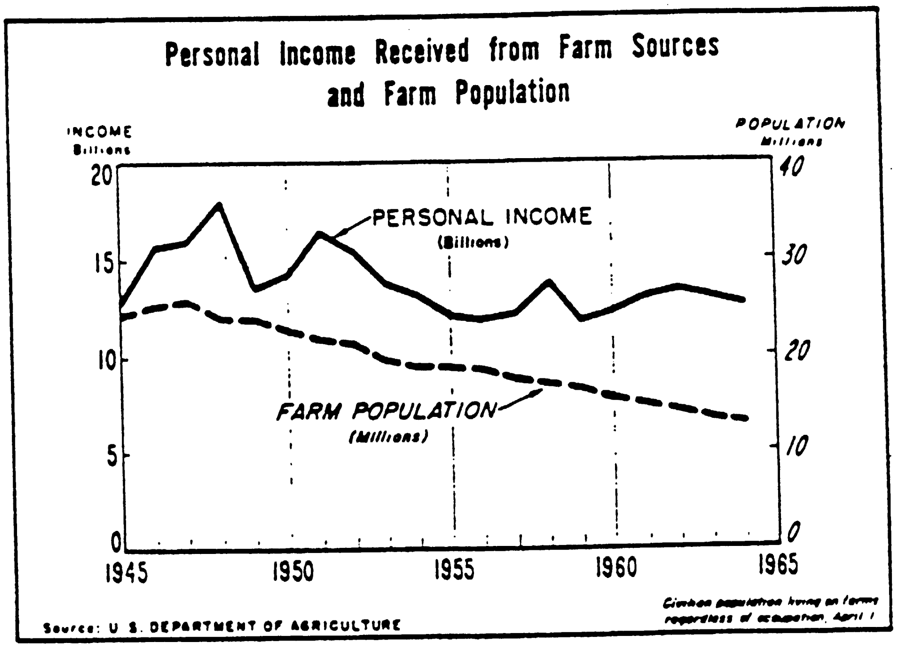


FIG. 3-9 Scrutinize the multiscale chart.

EXAMPLE 86A PERSPECTIVE

Perspective diagrams are hard to interpret. Fig. 86 is supposed to depict the change in the national debt from about 1860 to the present time.[23] This presentation grossly distorts the amplitude of the recent fluctuations. The visual impression is that the debt in 1948 is about $10\frac{1}{3}$ times the debt

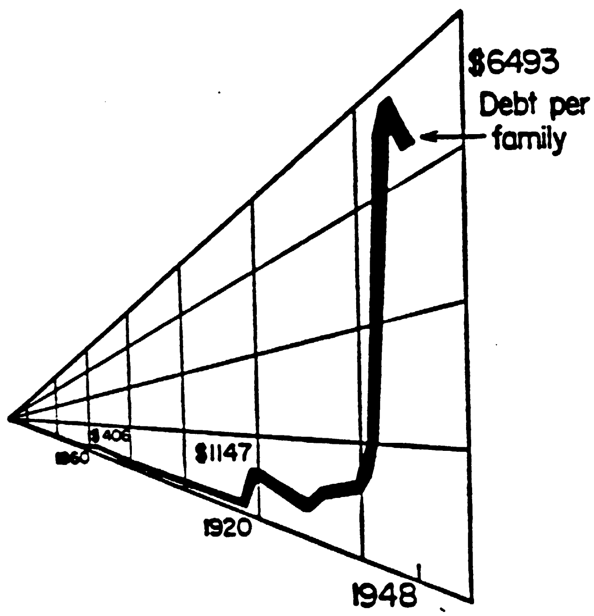


FIG. 86.

of 1920, but the ratio between 1948 and 1920 computed from the debt figures is only $5\frac{2}{3}$. The 1948 figure appears to be about 63 times the 1860 figure, but actually was only 16 times it. Thus, the chart gives two to four times the legitimate impact. The purpose of any chart is to present the facts clearly and simply. Such a perspective diagram does neither. It is easy to suspect that those who use charts that distort may not have a good case.

EXAMPLE 86B DECEPTIVE CHANGES OF SCALE

Fig. 87A sketches the general appearance of a misleading series of charts relating to sales of U. S. Government Series E bonds in the period 1941–1944. It was presented as a model of what "a lively imagination in selecting and compressing data" can do.[24]

---

23. This is the cover design used by the Committee on Public Debt Policy for its *National Debt Series*, issued between World War II and the Korean War.

24. J. A. Livingston, "Charts Should Tell A Story," *Journal of the American Statistical Association*, Vol. 40 (1945), pp. 342–350.

Often we can modify the perception of the graph (particularly for time series data) by choosing carefully the interval displayed. A precipitous drop can disappear if we choose a starting date just after the drop. Similarly, we can turn slight meanders into sharp changes by focusing on a single meander and expanding the scale. Often the choice of scale is arbitrary but can have profound effects on the perception of the display. Figure 11 shows a famous example in which President Reagan gives an out-of-context view of the effects of his tax cut. The *Times'* alternative provides the context for a deeper understanding. Simultaneously omitting the context as well as any quantitative scale is the key to the practice of Ordinal Graphics (see also Rule 4). Automatic rules do not always work, and wisdom is always required.

In Section 3 we discussed three rules for the accurate display of data. One can compromise accuracy by ignoring visual metaphors (Rule 3), by only paying attention to the order of the numbers and not their magnitude (Rule 4), or by showing data out of conte
We advocated the use of Tufte's measure of distortion as a way of measuring the extent t accuracy of the data has been compromised play. One can think of modifications that wo to be applied in other situations, but we leave such expansion to other accounts.

4. SHOWING DATA CLEARLY

In this section we discuss methods for badly displaying data that do not seem as serious as those de- tle (and not so subtle) techniques can be used to effectively obscure the most meaningful or interesting aspects of the data. It is more difficult to provide objective measures of presentational clarity, but we rely on the reader to judge from the examples presented.

Rule 6—Change Scales in Mid-Axis

This is a powerful technique that can make large differences look small and make exponential changes look linear.

In Figure 12 is a graph that supports the associated story about the skyrocketing circulation of *The New York Post* compared to the plummeting *Daily News* circulation. The reason given is that New Yorkers "trust" the *Post*. It takes a careful look to note the 700,000 jump that the scale makes between the two lines.

In Figure 13 is a plot of physicians' incomes over time. It appears to be linear, with a slight tapering off in recent years. A careful look at the scale shows that it

8. CHANGE SCALES IN MID-AXIS

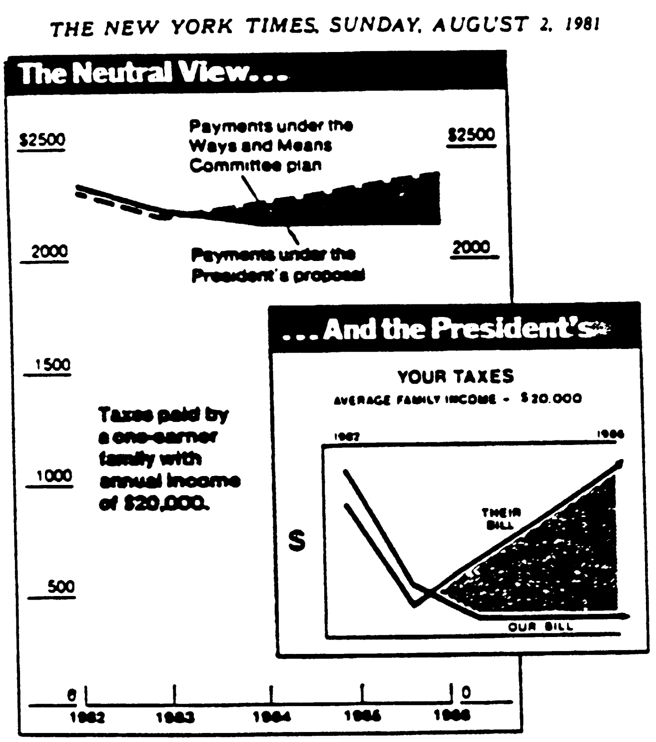


Figure 11. The White House showing neither scale nor context (© 1981, The New York Times, reprinted with permission).

