DOCUMENT RESUME

ED 279 726 TM 870 175

AUTHOR	Russ-Eft, Darlene F.
TITLE	Validity and Reliability in Survey Research. Technical Report No. 15.
INSTITUTION	American Institutes for Research in the Behavioral Sciences. Palo Alto, CA. Statistical Analysis Group in Education.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
PUB DATE	Aug 80
CONTRACT	300-78-0150
NOTE	143p.
PUB TYPE	Reports - Descriptive (141) -- Reference Materials - Bibliographies (131)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Concurrent Validity; Construct Validity; Content Validity; Data Collection; Educational Research; Questionnaires; *Research Design; *Research Methodology; Research Problems; *Surveys; *Test Reliability; *Test Validity
IDENTIFIERS	Internal Consistency; *Test Retest Reliability

ABSTRACT

With increasing reliance being placed on the results of their surveys, the National Center for Education Statistics (NCES) recognized a need for these survey results to be valid and reliable. As part of the work of the Statistical Analysis Group in Education (SAGE), an effort was undertaken to investigate validity and reliability in survey research. This document is the result of that effort. The first section provides an overview of the concepts of reliability and validity. Procedures for measuring sources of error are suggested. Several ways of approaching validity are mentioned including content validity, criterion-related validity, and construct validity. The second section presents some suggestions for approving the reliability and validity of survey data, with a focus on the data collection phase. These suggestions are based upon previous data collection experiences of the staff of SAGE. The final section, encompassing 106 pages, provides an annotated bibliography of selected materials relevant to validity and reliability in survey research. (JAZ)

Technical Report No. 15

Validity and Reliability in Survey Research

Darlene F. Russ-Eft

Prepared by

SAGE
STATISTICAL ANALYSIS GROUP IN EDUCATION

For the

National Center for Education Statistics

American Institutes for Research

Box 1113, Palo Alto, California 94302

TECHNICAL REPORT NO. 15

VALIDITY AND RELIABILITY IN SURVEY RESEARCH

Darlene F. Russ-Eft

Statistical Analysis Group in Education
American Institutes for Research
P.O. Box 1113
Palo Alto, California 94302

This work was done under Contract No. 300-78-0150 with the National Center for Education Statistics, Department of Health, Eudcation, and Welfare. However, the content does not necessarily reflect the position or policy of either agency, and no official endorsement should be inferred.

August 1980

Table of Contents

List of Tables

Preface

This report was prepared for the National Center for Education Statistics (NCES) as part of the work of the Statistical Analysis Group in Education (SAGE). The original SAGE effort was to focus on problems of validity and reliability in survey research. In addition to the preparation of an annotated bibliography on the topic, the SAGE staff were to work with selected NCES survey managers to design a guidebook on survey procedures. Before the annotated bibliography was completed, some critical needs in the area of imputing missing data arose at the Center requiring increased attention of SAGE staff. As a result, the work on survey validity and reliability was substantially reduced. The SAGE staff recognized, however, the Center's desire for some useful product in this area and proceeded to provide the available annotations, as well as some recommendations for survey researchers.

The report is divided into three sections. The first section provides a very basic overview of the concepts of reliability and validity. Included in this section is a discussion of some procedures that may be used to assess the threats to validity and reliability. The second section presents some suggestions for improving the reliability and validity of survey data, with a focus on the data collection phase. These suggestions are based upon previous data-collection experiences of SAGE staff. The final section provides an annotated bibliography of selected materials relevant to validity and reliability in survey research. Some readers may prefer to skip the first two sections and concentrate on the annotated bibliography.

If funds become available in the future to complete the original SAGE task, the next steps would be to link the overview, the discussion of assessment methods, and the procedural recommendations more directly to the Center's major survey efforts. The specific procedural suggestions should focus on all phases of the survey research effort, including the development of error assessment and error correction techniques.

The author wishes to thank the following persons for their contributions to this effort. Dr. Ronald Abeles helped to identify relevant articles. Dr. William Clemans reviewed a draft of the report and provided critical comments. Ms. Sibyl Anderson supervised the typing and editing of the manuscript.

Introduction

Important questions need to be answered concerning the current status of our nation's schools. The answers will provide information to critical decisions that need to be made concerning the future of our educational system, and that information will be useful only to the extent that the data are reliable and valid. Some typical questions posed by policymakers are:

- To what extent is the concept of equality of educational opportunity a reality? How can equality of educational opportunity be achieved?
- How serious is the current financial crisis facing the schools? Have school costs increased more rapidly than inflation? What is the relationship between revenue availability and demands for educational services?
- What is the current status of vocational education in our country? What form of federal support should be provided to such programs?
- What federal support is being received by private schools in the United States? Should this support be increased, and if so, how can this be accomplished most efficiently and effectively?

For over a century, the United States government has seen a need to collect and report data on the condition of American education. In 1870, three years after creation of the U.S. Office of Education, Congress granted the Commissioner three additional clerks and $3,000 for work in computing statistics and preparing reports. For over a century since that date, the Office of Education has continued to compile statistics on the nation's schools to inform Congress and the public about the status of the school system that they support. Since 1965, with the advent of large-scale federal aid for education and the accompanying need for assessments of relative educational needs, resources, and outcomes across the country, the need for valid and reliable gathering and reporting of school data has increased so rapidly that it threatens to clog the schools with paperwork. In 1970, after the failure of the so-called Belmont System to streamline school reporting, the responsibility for data gathering has been steadily shifted to the National Center for Education statistics. In 1974, as a part of P.L. 93-380, Congress established the purpose of NCES to be to "(1) collect, collate, and, from time to time, report full and complete statistics on the conditions of education in the United States; (2) conduct and publish reports on specialized analyses of the meaning and significance of such statistics; (3) assist State and local educational agencies in improving and automating their statistical and data collection activities; and (4) review and report on educational activities in foreign countries."

With increasing reliance being placed on the results of their surveys, NCES recognized the need for these survey results to be valid and reliable. Thus, as part of the effort of the Statistical Analysis Group in Education (SAGE), an effort was undertaken to investigate validity and reliability in survey research. Unfortunately, changing needs at the Center dictated that this effort be substantially reduced in scope.

The following report is the result of the reduced effort. It provides a very brief summary of the problem and suggests some methods for improving the validity and reliability in survey data. In addition, the report includes an annotated bibliography of some publications in this area. The annotations are lengthy and provide the reader with the major points as well as an overview of each article and book that is reviewed.

Validity and Reliability

Research indicators, data elements, or instruments must meet certain criteria before they can be considered effective and useful. In discussing considerations underlying the development of youth-specific social indicators, Rossi and Gilmartin (1980) identified and described several dimensions or criteria important for such social indicators. These authors then examined each dimension or criteria in relation to the role that an indicator might serve in a model (i.e., input, process, disposing condition, or outcome). They concluded that the importance of each criteria varies according to the indicator's role. If an indicator serves in more than one role (e.g., as an input for some analyses and as an outcome for other analyses), the most stringent criteria should be applied. After describing these criteria, the authors presented some details on the development of indicators.

Table 1 presents a summary statement on criteria to be applied to indicators as specified by Rossi and Gilmartin (1980). Several aspects of this table should be noted. First, the dimensions or criteria are divided into two categories: (1) quality of the data (validity, reliability, stability, responsiveness, availability, and scalability), (2) relation to other variables (disaggregatability, representativeness, and overlap with other indicators), (3) breadth of comparability (intertemporal comparability, intergroup comparability, and breadth of application), and (4) usefulness (understandability, normative interest, policy relevance, timing relative to the occurrence of a problem, and timeliness).