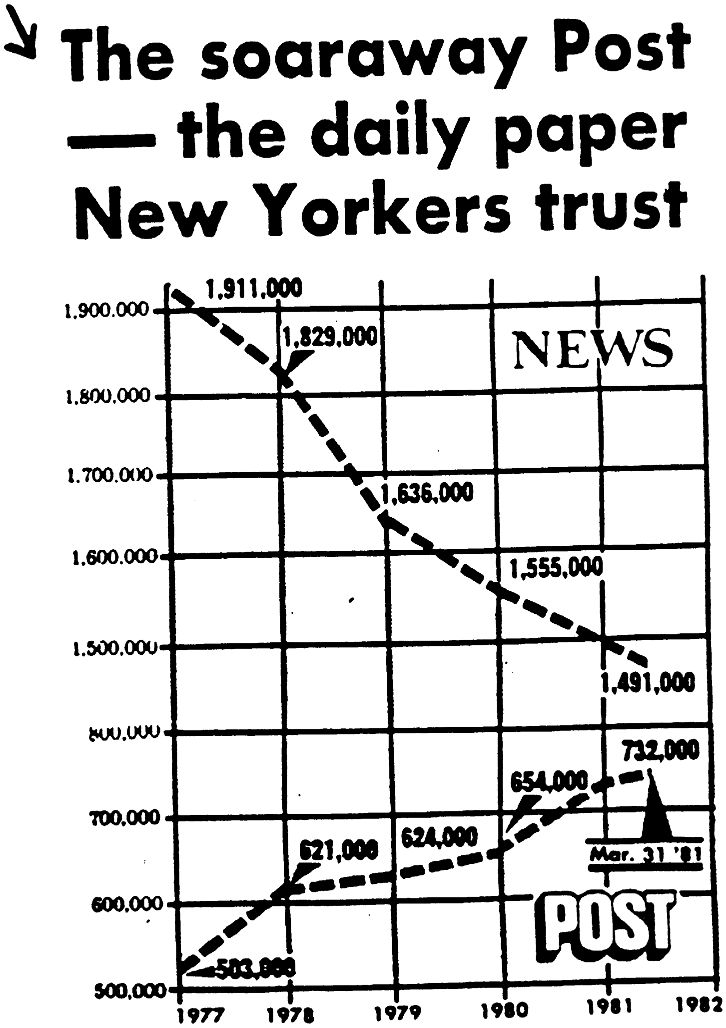


Figure 12. Changing scale in mid-axis to make large differences small (© 1981, New York Post).

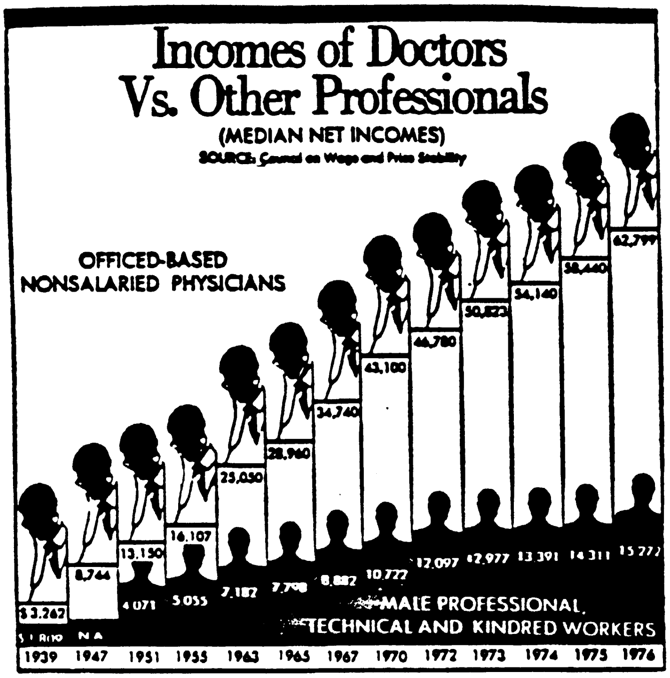


Figure 13. Changing scale in mid-axis to make exponential growth linear (© The Washington Post).

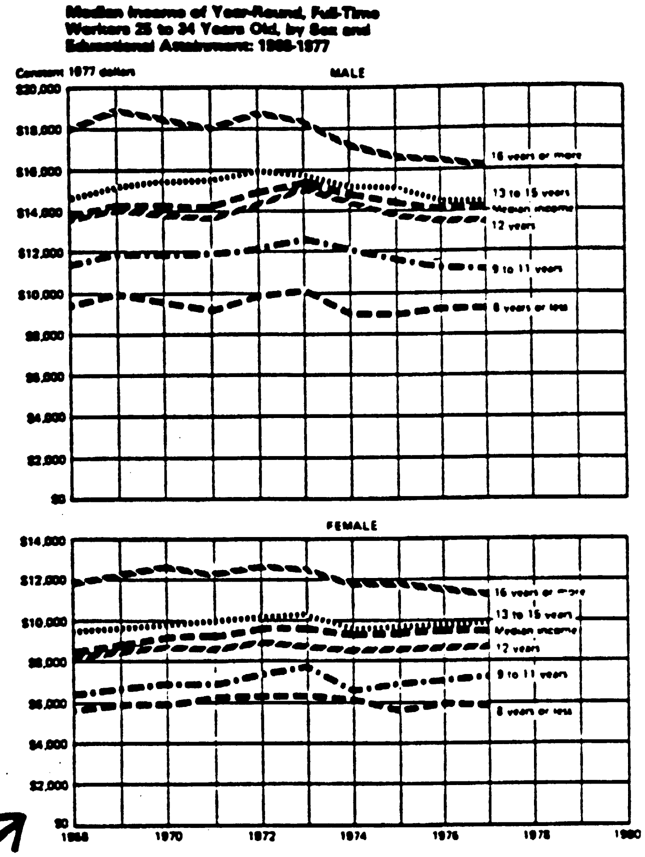


Figure 15. Emphasizing the trivial: Hiding the main effect of sex differences in income through the vertical placement of plots (from SI3).

Rule 7—Emphasize the Trivial (Ignore the Important)

Sometimes the data that are to be displayed have one important aspect and others that are trivial. The graph can be made worse by emphasizing the trivial part. In Figure 15 we have

income levels of m

It reveals the not s

individuals are pai

ones and that chan

9. EMPHASIZE TRIVIAL, IGNORE IMPORTANT

dollars are reasonably constant. The comparison of greatest interest and current concern, comparing salaries between sexes within education level, must be made clumsily by vertically transposing from one graph to another. It seems clear that Rule 7 must have been operating here, for it would have been easy to place the graphs side by side and allow the comparison of interest to be made more directly. Looking at the problem from a strictly data-analytic point of view, we note that there are two large main effects (education and sex) and a small time effect. This would have implied a plot that

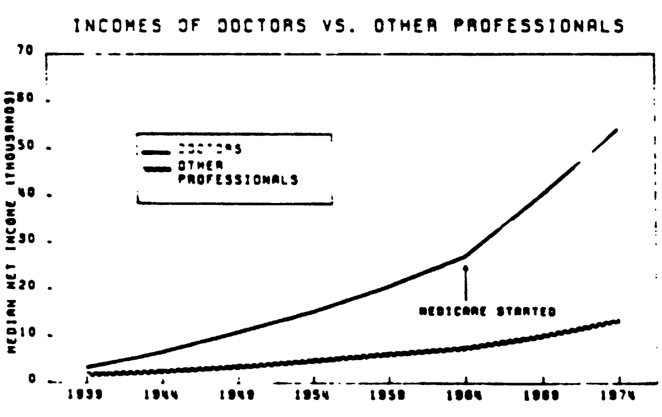


Figure 14. Data from Figure 13 redone with linear scale (from Wainer 1980).

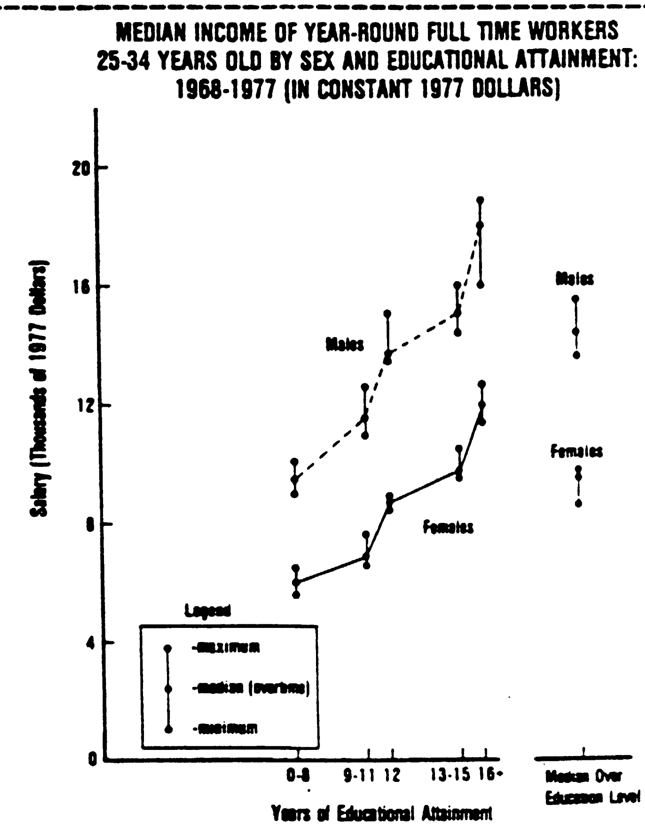


Figure 16. Figure 15 redone with the large main effects emphasized and the small one (time trends) suppressed.

APPRAISAL OF CLAIMS MADE ABOUT DATA

WHEN TO BELIEVE THEM
WHEN TO BE SKEPTICAL
WHEN THEY SHOULD BE IGNORED

Be skeptical about believing estimates or differences associated with:

1. **Large std errors**
2. **Wide confidence intervals**
3. **Results which are not statistically significant**

Not statistically significant does not mean "no difference".

Statistical significance is not the same as substantive significance.

Correlation does not imply causation.

Examine graphs carefully. Be skeptical.

SOME BASIC CONCEPTS OF RESEARCH DESIGN

- Operationalizing your terms
  E.g. "Motivated to Learn"

- Selection Bias
  E.g. Magazine Study
  E.g. Teacher Evaluation

- Need for Control Group
  E.g. Science Major
  E.g. Small Classes
  E.g. Persistence in School
  E.g. NAEP Reading Scores

- Nonresponse Bias
  E.g. Survey on attitudes toward marriage

- Confounding
  E.g. Television teaching
  E.g. Public/Private Schools

- Validity
  E.g. Motivated to learn
  E.g. Urbanicity codes

- Reliability
  E.g. Urbanicity codes
  E.g. Achievement tests

- Generalizability/External Validity
  E.g. Head Start

KEY PRINCIPLE IN EVALUATING RESEARCH CONCLUSIONS:

WHEN YOU COMPARE TWO GROUPS WHICH DIFFER ON SOME CHARACTERISTIC AND FIND THEY DIFFER IN OUTCOMES, YOU WANT TO BE ABLE TO CONCLUDE THAT THIS CHARACTERISTIC IS PROBABLY RESPONSIBLE FOR THE DIFFERENCE.

TO DO SO, YOU MUST EXAMINE AND RULE OUT OTHER COMPETING EXPLANATIONS.

Operationalizing Terms

Term = "motivated to learn mathematics"

Possible operationalizations:

1. As shown by enthusiasm in class
2. As judged by the student's math teacher using a rating scale she developed.
3. As measured by the "math interest" questionnaire.
4. As shown by attention to math tasks in class.
5. As reflected by achievement in math.
6. As indicated by records showing enrollment in math electives.
7. As shown by effort expended in math class.
8. As indicated by number of optional assignments completed.
9. As demonstrated by reading math books outside school.

STUDY

A professor did a study to evaluate student opinion of her performance in a large lecture course.

She asks all students who come to her office hours during a three-week period in the middle of the semester to fill out her questionnaire.

The students give her high marks for accessibility, openness, and willngness to talk to students.

WHAT'S THE PROBLEM?

Selection Bias:

This is a "convenience sample", not a random sample. Students who come to the professor's office have already decided she is accessible. By involving only these students, she is stacking the deck in her favor.

Selection bias refers to factors introduced into the selection of the study population that predetermine the outcome of the study.

Light, Willett, Singer

STUDY

A faculty member at a highly selective college was distressed to discover that nearly a third of students who entered his school as science majors switched to other fields before graduation. The colleague decided this dropout rate was too high and deserved immediate corrective action. He thought it reflected inadequacies in the science program, so he encouraged a curriculum reform committee to consider changes that might improve persistence.

It was later discovered that, in fact, this dropout rate was actually much lower than the rates at almost any similar school. Many felt that this college's program indeed may have been exemplary.

WHAT'S THE PROBLEM?

Lack of a control/comparison group.

Light, Singer, Willett

Understanding Relationships: Using Comparisons

- What is a comparison group?
- Why do you need a comparison group?
- What is an appropriate comparison group?

Understanding Relationships: Using Comparisons-cont.

- Why do you need a comparison group?

A comparison group provides a standard by which to judge your results. Without a comparison group, you CANNOT rule out rival explanations of the results you observe.

Example 1-Teacher satisfaction with small classes

In this hypothetical example, a researcher found that over 90 percent of teachers in elementary classroom with fewer than 15 students were "highly satisfied" with their teaching assignments. She recommended that elementary school uniformly adopt smaller class size, regardless of the expense.

Example 2-Dropout rates from science courses

A researcher found that nearly one-third of students who entered a highly selective college as science majors switched to other fields before graduation. He recommended that a curriculum committee consider changes that might improve persistence.

Understanding Relationships: Using Comparisons-cont.

- What is a comparison group?

A comparison group defines the interpretation of the result that you are reporting. It establishes the baseline against which research results are judged.

Example 1-Trends in Reading Profiency

The *Condition of Education, 1993* reports Trends in Reading Proficiency using three kinds of comparisons--historical comparisons, matched group comparisons, and comparisons against a standard.