The table points out that certain dimensions deserve special attention when longitudinal indicators are being employed. Longitudinal indicators are those that will be, or could be, used to gather measures over time. By measuring certain conditions and changes in these conditions over time, information is provided concerning long-term trends, periodic changes, and fluctuations in the rate of change. A well-known example of this type of indicator is the Gross National Product (GNP).

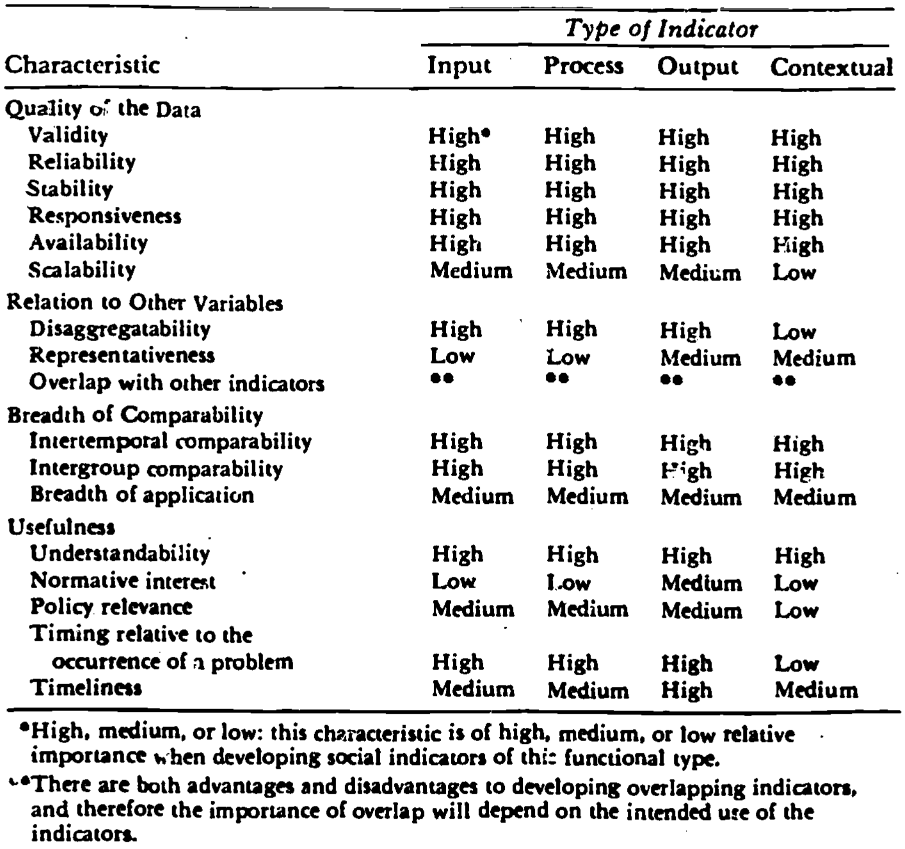
Finally, as can be seen in the table, the greatest number of constraints are placed on outcome indicators. Fewer constraints are applied to input and process indicators, and the least constraints are placed on contextual variables. Reliability and validity are viewed as of high importance to any measurement effort. Reliability is discussed first, because it is necessary to any consideration of validity.

Reliability

Reliability is a criterion important to all research indicators and instruments. Reliability refers to the degree of accuracy or consistency

Table 1

Relative Importance of Characteristics for Various Functional Types of Indicators



Characteristic	Type of Indicator			
	Input	Process	Output	Contextual
Quality of the Data				
Validity	High*	High	High	High
Reliability	High	High	High	High
Stability	High	High	High	High
Responsiveness	High	High	High	High
Availability	High	High	High	High
Scalability	Medium	Medium	Medium	Low
Relation to Other Variables				
Disaggregatability	High	High	High	Low
Representativeness	Low	Low	Medium	Medium
Overlap with other indicators	**	**	**	**
Breadth of Comparability				
Intertemporal comparability	High	High	High	High
Intergroup comparability	High	High	High	High
Breadth of application	Medium	Medium	Medium	Medium
Usefulness				
Understandability	High	High	High	High
Normative interest	Low	Low	Medium	Low
Policy relevance	Medium	Medium	Medium	Low
Timing relative to the occurrence of a problem	High	High	High	Low
Timeliness	Medium	Medium	High	Medium

*High, medium, or low: this characteristic is of high, medium, or low relative importance when developing social indicators of this functional type.

**There are both advantages and disadvantages to developing overlapping indicators, and therefore the importance of overlap will depend on the intended use of the indicators.

Note. From Handbook of Social Indicators by Robert J. Rossi and Kevin J. Gilmartin. Copyright 1980 by Robert J. Rossi and Kevin J. Gilmartin. Published by Garland STPM Press.

that an indicator measures whatever it measures. Simply stated, reliability assesses the degree to which an indicator will yield the same results on repeated application. It is a function of error variance in the measurement (i.e., the variability that is unrelated to changes in the actual status of the phenomenon being measured). Defined as the proportion of the indicator's variance that is "true" variance (or is not error variance), it can range from .00 (i.e., completely unreliable) to 1.00 (i.e., completely reliable).

Several methods exist for estimating the reliability of an indicator. These methods measure different sources of error. For example, reliability coefficients resulting from a single administration of a test, such as split-half reliability, do not measure response variability over time as a source of error. It is, therefore, recommended that, for longitudinal indicators, the test-retest or repeated measurement method be used for indicators composed of a single measure (as opposed to a weighted composite of many different measures). Much care should be exercised, however, because variability in the measurement with the passage of time may reflect actual changes in the status of the measured conditions. Another method for estimating reliability when conducting survey research involves the use of internal consistency checks. For example, in the Project TALENT survey, the student was asked to indicate the number of older brothers (SIB question 222) and the number of older sisters (SIB question 223) as well as the total number of brothers and sisters (SIB question 224). Thus, a consistency check can be made to determine whether the student responded that he or she had more older brothers and sisters than the total number of brothers and sisters.

Validity

Validity is considered to be the most important dimension of a research instrument, because it indicates the extent to which the instrument measures what it was intended to measure. If an instrument is not valid, it has little or no value for the task at hand, and, further, the researcher literally cannot be certain as to what is being measured. It should be noted that the validity of an indicator does not depend on a single measure. This will become evident in the following discussion on the various types of validity. Although correlation coefficients can be calculated to examine the validity of an indicator, validity can only be inferred.

The American Psychological Association's Standards for Educational and Psychological Tests (1974) indicates that validity can be approached in three ways. First, a research instrument should be examined for content validity. Content validity measures the extent to which items in the instrument reflect the purpose of the data collection effort. Critical for the process of examining content validity of the measure or indicator are three operations: (1) definition of the domain of concern, (2) specification of the objectives, and (3) development of a method for constructing or sampling items. (The third operation is most important in the development of test items. Some recommended procedures for the development of survey questionnaire items appear in the last section of this paper.) The process of establishing content validity involves evaluating the model developed by the operations, examining the characteristics of potential indicators, and deciding subjectively whether they are reasonable. It is this process that differentiates content validity from face validity. Content validity focuses on certain operations and is assessed in terms of the thoroughness by which these operations are accomplished; face validity is merely a judgment that the indicator appears to be relevant to the survey objectives. Some researchers may believe that "single item" measures such as ADA (average daily attendance) or number of faculty members have impeccable face validity. It is, however, necessary to follow the procedures for content validity to avoid errors. For example, reports of number of faculty may be accurate, but they do not match the domain of concern (e.g., number of tenured faculty).

The second approach to validity concerns criterion-related validity. Criterion-related validity indicates the degree to which an indicator correlates with or predicts some observable and quantifiable condition considered to be the criterion. The correlation between the indicator under scrutiny and an independent indicator measuring the same variable provides an assessment of the criterion-related validity of an instrument. This form of validity appears most often in the context of tests used in prediction.

The third and final approach to validity involves an investigation as to whether the instrument behaves as expected according to the model being used. This investigation determines the construct validity of a research instrument. Construct validity is usually inferred from a predicted network of relationships. The concern here is whether the indicator actually measures some construct, such as "reading readiness" or "limited English language ability."

The previous discussion has focused on validity as characteristic of instruments and indicators. Validity can be viewed in a different fashion when considering the survey or information-gathering effort as a whole. In their classical chapter in the Handbook of Research on Teaching, Campbell and Stanley (1963) define the terms "internal validity" and "external validity" and propose experimental and quasi-experimental designs to control for the threats to the validity of the research.

The internal validity of an information-gathering effort is defined as the extent to which it actually (correctly) answers the questions that it claims to answer. All data collection and analysis is carried out in the context of a model, or set of assumptions, about the process or phenomenon being observed. If those assumptions are wrong, the findings of the survey are meaningless. If those assumptions are correct, the survey results are internally valid, and the findings from the survey are meaningful. The main category of threat to internal validity is that unmeasured processes might account for the results that were observed. A second category of threat is that overt responses do not correctly reflect underlying dimensions. To achieve internal validity, a substantial level of control must be obtained over the information-gathering process.

The primary operations for obtaining internal validity are the operations that constitute the scientific method. If the research design calls for the assignment to treatment and control groups, that assignment must be random. The identification, measurement, and control of confounding variables is a second critical set of operations. Finally, the use of the multimethod approach provides for converging evidence in support of a particular finding. If a number of different methods of measurement and analysis all produce similar results, one is more likely to accept the result as being true rather than as being dependent upon the particular method that was used.

The external validity of an information-gathering effort is defined as the extent to which answers based on the observations correctly generalize to other unobserved situations. Generally, external validity is determined by sample selection, whereas internal validity is determined by sample assignment to treatment and control groups. To the extent that internal validity is achieved through obtrusive control, external validity will be reduced. Thus, rigorous experimental designs with artificially contrived situations may decrease the external validity or the generalizability of the findings.

Increasing external validity depends upon drawing a representative probability sample. To the extent that nonresponse occurs in the survey population or sample, bias may enter, decreasing the external validity. Poorly worded questions and poorly designed questionnaires can decrease the accuracy of the survey through nonresponse or incorrect response. Cross-validation methods should be used to check on external validity, and efforts to achieve internal validity should avoid obtrusive control.

Threats to Validity and Reliability in Survey Research

Much of the formal work on validity and reliability has focused on test development (e.g., Armor, 1974; Bohrnstedt, 1969; Cronbach, 1951; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; Cureton, Cook, Fischer, Laser, Rockwell, & Simmons, 1973; Kaufman, 1977; Lord, 1974; Nunnally, 1967; American Psychological Association, 1974; Subkoviak & Levin, 1977). The concern for validity and reliability of measurement exists in survey research efforts as well. A few examples of some efforts to examine and deal with the problems of reliability and validity in survey research are described below.

Withey (1954) investigated the reliability of reports on income. His research focused on whether annual income is reliably recalled over a one-year period and whether the direction of the recall errors is systematic. He found that the income reports were affected by several things. First, the information was of a personal nature and subject to distortion. Second, income was received from several sources, and this resulted in confusion. Third, the fact that incomes are shared created response problems. The evidence indicated that recall concerning income was not accurate or random, and the correlation between the magnitude of income change was positively related (.58) to the magnitude of unreliability. In other words, as income change increased, the unreliability of the recall report also increased. Ferber (1966) obtained similar results in a study of the reliability of reports on demand deposits. In particular, very small accounts were overestimated and very large accounts were underestimated. These studies indicate that, even when reporting "factual" information, errors can exist in the data; however, approaches do exist for assessing the quality of survey data.

The ability of a measure can vary widely with the ability and maturity of the population being assessed. Data collected from students can be significantly less reliable than similar data collected from adults; similarly, data collected from adults in a professional capacity (e.g., school administrators) can be significantly more reliable than data collected from a general sample of adults. Data from Project TALENT, for example, indicated a significantly higher rate of "spurious" responses to biographical questionnaire items among 9th- and 10th-grade males than among females or 11th- and 12th-grade males. This difference is evident in inconsistency rates in items concerning number of siblings (e.g., total number of brothers and sisters as compared with the number of older brothers and sisters) and in the overall rates of reported twins that were significantly above known national rates for the 9th- and 10th-grade males. The reasons for lower reliablity rates in this group include (1) a greater proportion of respondents who did not understand the question (e.g., did not know what a twin is; (2) a small proportion who appeared to be answering randomly; and (3) another small proportion who intentionally gave false responses. By developing an index of infrequent and inconsistent responses, it was possible to screen out, post hoc, unreliable respondents to key items. After this screening, the proportion of twins agreed with known rates for all groups.

Zarkovich (1966) presented an overview of the quality problem of statistical survey data for the Food and Agriculture Organization of the United Nations. He grouped the various kinds of errors into three major types: (1) errors resulting from inadequate preparations, (2) errors committed during data collection, and (3) processing errors. He then presented a detailed discussion of the types of errors. To reduce errors and to examine the quality of the data, Zarkovich proposed two approaches: (1) the use of post hoc techniques and (2) the use of sampling methods. In the first approach, the researcher can compare the results with data from independent sources, can compare the results with some generally accepted knowledge about the characteristics or their relationships (i.e., consistency checks), can examine the internal consistency or the degree to which estimates of different characteristics describe the same phenomenon in the same way, and can determine the expected attrition. Unfortunately, post hoc techniques have several limitations: (1) independent sources may differ in the units covered and/or in the precise item definition; (2) data from independent sources may be subject to error; (3) previously collected data on the same topic may not be available; (4) application of such techniques results in impressions about the errors or about the quality rather than in numerical measures; and (5) techniques refer to final survey results and, thus, individual errors are lost. Quality measurements based on sampling methods can overcome many of these problems: (1) they can be applied whether or not accurate data have been collected previously on the topic; (2) they can provide numerical estimates; and (3) they can provide estimates of the quality for any part of the population and for any item in the survey.

In the National Longitudinal Study of the 1972 high school graduating class, Conger and his associates (1976) discussed procedures for conducting reliability and validity pilot studies prior to the main survey. Based on their findings, the authors stated that generalizing results across populations with different respondent characteristics is highly problematic. They suggested that demographic variables should be used not only as control variables but also as moderating variables when analyzing survey data. The authors also stated that path analyses can be conducted with factually oriented data with high ability or high SES respondents, otherwise structural modeling with error measurement estimates based on a similar population is required. For these reasons, Conger and his associates recommended conducting pilot studies prior to the main survey to permit modifications to improve its reliability and validity.

The notion of using multiple indicators with structural equation modeling is an approach that can be used in survey research (Blalock, 1969; Heise & Bohrnstedt, 1970). It does, however, require the development of a structural equation model (either recursive or nonrecursive), the identification of several indicators of the same concept, and the measurement of those indicators. Analysis of these indicators within the structural equation model provides some assessment of the validity and reliability of the measurement and the model.