Average reading proficiency has increased for 17 year olds since 1971, but not for 9 and 13 year olds.

The gap between the reading proficiency of black and white 13 and 17 year olds has narrowed since 1971.

On average, 9 year olds do not demonstrate reading proficiency at the level where they can interrelate ideas and make generalizations. (anchor point)

Trends in the reading proficiency of 9-, 13-, and 17-year-olds

- **Overall, average reading proficiency for 9- and 13-year-olds was the same in 1990 as in 1971; for 17-year-olds it was somewhat higher in 1990 than in 1971.**
- **Average reading proficiency of black students at all three ages was higher in 1990 than in 1971.**
- **Hispanic 17-year-olds were reading better in 1990 than in 1975.**
- **Between 1971 and 1988, 13- and 17-year-old blacks narrowed gaps between their reading proficiency scores and those of their white counterparts. Similarly, between 1975 and 1988, 17-year-old Hispanics also narrowed gaps between their scores and those of whites. However, large gaps remain, and among black students, the gap did not continue to narrow in 1990.**

Reading skills are basic to the educational process. When students fall behind in their reading proficiency, they may find it difficult to benefit from other aspects of the curriculum. In the future, poor readers may also find it difficult to participate effectively in an economy requiring increasingly sophisticated job skills.

Average reading proficiency (scale score), by age and race/ethnicity: 1971–1990

	Age 9				Age 13				Age 17			
Year	All races	White	Black	Hispanic	All races	White	Black	Hispanic	All races	White	Black	Hispanic
1971	208	214	[1]170	—	255	261	[1]222	—	[1]285	[1]291	[1]239	—
1975	210	217	[2]181	183	256	262	[1]226	233	[1]286	293	[1]241	[1]252
1980	[1,2]215	[2]221	[2]189	190	259	[2]264	[1,2]233	237	286	293	[1]243	[1]261
1984	211	[2]218	[2]186	187	257	263	[2]236	240	289	[2]295	[2]264	[2]268
1988	212	218	[2]189	194	258	261	[2]243	240	[2]290	295	[2]274	[2]271
1990	209	217	[2]182	189	257	262	[2]242	238	[2]290	[2]297	[2]267	[2]275

Average reading proficiency (scale score), by age and sex: 1971–1990

	Age 9		Age 13		Age 17	
Year	Male	Female	Male	Female	Male	Female
1971	201	214	250	261	279	291
1975	204	216	250	262	280	[1]291
1980	[1,2]210	[1,2]220	[2]254	263	282	[1]289
1984	[2]208	214	253	262	[2]284	294
1988	[2]208	216	252	263	[2]286	294
1990	204	215	251	263	284	297

— Not available.

[1] Statistically significant difference from 1990.

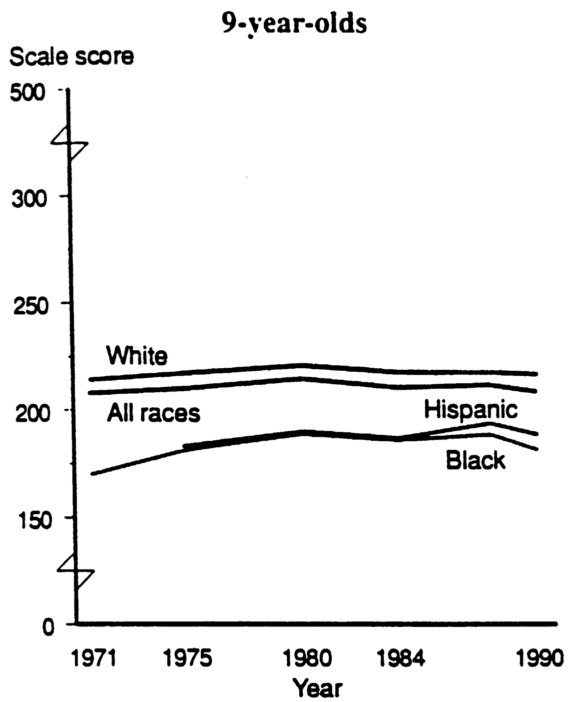
[2] Statistically significant difference from 1971 for all except Hispanics. Statistically significant difference from 1975 for Hispanics.

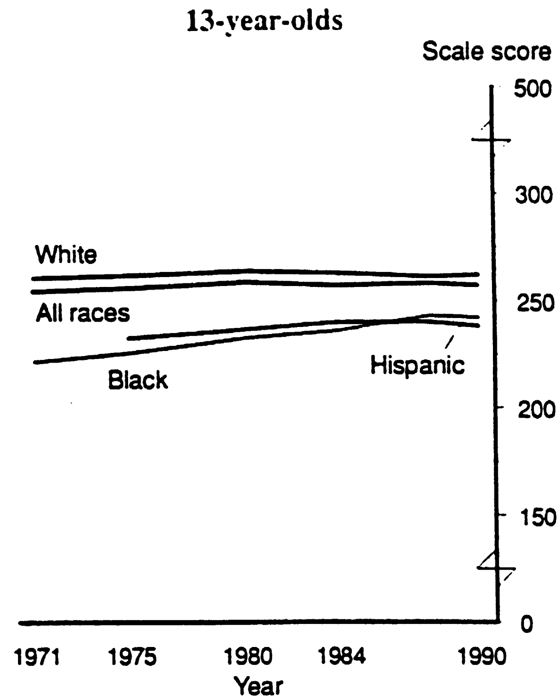
NOTE: Reading Proficiency Scale has a range from 0 to 500

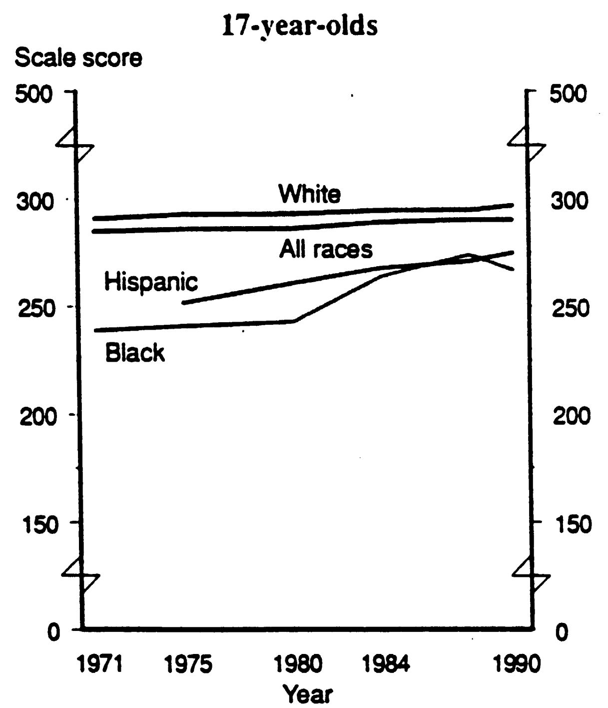
Level 150: Simple discrete reading tasks
Level 200: Partial skills and understanding
Level 250: Interrelate ideas, and make generalizations
Level 300: Understands relatively complicated information
Level 350: Learns from specialized reading materials

SOURCE: National Assessment of Educational Progress, *Trends in Academic Progress: Achievement of American Students in Science, 1969-70 to 1990, Mathematics, 1973 to 1990, Reading, 1971 to 1990, and Writing, 1984 to 1990,* 1991.

Average reading proficiency, by age and race/ethnicity: 1971–1990







SOURCE: National Assessment of Educational Progress, *Trends in Academic Progress: Achievement of American Students in Science, 1969-70 to 1990, Mathematics, 1973 to 1990, Reading, 1971 to 1990, and Writing, 1984 to 1990*, 1991.

Understanding Relationships: Using Comparisons-cont.

Example 2-Persistence in School

The *Condition* also reports on students' persistence in school using both historical and group comparisons.

Persistence rates among college students increased between 1972 and 1991.

The high school persistence rate for students from high income families is about 10 percent higher than the rate for students from low income families.

Persistence in school

- **Between 1990 and 1991, 96 percent of 15- to 24-year-olds in grades 10-12 stayed in school or completed high school. The other side of this statement is that 4 percent dropped out before completion (although some of these dropouts may have re-enrolled during a subsequent school year).**
- **The high school persistence rate for students from high income families is about 10 percent higher than the rate for students from low income families. The difference in rates between students from high and middle income families is small, about 3 percent (see supplemental table 5-2).**

A measure of persistent attendance is the proportion of students enrolled in 2 consecutive years. Students who do not complete high school face a decreased opportunity for assuming a successful and fully functional place in the American workplace and society at large. Persistent attendance is strongly associated with completing high school. In college, both persistent attendance and full-time attendance are strongly associated with completion of a 4-year degree. Those who attend part-time or stop out (i.e., have periods of nonattendance) are less likely to complete a degree.

- **In October 1991, 84 percent of college students who had been enrolled in their first, second, or third year of college the previous October were still enrolled.**
- **Persistence rates among college students at each level increased between 1972 and 1991 (supplemental table 5-4).**

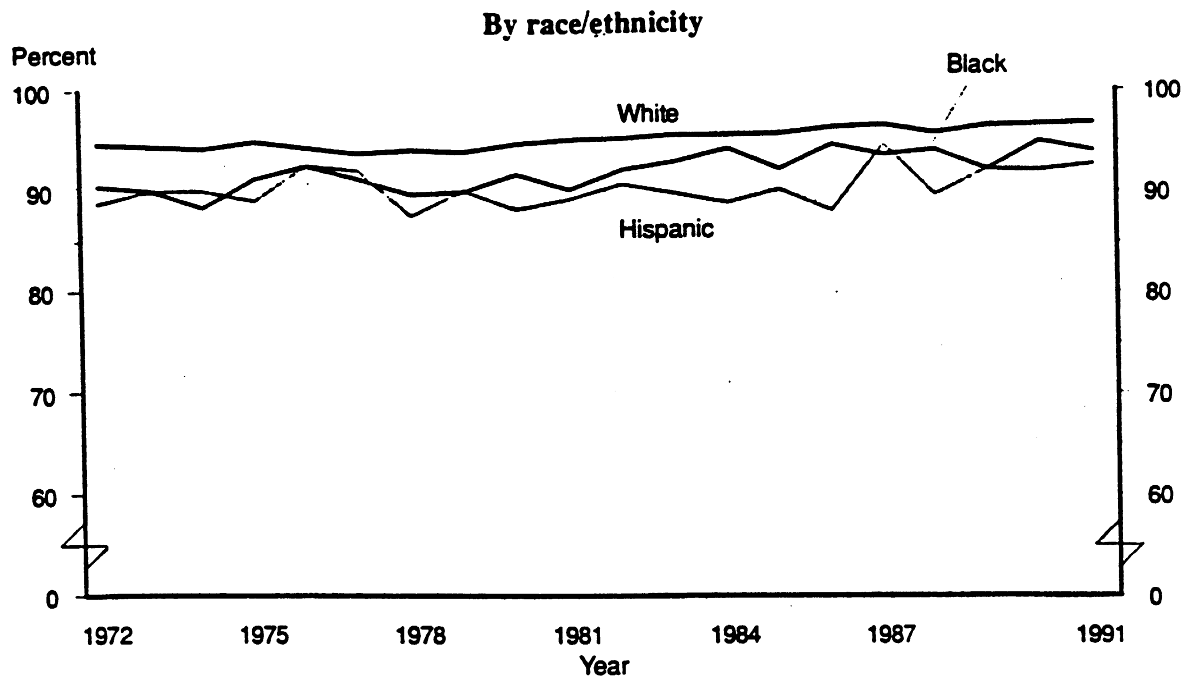
Percentage of high school and college students enrolled the previous October who are enrolled again the following October: 1972–1991

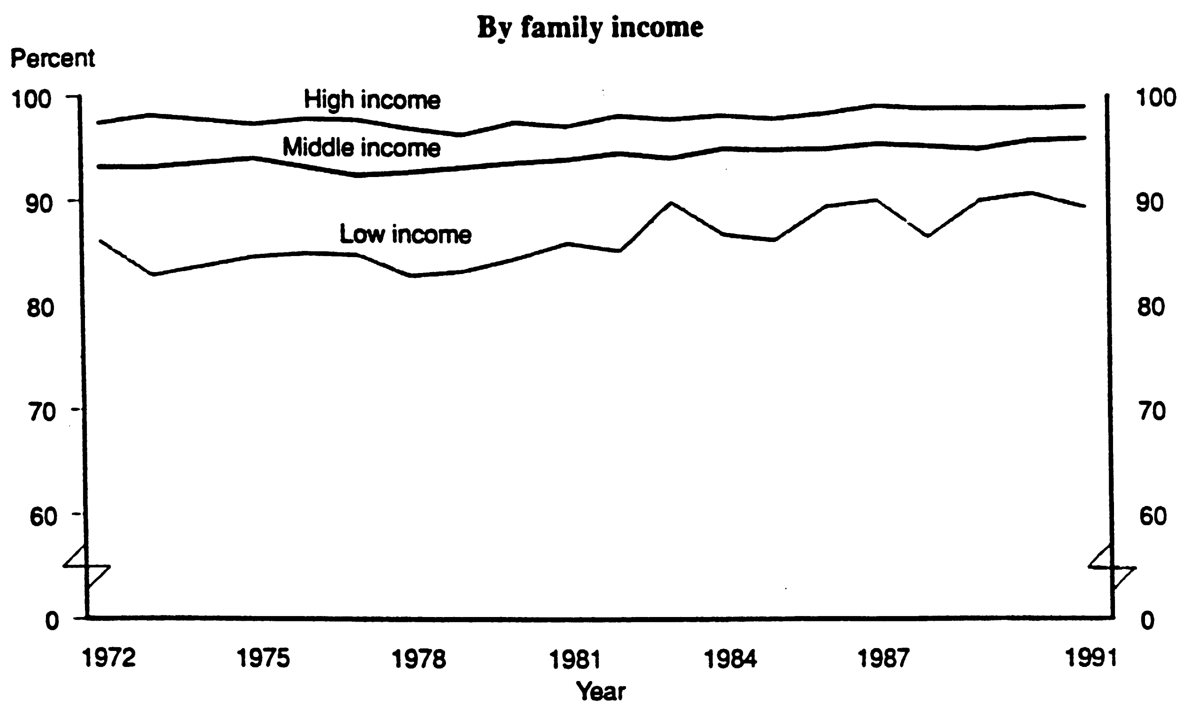
	High school students, grades 10–12, ages 15–24				College students, 1st–3rd years, ages 16–24			
October	Total	White	Black	Hispanic	Total	White	Black	Hispanic
1972	93.9	94.7	90.5	88.8	77.7	78.1	71.3	78.1
1973	93.7	94.5	90.1	90.0	76.7	76.8	77.2	73.8
1974	93.3	94.2	88.4	90.1	77.5	77.4	74.3	76.0
1975	94.2	95.0	91.3	89.1	79.3	79.9	77.0	72.8
1976	94.1	94.4	92.6	92.7	79.2	79.3	81.3	74.9
1977	93.5	93.9	91.4	92.2	79.2	79.3	79.1	75.9
1978	93.3	94.2	89.8	87.7	77.7	77.8	75.3	76.7
1979	93.3	94.0	90.1	90.2	77.8	78.4	73.6	72.4
1980	93.9	94.8	91.8	88.3	79.0	80.2	71.0	69.2
1981	94.1	95.2	90.3	89.3	78.0	79.4	72.3	72.5
1982	94.5	95.3	92.2	90.8	80.4	81.2	74.6	77.4
1983	94.8	95.6	93.0	89.9	80.3	81.1	74.8	74.4
1984	94.9	95.6	94.3	88.9	79.1	79.8	74.2	72.8
1985	94.8	95.7	92.2	90.2	79.7	81.0	71.4	67.7
1986	95.3	96.3	94.6	88.1	80.2	80.5	74.4	81.7
1987	95.9	96.5	93.6	94.6	81.3	82.9	69.6	74.9
1988	95.2	95.8	94.1	89.6	83.0	83.7	78.0	77.0
1989	95.5	96.5	92.2	92.2	83.8	84.3	79.0	81.1
1990	96.0	96.7	95.0	92.1	81.8	81.7	79.4	79.7
1991	96.0	96.8	94.0	92.7	84.1	84.4	77.8	80.8