An approach to examining the quality of the data has recently been proposed by the Subcommittee on Nonsampling Error of the Federal Committee on Statistical Methodology. It takes advantage of alternative approaches, such as those discussed above, and integrates and summarizes them into an error profile. Thus, the survey research team constructs an error profile by examining possible sources of error arising in each phase of the survey operation. Under the auspices of the Subcommittee on Nonsampling Errors, Brooks and Bailar (1978) of the U.S. Bureau of the Census prepared an error profile of employment statistics as measured by the Current Population Survey. In this error profile, the authors reviewed

- the sampling design and implementation (including the sampling frame, the sample selection procedure, and the quality control of the sampling procedure);

- the observational design and implementation (including the data collection procedure--listing sampling, and interviewing, the questionnaire design, the data collection staff, the interviewer training, and the quality control of the field work);

- the data processing (including the FOSDIC/Microfilming procedure, the editing and imputation, and the quality control of data processing);

- the estimation procedures (including the weighting procedure, the production of estimates, the specification of errors for the estimates, and the quality control of the estimation procedure);

- the analysis and publication.

Within each facet of the data collection and presentation effort, they described the specific procedures that were followed and indicated potential sources of error. Where relevant studies of these error sources existed, the studies were presented along with some indication of the impact of this error on the resulting statistic.

Improving the Validity and Reliability in Surveys

Various categorizations of threats to the reliability and validity of data appear in the literature (e.g., Webb, Campbell, Schwartz, & Sechrest, 1966; Berdie & Anderson, 1974; NCES, 1976). As an example, some threats to validity provided in a previous document by the Center are summarized in Table 2.

In dealing with problems of reliability and validity, such threats as those presented in Table 2 should be considered within each functional phase of survey design and management. Indeed, this is the major point in the approach taken through the conduct of an error profile. These functional phases are (1) selection of survey objectives, (2) selection of a respondent sample, (3) instrument development, (4) respondent contact and data collection, (5) data transfer and editing, (6) data analysis, and (7) reporting. The following sections provide some guidelines for improving validity and reliability, focusing on Phase 3--instrument development and Phase 4--respondent contact and data collection. The rationale for this focus is that other SAGE tasks have addressed problems in each of the other phases. (Task 1 of SAGE focused, somewhat obliquely, on Phase 1--selection of survey objectives; portions of Task 2c concerned problems related to Phase 2--sample selection; Task 2d addressed problems in Phase 5--data transfer and editing; Task 5 focused on a key threat to validity in Phase 6--imputation of missing data prior to analysis; and portions of Task 2b addressed problems related to Phase 7--reporting.)

Instrument Development

Instrument development poses key questions that need to be answered for survey designers and managers: How can the survey items be worded to ensure that responses have the meaning intended for them? How can the accuracy of the survey responses be assessed? How can the instrument be refined to reduce response burden and to increase response rate without losing valuable information? The following paragraphs will attempt to provide some answers to these questions. Lipset (1976) in "The Wavering Polls" discusses the lack of reliability of most survey research results. Examples are cited from popular polls such as Gallup, Harris, and California to show the differing response rates to similar questions. The power of the word in effecting response cannot be taken lightly. Response rates have been found to vary by as much as 20 percentage points when the words "should," "could," and "might" were interchanged in a sentence. A change in wording in one census survey brought an increase of 1,400,000 responses (Payne, 1951). In his book The Art of Asking Questions, Payne (1951) states that "the most critical need for attention to wording is to make sure that the particular issue that the questioner has in mind is the particular issue on which the respondent has given his answers." A checklist of 100 considerations to be used in wording a question is contained in the final chapter of this book. Payne stresses the importance of not "talking down" to the respondent, the use of common sentence construction in questions, the avoidance of overelaboration and trick questions, and the right of the respondent to refuse to answer a question. Duckworth

(1973) highly recommends that individuals who will be wording questions to be used in a survey (1) gain some knowledge of the topic area, (2) develop the questions with help of subject-matter specialists, (3) devise more than one question dealing with a specific subject matter topic, and (4) maintain a careful record of all reviews of a question.

Question types. When constructing a new instrument, the first decision involves choosing the type of question that will be asked. Several question types can be used, but this discussion will be limited to the most frequently used types. These question types include open-ended, two-choice, and multiple-choice. Four factors need to be considered when choosing from among these types:

- Who will be answering the questions?
- How much time will these respondents be willing to spend?
- How many respondents will be involved?
- How much is known about the range of possible answers?

Each of the question types will now be considered.

a. Open-ended questions. An example of an open-ended is:

What is the chief goal of your education program?

An open-ended question is easy to construct and ask. When uncertainty exists as to the entire range of alternative answers, the use of an open-ended question may be warranted. The major disadvantage of using open-ended questions involves time for both the respondent and the evaluator. Particularly if the answer must be written, the respondent may not be willing to devote the needed time and effort to provide a complete response. From the standpoint of the researcher, open-ended questions require extensive time and effort for coding and analysis. With a large sample size, analysis of such questions may become a major problem. Another disadvantage in using open-ended questions is that the respondent may misinterpret the question or focus on an irrelevant aspect, because a set of choices is not available to guide the response. (For such reasons, open-ended questions should be avoided with respondents who have less than an eighth-grade education.)

b. Two-choice questions. Such questions permit the choice between straightforward alternatives, as in the following example:

Does your school receive Title I funds?
1. Yes
2. No

Unlike the open-ended question, the two-choice question permits a rapid response and provides for fast economical data analysis. Some of the disadvanges of this type of item are that (1) dichotomous responses convey limited information, (2) respondents may choose an answer based on response set (e.g., some people may tend to choose the last alternative or the affirmative); and (3) the alternatives may not be clear cut.

c. Multiple-choice questions. As with the two-choice type, multiple-choice questions can be answered and analyzed quickly and easily. Thus, they are efficient when a large sample size is being considered. To obtain valid information, however, such questions and their alternatives must be carefully worded. Furthermore, the full range of alternatives must be presented. The question writer should pose the question in an open-ended fashion to a few respondents in advance. Alternative options for the multiple-choice question to be included in the final survey can be selected from the open-ended responses. It is sometimes useful to include an open-ended alternative, as in Option 5 in the following example:

Who paid for this course or activity?
1. Self or family
2. Employer
3. Public funding
4. Private organization (church, professional organization)
5. Other (Describe) ______________________________
6. Don't know

Item frame. The next step involves the actual writing of the question. To obtain valid information from respondents, the question must be worded carefully. The following are some suggestions and cautions that can lead to improved item writing:

- Use terms that respondent will understand. Although some terms may seem to be common, they may not be part of the vocabulary of the respondents. Attention to wording problems both in writing and in pilot testing items will result in the collection of more valid information.*

- Do not confuse the respondent with tricky or ambiguous wording. An easy way to create problems for the respondent is to use confusing wording. With the double negatives in the following example, the item writer has created a question that respondents may find difficult to answer.

    Are you against not requiring mandatory busing of elementary school students?

    1. Yes
    2. No

- Avoid wording that suggests answers or "loads" responses in one direction. The following example clearly suggests an answer.

    Isn't it true that high school vocational training should be improved?

    1. Yes
    2. No

    This problem may be more subtle, as in this next example (which is a variant of an item from an actual survey). Here we can see that the writer believes that at least one ethnic minority person should be on the museum's board of directors.

    How many members of your museum's board of directors are ethnic minority persons?

    1. 1 person
    2. 2 persons
    3. 3 persons
    4. 4 or more

    (No option was included for "no persons.")