NOTE: High school students were either enrolled again the following October or had graduated. See supplemental note to *Indicator 4* for details on how the persistence rates in this table are calculated. Not shown separately but included in the total are non-Hispanics who are neither black nor white. Data for 1987 through 1991 reflect new editing procedures instituted by the Bureau of the Census for cases involving missing school enrollment items.

SOURCE: U.S. Department of Commerce, Bureau of the Census, October Current Population Surveys.

Percentage of high school students in grades 10–12 and from ages 15–24 enrolled in the previous October and again the following October*: 1972–1991





* Or who had completed high school

NOTE: Low income is defined as the bottom 20 percent of all family incomes; high income is defined as the top 20 percent of all family incomes; and middle income is defined as the 60 percent of family incomes between high and low incomes.

SOURCE: U.S. Department of Commerce, Bureau of the Census, October Current Population Surveys.

STUDY

In 1987, author and social investigator Shere Hite published her third book on men and women. Her latest findings on women's attitudes about men, sex, and personal and marital relationships put her on the cover of Time and lauched a flood of news stories and TV talk.

100,000 detailed questionnaires,
127 questions
women in groups of many kinds all over the country

4500 replies

Report

84 percent of the women in her study were dissatisfied with their marital or other intimate relationships,
78 percent said they were generally not treated as equals by men,
70% of those married more than five years had had affairs.

And so on, with a number of answers and Hite's elaborations indicating that women in general are mainly unhappy with their relationships.

WHAT'S THE PROBLEM?

NONRESPONSE BIAS

Women in general? At one point, she said "no one can genralize" from her findings. Yet, she also claimed that her respondents were typical.

Critics said her sample was almost certainly heavily weighted with the unhappiest women, those who took the time to answer the lengthy questinnaire. Many women probabably feel the same way - but we have no idea how many.

Washinton Post - ABC polling team questioned by phone a representative sample of women and men across the nation. They found that

> 93 percent of the married and single women said they were satisfied with their relationships,
>
> 81% said they were treated as equals most of the time,
>
> only 7 percent reported affairs.

STUDY

A study is done to compare the effectiveness of televised instruction versus regular classroom instruction. Students were randomly assigned to one of the two groups. At the end of the course, the investigator compared the progress of students in the two groups, found students in the television group performed better, and concluded that the television approach was more effective.

POINTS TO CONSIDER

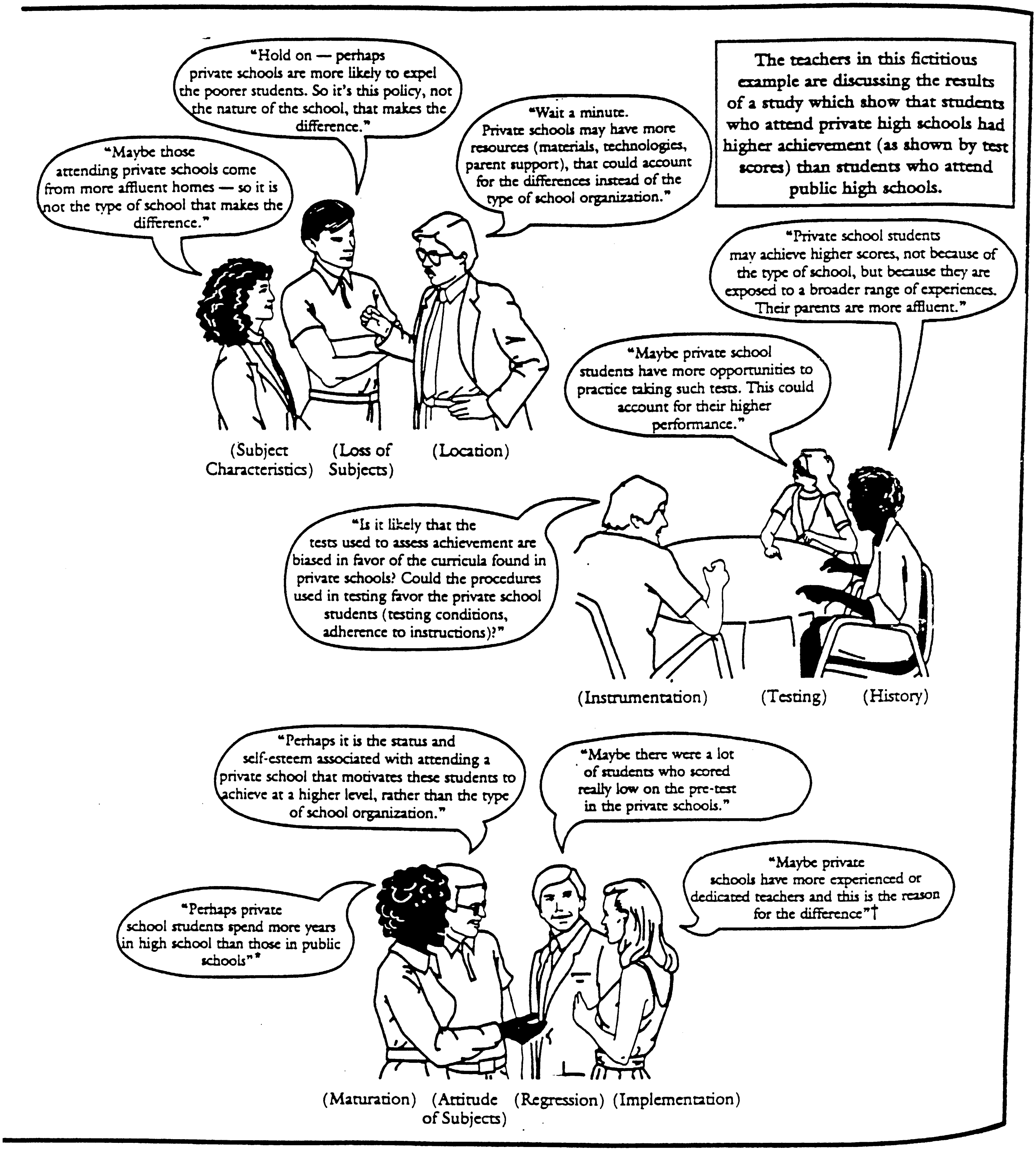
Confounder - a factor which differs between the treatment and control group and is likely to affect the outcome.

Confounder here is the quality of the teacher. When this type of research was done, the usual procedure was to select the best teacher available and give this person the full day to prepare the lesson.

Better controlled studies found no difference between the two groups.

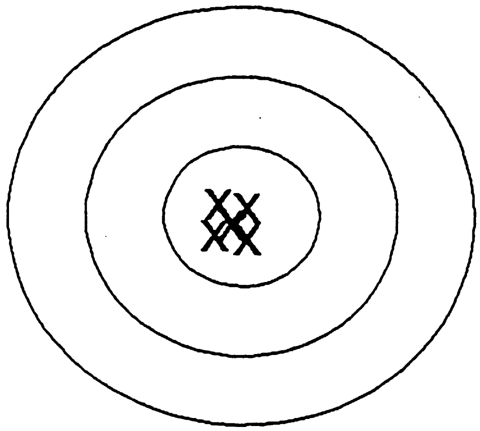
FIGURE 12.1

Illustration of Threats to Internal Validity

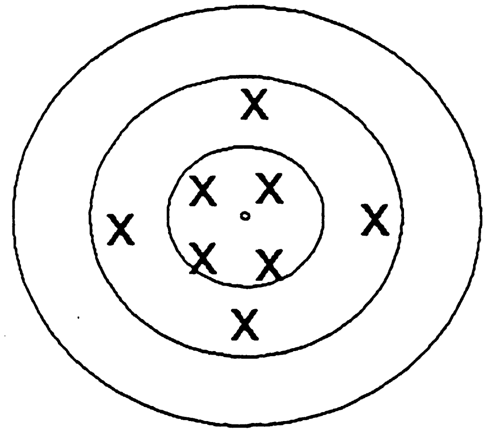


Note: We are not implying that any of these statements are necessarily true; our guess is that some are and some are not.
*This seems unlikely.
†If these teacher characteristics are a *result* of the type of school, then they do not constitute a threat.

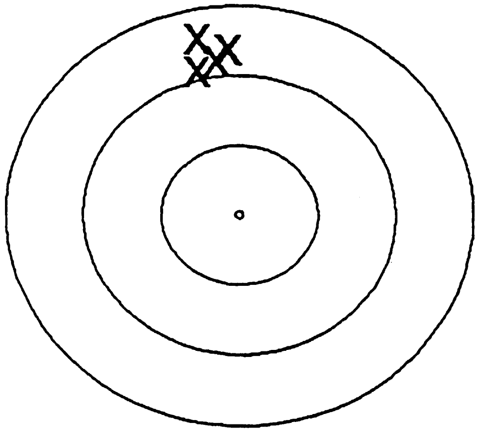
VALIDITY AND RELIABILITY



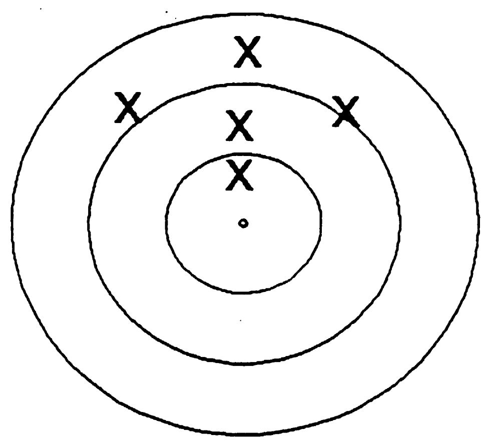
valid and reliable



valid, not reliable



not valid, reliable



not valid, not reliable

reliability: the reproducibility of a result when a test or study is repeated

validity: how well a measure actually assesses what you want it to.

Validity and Reliability-cont.

Example: SASS Community Type

VALIDITY:

Is the community that the school is located in really the level of urbanicity that the principal reports? (Example of Fairfax City Schools)

RELIABILITY:

If you were to readminister the questionnaire tomorrow, would the principal respond differently?

Locale Codes (columns) Versus Self-Report (rows)

	Large City	Mid-size City	Large City Fringe	Midsize City Fringe	Large Town	Small Town	Rural
Small City	.27%	7.3%	6.7%	11.8%	7.8%	51.1%	15.0%
Suburb of Med. City	.50%	20.8%	16.9%	34.5%	2.2%	11.9%	13.2%

Collapsed Locale codes (columns) Versus Self Report (rows)

	Urban	Suburban	Small Town/Rural
Small City	7.6%	26.3%	66.1%
Suburb of Med. City	21.3%	53.6%	25.1%

Understanding Relationships: Generalizability

- Defines the target population of the research

- Are the results applicable to a broad target population or are they too specific to a particular set of places, person, and times to be useful for general policy making?

- Narrow target populations mean less generalizability, but may mean more ability to detect effects

  Example: Shy Females Study

- Broad target populations mean more generalizability, but may be less feasible

  Example: Introductory Psychology

- KEY ISSUE--Don't generalize beyond what your sample allows!

Understanding Relationships: Generalizability-cont.

What are the pitfalls of overgeneralization?

Example--Head Start Study

Many policy discussions about the efficacy of Head Start and decisions about funding of Head Start have been based upon a study conducted in Ypsilanti, Michigan of a model Head Start program.

What were the characteristics of the program?

How many children were in the study?

What were the results of the study?

How have they been used?

A MORE IN-DEPTH LOOK AT ONE EXAMPLE: THE CASE OF RESEARCH IN BILINGUAL EDUCATION

References:

Ann Willig, "A Meta-Analysis of Selected Studies on the Effectiveness of Bilingual Education",RER, 1985, Vol 55, No. 3

Keith Baker, "Comment on Willig's "A Meta Analysis of Selected Studies on the Effectiveness of Bilingual Education", RER, 1987, Vol. 57, No. 3

Ann Willig, Response to Baker, RER, 1987, Vol. 57, No. 3

WHAT IS THE OUTCOME VARIABLE?

- Different interpretations of what constitutes success.

  - Successful as long as it does not hinder children in the learning of English while it promotes learning of the nonlanguage subjects.

  - Successful if it improves achievement in school.

  - Successful if the children can be taught in the second language and still maintain grade level in nonlanguage subjects.

  - Successful if it accelerates children's learning of English over what it would have been without the program.

WHAT GROUPS ARE BEING COMPARED (TREATMENT/CONTROL)?

WHAT ARE THE CHARACTERISTICS OF THE CHILDREN BEING STUDIED AND THEIR COMMUNITIES?

RESEARCH STRATEGIES

PROBLEMS

LACK OF RANDOM ASSIGNMENT LEADING TO PROBLEMS WITH CONFOUNDING FACTORS

- Uncontrolled differences between the experimental and control groups when random assignment is not used which contribute significantly to the results.

CONFOUNDING FACTORS:

When random assignment was not used, bilingual students differed from control students in several ways:

(a) in language dominance and/or their need for a bilingual program.

When both groups were Spanish dominant, there is an effect of almost one half of a std dev favoring the experimental group.

When the experimental group was Spanish dominant and the comparison group was English dominant, there is little or no difference between the groups.

When both groups were English dominant, there is little difference.

(b) Some comparison groups contained students who were not qualified for a bilingual program, were not deemed limited English proficient.