*In Project TALENT, the accurate reporting of being a "twin" correlates -.39 with academic aptitude.

- Avoid asking questions that presuppose a certain condition. An example from Berdie and Anderson (1974) illustrates this point.

  Do you still design bad questionnaires?

  1. Yes
  2. No

- Screen the question for any "dead giveaway" words. In the example below, the wording will lead the careful respondent to answer in the negative and the less careful one to respond in the positive.

  Do you feel that your volunteer aides did their best in all project work during the past month?

  1. Yes
  2. No

  One may first ask whether any group of aides can do its actual best, except on rare occasions. In addition, the word "all" makes this question even more difficult to answer positively.

- Avoid asking respondents to make undesirable choices. As an example, consider the fact that most people are reluctant to criticize. A question that allows respondents to avoid criticism will lead to biased responses. In the following example, many respondents may simply avoid criticism by answering in the negative.

  Has your project ever violated state regulations and guidelines?

  1. No
  2. Yes

  If yes, in what way? ____________________

Some other guidelines to follow include:

- Make the question and any choices as short and simple as feasible.

- Do not ask the respondent to perform many activities in one question. (Use a separate item for each distinct bit of information desired.)

- Avoid using words with vague meanings, such as "country" (whose country?) or "population" (how much of which population?).

- Do not ask for fine distinctions, unless you are certain that they will be meaningful to the respondents.

Response options. For two-choice or multiple-choice questions, attention must be given to the response options. Confusing options may lead to unreliable results and low response rates. Response options must be examined according to the following guidelines.

- Include one option for every possible response.
- Include a "don't know" option whenever a respondent may be unable to answer.
- Construct response options that are mutually exclusive and independent.
- Include an equal number and degree of options on each side of a middle position for all rating scales.

Format considerations. As in developing the items, it is important to consider the respondent population. To keep the form both simple and interesting, the following suggestions should be addressed.

- Begin with interesting and nonthreatening questions.
- Group items into logical sections--by topic area.
- Include smooth transitions between sections.
- Avoid putting the most important questions at the end.
- Arrange response options vertically rather than horizontally.

For questionnaires, several other considerations should be kept in mind.

- Include an attractive cover.
- Number the items and pages to reduce confusion for the respondent.
- Include clear, brief instructions for completing the questionnaire items. If necessary, provide examples.
- Place an identification code on each page of the form so that if pages get separated they can be identified.

Finally Berdie and Anderson (1974) recommend that certain guidelines be followed for questionnaires being sent directly to be keypunched.

- Group items that have the same response options.
- Place spaces for answering questions in columns so that the keypunch operator will not have to search for the response.
- Place card column numbers near the answer spaces.

Internal Screening Scales

One of the unfortunate problems associated with large-scale surveys is that some small percentage of the respondents cannot or will not respond reliably to the questions posed to them. For some analyses, it may be preferable to eliminate these unreliable respondents from the sample. To this end, a screening scale can be included in the instrument or can be constructed from appropriate items in the questionnaire.

One approach is to develop a screening device to eliminate persons prior to the administration of the survey. This may be used to screen for persons who are appropriate or inappropriate for the survey. For example, the Children's English Services Study used items to identify whether households were non-English speaking and had children of certain ages. Such a screening device may be used to eliminate persons who are likely to give unreliable responses. For example, some studies on the elderly use senility screens.

A second approach involves the development of a screening measure that is used after the collection of the data to identify respondents whose data are probably spurious. The method is similar to the development of the Tie Scale on the Minnesota Multiphasic Personality Test (MMPT). For example, the Project TALENT test battery included a screening scale consisting of 12 items that every student should have answered correctly (e.g., "How many months are there in a year?"). This scale was used to flag respondents with questionable data and to eliminate such cases from most further analyses.

A third approach focuses on construction of screening indexes based on the available data. The screening index can use material available from the responses of one individual at one time to determine consistency in the responses. For example, the report on the number of male students in a school plus the number of female students in that same school should equal the report on the total number of students in the school. The data could also be examined for credibility. For example, in Project TALENT, respondents who marked that their parents spoke Hebrew and an Oriental language were also likely to have provided questionable responses on other items. These same kinds of checks can be made using longitudinal data (e.g., reports on school size for the prior year) or using relational data (e.g., calculating pupil/teacher ratios using reports on the number of pupils and the number of teachers). In such cases, the researcher might have to develop a measure of change and a measure of dispersion and to select fences for outliers. Fingerman (1980) provides greater detail on longitudinal and relational editing.

Pilot Testing the Items and Instruments

Even though the items have been carefully examined to eliminate major problems, the items and instruments must be pilot tested to identify problems that may have been overlooked. Respondents selected for pilot testing should be representative of the eventual target sample. If nine or fewer persons from each of the respondent groups are used, forms and procedures for the pilot testing will not have to undergo OMB clearance.

Pilot testing should be conducted by the question writer or someone else who is knowledgeable about the subject area. During pilot testing, the respondent should be asked to comment on the content and wording of the questions. Three useful techniques involve (1) asking the respondent to read and explain the question, (2) asking him or her to explain the reasons for his or her choice, and (3) asking him or her for other answers that could be given. These probes may reveal incorrect assumptions or alternative rationales that were never anticipated.

The survey staff should revise the items using the results of the pilot testing. If necessary, further pilot testing and revision should be undertaken.

Respondent Contact and Data Collection

In respondent contact and data collection, the key questions are (1) What are the most efficient and effective methods for maximizing response rates? (2) What procedures should be avoided because they affect the meaning that can be attached to item responses? (3) How can threats to the survey's validity and reliability be ascertained? The following paragraphs will present some answers to these questions.

Data collection procedures basically differ along two dimensions: (1) the degree of structure in the questions (e.g., closed versus open-ended questions) and (2) the amount of researcher contact with the respondent (e.g., individual interviews, telephone interviews, group-administered questionnaires, and self-administered questionnaires [Herriot, 1969]). For each type of data collection procedure, specific requirements exist that should be met to ensure a high level of validity and reliability.

The most common form of large scale survey conducted today is the mail survey. The major advantages to a mail survey are wider distribution, less distribution bias in connection with the individual, no interviewer bias, better chance of a truthful reply, better chance of a thoughtful reply, time saving (under certain circumstances), centralized control, and cost saving (Erdos, 1970). To ensure that a mail survey be properly and soundly designed, the following steps are proposed in Professional Mail Surveys (Erdos, 1970):

1. Outline the problem.
2. Define the research objectives.
3. Investigate existing research on the same problem or with the same objectives.
4. Define the universe.
5. Decide on the degree of reliability aimed at within a realistic budget.
6. Define the sample and scope.
7. Decide on the survey method.
8. Decide who will conduct the survey.
9. Establish the techniques that will be needed to achieve the research objectives.

10. Outline the type of tabulation, analysis, and report desired.

11. Make up a time schedule.

12. Make up a cost estimate.

It recommended that each survey effort follow these steps.

Pilot Testing the Data Collection Procedures

One of the most useful aids in making survey design decisions is the pilot study. The pilot study can be used tc check the percentage of returns, to check the occurrence of bias resulting from the wording of the questionnaire, to check how well questions are understood and answered, to check the usefulness of information received, and to check or even establish a cost estimate. To be effective, a pilot study must meet the following criteria: (1) it must be conducted among a random sample of the universe surveyed, and (2) the number of questionnaires mailed should by large enough to provide meaningful results (Erdos, 1970).

Increasing the Response Rate for Mailed Questionnaires

In any research or evaluation study that must depend upon a mail survey, the researcher must face the problems posed by a less-than-100% response rate. The surest way to minimize the problem is to use whatever methods are available and reasonable (given financial and time constraints) to increase the response rate. The following guidelines briefly outline methods that have been shown to have at least some positive effect on response rate and that may be applicable to NCES surveys. (See Erdos, 1979 and Scott, 1961 for further suggestions.)

Follow-up reminders. This is the method that most dramatically increases the response rate. In addition to a reminder by mail or telephone, a replacement questionnaire sent following telephone calls to delinquent projects, while adding to overall costs, can also ensure a higher response rate.

Threat of follow-up. A clear message in the initial cover letter, preferably in BOLD type, mentioning the possibility of follow-up can foster more prompt replies. ("TO SAVE YOU THE ANNOYANCE OF FURTHER REMINDER LETTERS, PLEASE REPLY PROMPTLY.")

Facilitating the return of questionnaires. A minimum of folding of questionnaires, repacking, and sealing of returns should be required. Of course, stamped, self-addressed return envelopes should be included.

Postage stamps rather than franking should be used on return envelopes. Not only does this give the impression of a more personal approach, but the thought of having an unused stamped envelope lying around can increase returns.

Certified or Special Delivery mail makes the survey seem more important.

Handwritten signatures, names, and addresses, or postscripts urging a prompt reply can partially redeem an otherwise impersonal questionnaire.

Official sponsorship. It should be made clear that it is NCES and the Center program staff who are collecting this information. The information will be used to report to Congress and to the American people on the condition of education.

Short personal letter. This should convince the person of his or her importance to the survey, tell why the survey will benefit the respondent, and assure the respondent of anonymity. A message on the questionnaire has been found to be as effective as a separate letter.

Offer of a copy of the report. An offer to provide the respondent with a copy of the report or with a short summary of the results might be included in the letter. (The choice is a function of the survey and the respondents.) This demonstrates the researcher's good intentions and may elicit enough of the respondent's interest to provoke completion and return of the questionnaire.

Assurance of anonymity or confidentiality. A respondent must be assured that his or her responses will be actively protected. Also, careful thought must be given to the manner in which the identification of nonrespondents is protected.

Additional "interesting" questions, especially at the beginning, can help an otherwise uninteresting questionnaire. A higher "interest level" more than compensates for the additional length.

Difficulty of questions and length of questionnaire. Each of these factors can seriously affect the completion rate and overall return rate of a questionnaire. Complex questions requiring several readings and much consideration can affect reliability as well as completion rates. Keep the questionnaire as short and simple as possible.

Appearance of questionnaire. Questions, choices, and answers should not be crowded together, even at the expense of extra paper.

Deadlines. Care should be taken with the use of deadlines as a motivation for respondents. A description (threat) of the nonrespondent follow-up procedure, coupled with an explanation of the importance of reporting the findings in the not-too-distant future, is a better approach than implying that responses cannot be used after a certain fixed date.

A reward. This method is included last, as it may not be necessary, appropriate, or even possible to provide a financial or other incentive to respondents. If the final response rate is quite low, however, some creative thought may be given in the future to this form of incentive.

References

American Psychological Association. Standards for educational and psychological tests. Prepared by a joint committee of the American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education. Washington, D.C.: American Psychological Association, 1974.

Armor, D. J. Theta reliability and factor scaling. In H. L. Costner (Ed.), Sociological methodology: 1973-1974. San Francisco: Jossey-Bass, 1974.

Berdie, D. R., & Anderson, J. F. Questionnaires: Design and use. Metuchen, N.J.: Scarecrow Press, 1974.

Blalock, H. M., Jr. Multiple indicators and the causal approach to measurement error. American Journal of Sociology, 1969, 75, 264-272.

Bohrnstedt, G. W. A quick method for determining the reliability and validity of multiple-item scales. American Sociological Review, 1969, 34, 542-548.

Brooks, C. A., & Bailar, B. A. An error profile: Employment as measured by the current population survey (statistical policy working paper 3). Washington, D.C.: U.S. Department of Commerce, 1978.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Conger, A. J., et al. Reliability and validity of National Longitudinal Study measures: An empirical reliability analysis of selected data and a review of the literature on the validity and reliability of survey research questionnaires (National Longitudinal Study of the High School Class of 1972). Research Triangle Park, N.C.: Center for Educational Research and Evaluation, Research Triangle Institute, 1976. (ERIC Document Reproduction Service No. ED013 371)

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. Theory of generalizability: A liberalization of reliability theory. The British Journal of Statistical Psychology, 1963, 16, 137-163.

Cureton, E. E., Cook, J. A., Fischer, R. T., Laser, S. A., Rockwell, N. J., & Simmons, J. W., Jr. Length of test and standard error of measurement. Educational and Psychological Measurement, 1973, 33, 63-68.

Duckworth, P. A. Construction of questionnaires: A technical study. Washington, D.C.: Civil Service Commission, Personnel Measurement Research and Development Center, July 1973.

Erdos, P. L. Professional mail surveys. San Francisco: McGraw-Hill, 1970.

Erdos, P. L. High response rates in mail surveys. Paper presented at the annual meeting of the American Psychological Association, New York, September 1979.

Ferber, R. The reliability of consumer surveys of financial holding: Demand deposits. Journal of the American Statistical Association, 1966, 61, 91-103.

Fingerman, P. Relational/longitudinal editing for the common core of data (Technical Report). Palo Alto, Calif.: American Institutes for Research, 1980.

Heise, D. R., & Bohrnstedt, G. W. Validity, invalidity, and reliability. In E. F. Borgalla (Ed.), Sociological methodology, 1970. San Francisco: Jossey-Bass, 1970.

Kaufman, A. S. Should short-form validity coefficients be corrected? Journal of Consulting and Clinical Psychology, 1977, 45(6), 1159-1161.

Lipset, S. M. The wavering polls. The Public Interest, 1976, 43, 70-89.

Lord, F. M. The relative efficiency of two tests as a function of ability level. Psychometrika, 1974, 39, 351-358.

National Center for Education Statistics. Statistical survey of elementary schools: Development of large-scale survey 1972-74 (NCES 76-303). Washington, D.C.: U.S. Department of Health, Education, and Welfare/ Education Division, 1976.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Payne, S. L. The art of asking questions. Princeton: Princeton University Press, 1951.

Rossi, R. J., & Gilmartin, K. J. Handbook of social indicators sources, characteristics, and analysis. New York: Garland, 1980.

Scott, C. Research on mail surveys. Journal of the Royal Statistical Society, 1961, 124, 143-205 (Series H).

Subkoviak, M. J., & Levin, J. R. Fallibility of measurement and the power of a statistical test. Journal of Educational Measurement, 1977, 14, 47-52.

Webb, E., Campbell, D., Schwartz, R., & Sechrest, L. Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally, 1966.

Withey, S. B. Reliability of recall of income. Public Opinion Quarterly, 1954, 18, 197-204.

Zarkovich, S. S. Quality of statistical data. Rome: Food and Agriculture Organization of the United Nations, 1966.

ANNOTATED BIBLIOGRAPHY

ON

VALIDITY AND RELIABILITY

American Psychological Association. Standards for educational and psychological tests. Prepared by a joint committee of the American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education. Washington, D.C.: American Psychological Association, 1974.

This booklet presents standards for test use and for test manuals, and it is intended as a guide both for test developers and for test users. This abstract is confined to a discussion of the standards for reports of research on reliability and validity.

Validity is concerned with two questions: (1) What can be inferred about what is being measured by the test? (2) What can be inferred about other behavior? The first question inquires into the intrinsic nature of the measurement itself. The second inquires into the usefulness of the measurement as an indicator of some other variable as a predictor of behaviors. The interdependent kinds of validation are criterion-related validities (predictive and concurrent), content validity, and construct validity.

Criterion-related validities apply when an individual's standing on a variable called a criterion is inferred from a test score. The two types of criteria are predictive and concurrent. Statements of predictive validity indicate the extent to which an individual's future level on a criterion can be predicted from knowledge of prior test performance; statements of concurrent validity indicate the extent to which a test can be used to estimate an individual's present standing on a criterion.

Content validity is required when a test user wants to estimate how an individual will perform in the universe of situations the test is intended to represent. Definitions of the performance domain, the user's objectives, and the method of sampling are critical to claims of content validity.

Construct validity is implied when a test (or other set of operations) is evaluated in light of a specified construct. A psychological construct is a theoretical idea developed to explain or to organize some aspects of existing knowledge. Terms such as intelligence, clerical aptitude, and reading readiness are constructs.

The general principles concerning validity that should be presented in a test manual or in a research report are presented below.

1. Evidence of validity should be presented for each type of inference for which use of the test is recommended. If validity for some suggested interpretation has not been investigated, that fact should be made clear.

2. A test user is responsible for marshaling evidence to support his or her claims of validity and reliability. The use of test scores in decision rules should be supported by evidence.

Criterion-Related Validity

3. All measures of criteria should be described completely and accurately. The adequacy of each criterion should be discussed. When feasible, significant aspects of performance that the criterion measure does not reflect and irrelevant factors likely to affect it should be discussed.

4. A criterion measure should be studied, and that evidence should be presented.

5. Information on the appropriateness of or limits to the generalizability of validity information should be provided.

6. The sample employed in a validity study and the conditions under which testing is done should be consistent with recommended test use and should be described sufficiently for the reader to judge its pertinence to his or her situation.

7. The collection of data for a validity study should follow procedures consistent with the purposes of the study.

8. Statistical analysis of criterion-related validity should be reported in a form that enables the reader to determine how much confidence is to be placed in judgments or predictions regarding the individual.

9. A test user should investigate the possibility of bias in tests or in test items. The test user should try to investigate possible differences in criterion-related validity for ethnic, sex, or other subsamples that can be identified when the test is given. The results for each subsample should be reported separately or information presented that no differences were found.

10. When a scoring key, the selection of items, or the weighting of tests are based on one sample, the manual should report validity coefficients based on data obtained from one or more independent cross-validation samples. Validity statements should not be based on the original sample.

11. To the extent possible, a test user who intends to continue using a test over a long period of time should develop procedures for gathering data for continued research.

Content Validity

12. If test performance is to be interpreted as a representative sample of performance in a universe of situations, the universe represented should be clearly defined and the sampling procedures should be described.

Construct Validity

13. If an author proposes to interpret test scores as measuring a theoretical variable (ability, trait or attitude), the proposed interpretation should be fully stated, and the theoretical construct should be distinguished from interpretations based on other theories.

Reliability refers to the degree to which the results of testing are attributable to systematic sources of variance. Classical methods of estimating reliability coefficients call for correlating at least two sets of similar measurements. Four methods can be used to compute reliability coefficients: (1) test-retest, (2) administration of parallel forms of the test; (3) split-half testing; and (4) analysis of variance procedures. Each method takes different sources of error into account.

The general principles recommended concerning reliability that should be presented in a test manual or in a research report are presented below.

1. Evidence should be presented of reliability, including estimates of the standard error of measurement, that permits the reader to judge whether scores are sufficiently dependable for the intended uses of the test. If any of the necessary evidence has not been collected, the absence of such information should be noted.

2. The procedures and samples used to determine reliability coefficients or standard errors of measurement should be described sufficiently to permit a user to judge the applicability of the data reported to the individuals or groups with which he or she is concerned.

3. Reports of reliability studies should ordinarily be expressed in the test manual in terms of variances of error components, standard errors of measurement, or product-moment reliability coefficients. Unfamiliar expressions of data should be clearly described, with references to their development.

4. If two or more forms of a test are published for use with the same examiners, information on means, variances, and characteristics of items in the forms should be presented in the test manual along with the coefficients of correlation among their scores. If necessary evidence is not provided, the readers should be warned against assuming equivalence of scores.

5. Evidence of internal consistency should be reported for any unspeeded test.

6. The extent to which test scores are stable should be reported (i.e., how nearly constant the scores are likely to be if a parallel form of a test is administered after time has elapsed), and the effect of any such variation on the usefulness of the test should be described. The time interval to be considered depends on the nature of the test and on what interpretation of the test score is recommended.

Armor, David J. Theta reliability and factor scaling. In H. L. Costner (Ed.), Sociolological Methodology: 1973-1974. San Francisco: Jossey-Bass, 1974.

Reliability (in terms of internal-consistency) is usually assessed with Cronbach's alpha. It can be obtained as follows:

$\alpha = p\bar{r}\left[1 + \bar{r}\ (p-1)\right]$ where $\bar{r}$ is the mean inter-item correlation and p is the number of items.

Thus, each item in a composite is treated as a parallel variable. The coefficient incorporates much of the split-half and Kuder-Richardson coefficients, and it appears to be a lower bound to the true reliability. The method relies on the assumption that all items in a composite are parallel. Thus assumption implies that a set of real data may violate two major conditions: (1) the items may measure a single property but do so unequally; and (2) the items may measure two or more independent properties. The techniques of covariance scaling (i.e., item analysis, summated ratings, and Likert scaling) may overcome some aspects of the first problem, but they cannot assist in the second problem.

A solution to this problem is an approach to reliability based on principal-component factor analysis, and this coefficient is labeled theta. The reliability of a composite score with p items and a single factor solution is:

$\theta = \left[p/(p-1)\right]\ \left[1 - (1/\lambda_1)\right]$ where λ_1 is the first root of a principal-component solution.

For a multiple-factor solution with rotated factors,

$\theta_k^* = \left[p/(p-1)\right]\ \left(1 - \sum_{h=1}^{m} \phi_{hk}^2/\lambda_h\right)$ where ϕ_{hk}^2 is the squared correlation between the original unrotated scores for factor h and the new factor k.

Theta and alpha are equal only when all inter-item correlations are equal; otherwise, theta expresses a more accurate ratio of covariance to scale variance by weighting items according to factor loadings. Thus, a substantial difference between theta and alpha occurs only when some items have consistently lower correlations with all the remaining items in a set.

The author introduces the method of factor scaling to increase reliability. Using this method, a scale is formed by including only items that load highest on a given factor (as a rule of thumb, items with loadings below 0.3 should be excluded, and items with loading between 0.3 and 0.4 should be considered borderline). Since this procedure may produce a reliable scale for a singl: sample, the author recommends that many samples and many factor analyses be taken before deriving a final scale for wide usage.

When a set of items measures more than one dimension, two major methods exist for constructing scales and computing reliabilities: (1) basing the calculations on the rotated factor scores and (2) using factor scales based on the highest loading items. For the first method, the author suggests using a varimax rotational solution so that the factor loadings on all factors attain maximum variance. To select the number of factors to extract and rotate, the author suggests that (1) all factors with a root near to or less than 1 should be excluded; (2) a large drop from one root to another followed by gradually decreasing roots usually indicates the end of the meaningful factors; and (3) with inter-item correlations ranging from 0.3 to 0.5, a good solution will involve factors accounting for 40% to 60% of the total variance. After obtaining the rotated factors scores, a reliability coefficient can be computed for each factor by applying the general formula, theta-star (θ*). These reliabilities are not unique (since they depend on a given number of factors and a certain rotational method), and they are generally lower than those obtained by factor scaling. Applying the second method, better reliabilities can be obtained by forming scales that only use items loading highly on a given factor (i.e., above 0.40 with no higher loadings on other factors). Then, the set of items that loads highly on each factor is used to form a scale. Simple unweighted summed scales can be constructed, since relatively equal factor loadings result. Then alpha can be computed as a close approximation to theta. The major disadvantage with this method is that scales will not necessarily be orthogonal (uncorrelated). The author provides an example that illustrates that factor scaling can substantially improve scale reliability, especially with the presence of multiple dimensions.

The fact that high reliability (in terms of internal consistency) does not guarantee a meaningful construct must be recognized. However, as shown in an example, increasing the reliability of a scale can lead to increases in the validity.

The bibliography includes 16 references.

Barber, Theodore X. Pitfalls in research: Nine investigator and experimenter effects. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally, 1973.

The author begins by distinguishing the role of investigator from that of experimenter. The former is responsible for the design, conduct, analysis, and interpretation of a study, while the latter is responsible for the data collection. Misleading results and conclusions can be based on the work of the investigator, the experimenter, or both.

Four major pitfalls can be associated with investigator effects. The Investigator Paradigm Effect refers to bias introduced by an investigator because of the basic paradigm being used; thus, bias enters in terms of the questions being asked, the hypotheses, the experimental design, and the analysis interpretations. The Investigator Loose Protocol Effect refers to the degree of imprecision in experimental design and protocol; as imprecision increases (in the instructions to subjects or in the questions asked in an interview), reliability decreases. The Investigator Analysis Effect refers to biases resulting from selective focusing of analyses or selective reporting of results that confirm the investigator's hypotheses. Finally, the Investigator Fudging Effect results from a fudging of the data analyses, and it may appear in a few instances where a strong motivation exists to obtain certain results.

Turning to the role of the experimenter, the author identifies five major biasing effects. The Experimenter Attributes Effect refers to the effect of such personal attributes as sex, age, race, prestige, and social status on the subject's performance. The Experimenter Failure to Follow the Protocol Effect can result in misleading results, even when the protocol is standardized and specific; instances are cited of slight deviations in the procedures that may have led to unreliable results. The Experimenter Misrecording Effect, although rare, may lead to results biased toward the expectations of the experimenter. The Experimenter Fudging Effect has been found in many different kinds of studies and can result from deviations in the prescribed procedures, a "cutting corners" response to reduce effort, or an actual changing of the data. The Experimenter Unintentional

Expectancy Effect derives from the expectations and desires of the experimenter influencing the subject's responses.

After reviewing these various pitfalls, the author reviews several studies pointing out examples of these effects. Because of these problems, he recommends the following changes:

1. Investigators should become more explicit in presenting their assumptions and more aware of the effect of their paradigm on the research.

2. The person who plans the study should not be the one who collects the data or who analyses it.

3. Investigators should develop more detailed and specific protocols for use by the experimenters.

4. Investigators should provide experimenters with more supervised practice and should continue to check their work throughout the study.

5. Greater emphasis should be placed on the complexities of data analysis during the training of researchers.

6. Research should be judged on the validity of the design and procedures rather than on the outcomes. In addition, research training should emphasize the value of carefully following prescribed procedures and carefully and honestly recording data.

7. Investigators should use experimenters differing in personal attributes.

8. Where possible, experiments should be conducted in a "blind" fashion so that the data collector does not know what treatment the subject has received.

9. Because of all the pitfalls, results should be replicated by other researchers before they are accepted.

Bell, Roger A., Lin, Elizabeth, & Warheit, George J. Issues in need assessment data collection strategies. Paper presented at the annual meeting of the American Psychological Association, San Francisco, August 1977.

This paper discusses conceptual and methodological issues in the use of different need assessment strategies: (1) key informant, (2) community forum, (3) service utilization, (4) social indicators, and (5) citizen survey. Since the SAGE effort is concerned with survey methodology, this abstract will focus on the discussion of the citizen survey.

The first step in the conduct of a need assessment involves the development of a need model. This development should be based on inputs from three sources: (1) both lay and professional consultants concerned about the problem, (2) the general framework or environment of the problem, and (3) research data from related studies. After creating the need model, its measurable aspect must be identified. The following questions must then be asked:

- What kind of data is dictated by the model?
- Are such data presently available? If so, where? If not, what kind of procedures could obtain it? By what procedures can the most valid and accurate data be obtained?
- What kinds of practical limitations (e.g., money and time) must be dealt with?

Three major advantages of citizen surveys were discussed: (1) they provide the most direct scientifically valid and reliable information; (2) they expand the utility of other need assessment approaches; and (3) they often provide greater flexibility than other techniques. Unfortunately, such surveys pose a number of problems: the validity of the data may be questionable; the validity and reliability of the responses to survey questions may be difficult to ascertain; large surveys designed to measure prevalence or incidence may be difficult to define, conceptualize, and operationalize; and respondents may not be willing to cooperate with or support the survey effort. This last problem may be overcome by enlisting the cooperation and participation of community organizations and agencies. Multi-organization participation during the early stages of the survey development can broaden the scope of the survey to include questions that are meaningful to a large audience. In addition, these organizations can provide needed approval and legitimization for the survey.

The bibliography contains 28 references.

Berdie, Douglas R., & Anderson, John F. Questionnaires: Design and use.
Metuchen, N.J.: Scarecrow Press, 1974.

The purpose of this book is to teach individuals the basis skills necessary to design and use questionnaires. The authors believe the answers to many questions concerning questionnaire use depend on "study-specific" variables and differ from one study to another. Therefore, the goal of this book is to enumerate the issues so that the reader will be aware of them and consider them in a manner appropriate to his specific project.

The first chapter is devoted to defining the questionnaire. The authors point out that the use of questionnaires in research is based on one basic, underlying assumption: that the respondent will give truthful answers. This means that the respondent will be both willing and able to give truthful answers. Consideration of this assumption is essential throughout any study using questionnaires. People may not have the time necessary to look up requested information. Their memories may be faulty in relation to certain facts. For these reasons, questionnaire use should be limited to asking for information that is not directly available from other sources.

In the second chapter, the authors discuss the advantages and limitations of using a questionnaire in research. Some of the advantages include minimal cost, ease in collecting data from an extremely large sample, ability to cover large geographic areas, ease of completion, less bias, ease of tabulation, uniform question presentation, familiarity with questionnaire format, and structure making completion easier. The limitations include the historically low response rate (discussed in detail in Chapter V.), limited ways of checking reliability and validity, impersonalization, sample limitations, question limitations, prejudice against questionnaires, and disputed authenticity of respondent completing the form.

Other topics covered in the book include thoughts about study design, the appearance and arrangement of a questionnaire, how to stimulate response, and analyzing the results.

The annotated bibliography contains 137 entries. Included in the appendices are four sample questionnaires, a case history of a study using questionnaires, sample follow-up letters, and a sample check-off list.

Blalock, Hubert M., Jr. Multiple indicators and the causal approach to measurement error. American Journal of Sociology, 1969, 75, 264-272.

This article extends the work of Costner (1969) using a multiple indicator approach in path analysis. First, the author discusses the case of the general (linear additive) recursive system. The three-variable model presented below is considered.