(c) some comparison groups contained students who had exited from bilingual programs. These studies tended to show no benefit for the bilingual group.

(d) some comparison groups contained schools having no bilingual program. It is most likely that in these schools there is an insufficient number of non-English speaking children in the attendance center. Children in such schools tend to be exposed to more English from their peers, teachers, and neighbors.

Generally, when one has a nonrandomized study and is concerned about the influence of possible extraneous variables, one tries to adjust statistically for these differences. Many researchers believe, however, that in program evaluation research, such adjustments will be underadjustments and will make the program look less effective than it really is.

PROBLEMS WITH THE MAINTENANCE OF DEFINITION OF TREATMENT AND CONTROL GROUPS

In addition, treatment and control programs failed to maintain their unique identity

(a) some treatment groups changed in composition such that, subsequent to the pretest and prior to the posttest, the better students exited and more needy students entered.

(b) stability of the treatment program (e.g. teacher turnover, reorganization of program)

(c) some comparison programs contained elements of bilingual programs such as bilingual teachers or aides who had previously taught in bilingual programs.

PROBLEMS WITH THE RELIABILITY AND VALIDITY OF THE LANGUAGE TESTS

Many claim that the language tests used to determine entry into bilingual programs have low reliability and validity. Individuals possess a variety of language skills and competence and performance will vary depending on the context or setting of language use, the interactants, their relationships and relative statuses, the domain of the communicative intent, and the topic.

"FINE" ANALYSIS CRITERIA FOR QUANTITATIVE STUDIES

I. Introduction to Problem
   A. Stated problem clear and researchable?
   B. Thorough review of literature
   C. Clear hypotheses/research questions

II. Research Procedures
   A. Representative sample
      1. Characteristics of sample described
      2. Did sample selection methods produce unbiased sample?
      4. Numbers of participating and nonparticipating given
      5. Sample size large enough?
   B. Data Gathering Techniques
      4. Validity/reliability
   C. Research design and procedures appropriate/replicative
      1. Research design appropriate for question
      2. Procedures described
      3. Research design eliminated confounders

III. Discussion
   A. Results appropriate and clear
      1. Statistical techniques appropriate
      2. Results presently clearly
      3. Levels of significance and degrees of freedom
      4. Graphs and tables discussed
      5. Every hypothesis tested.
   B. Results of analysis support conclusions
      4. Limitations of study discussed
   C. Recommendations for future action

IV. Method Specific Criteria
   A. Surveys/Questionnaires
   B. Correlational Studies
   C. Causal-Comparative Studies
      E.g. SES and GPA
      2. Extraneous variables identified and controlled
      3. Caution in causal statements

4. Alternative hypotheses discussed

D. Experimental Studies
   1. Group formation methods described
   2. Participants selected randomly
   3. Random assignment
   4. Extraneous variables identified
   5. Control for extraneous variables

E. Quasi-Experimental Studies
   1. Groups compared such that relatively similar
   2. Extraneous variables controlled
   3. Caution in causal statements

SOME QUESTIONS TO ASK
(Victor Cohn)

How do you know?

Are there studies supporting the claims?

Were the studies acceptable ones, by general agreement?

Were there enough people in the study?

Were appropriate control groups used?

Was the sample studied representative of the population?

Have results been fairly consistent from study to study?

Do the results hold across subgroups or only for particular subpopulations?

If the results are based on questionnaires, were the questions likely to elicit accurate, reliable answers?

What was the response rate? Were the nonrespondents different from the respondents?

Do you have a conclusion or suggestion for further study?

Are there other possible explanations for the differences or relationships you are seeing?

Have the findings resulted in consensus among others in the same field? Do at least the majority of informed persons agree? Or should we withhold judgment until there is more evidence?

ARE THE CONCLUSIONS BACKED BY BELIEVABLE STATISTICAL EVIDENCE?
What is the degree of uncertainty? How sure can you be? Could these results have occurred by chance?

To whom do the results apply? Who can you generalize to?

Did the investigator frankly discuss possible biases or flaws in the study?

Have the results been reviewed by unbiased parties?

Do the results make sense?

SOME SLIPPERY STATISTICS

(Nancy Spruill, Post)

1. The Everything's Up Statistic
   Uses numbers rather than rates.

2. The Best Foot Statistic
   Choose what fits your story: median vs mean; year of comparison

3. The Half Truth Statistic
   Statistic based on special subgroup

4. Anecdote statistic

5. Everyone is averge statistic

6. Coincidence statistic

7. Meaningless statistic: e.g. "overall cleanliness of NY streets up from 56 to 85 % in last 5 years"

8. Unknowable statistic

This page intentionally left blank.

Listing of NCES Working Papers to Date

Please contact Ruth R. Harris at (202) 219-1831
if you are interested in any of the following papers

Number	Title	Contact
94-01 (July)	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02 (July)	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03 (July)	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04 (July)	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-05 (July)	Cost-of-Education Differentials Across the States	William Fowler
94-06 (July)	Six Papers on Teachers from the 1990-91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
94-07 (Nov.)	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
95-01 (Jan.)	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02 (Jan.)	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03 (Jan.)	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95-04 (Jan.)	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
95-06 (Jan.)	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-08 (Feb.)	CCD Adjustment to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09 (Feb.)	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10 (Feb.)	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11 (Mar.)	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12 (Mar.)	Rural Education Data User's Guide	Samuel Peng
95-13 (Mar.)	Assessing Students with Disabilities and Limited English Proficiency	James Houser
95-14 (Mar.)	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15 (Apr.)	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16 (Apr.)	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17 (May)	Estimates of Expenditures for Private K-12 Schools	Stephen Broughman
95-18 (Nov.)	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01 (Jan.)	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-02 (Feb.)	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96-03 (Feb.)	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
96-04 (Feb.)	Census Mapping Project/School District Data Book	Tai Phan
96-05 (Feb.)	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96-06 (Mar.)	The Schools and Staffing Survey (SASS) for 1998-99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96-07 (Mar.)	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96-08 (Apr.)	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96-09 (Apr.)	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998-99 SASS	Dan Kasprzyk
96-10 (Apr.)	1998-99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96-11 (June)	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96-12 (June)	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96-13 (June)	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96-14 (June)	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-15 (June)	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96-16 (June)	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96-17 (July)	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
96-18 (Aug.)	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
96-19 (Oct.)	Assessment and Analysis of School-Level Expenditures	William Fowler
96-20 (Oct.)	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-21 (Oct.)	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96-22 (Oct.)	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96-23 (Oct.)	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96-24 (Oct.)	National Assessments of Teacher Quality	Dan Kasprzyk
96-25 (Oct.)	Measures of Inservice Professional Development: Suggested Items for the 1998-1999 Schools and Staffing Survey	Dan Kasprzyk
96-26 (Nov.)	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96-27 (Nov.)	Intersurvey Consistency in NCES Private School Surveys for 1993-94	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-28 (Nov.)	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
96-29 (Nov.)	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
96-30 (Dec.)	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-01 (Feb.)	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk
97-02 (Feb.)	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-03 (Feb.)	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97-04 (Feb.)	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-05 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-06 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-07 (Mar.)	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-08 (Mar.)	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-09 (Apr.)	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97-10 (Apr.)	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993-94 School Year	Dan Kasprzyk
97-11 (Apr.)	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97-12 (Apr.)	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97-13 (Apr.)	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
97-14 (Apr.)	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97-15 (May)	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97-16 (May)	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97-17 (May)	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
97-18 (June)	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97-19 (June)	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97-20 (June)	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97-21 (June)	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed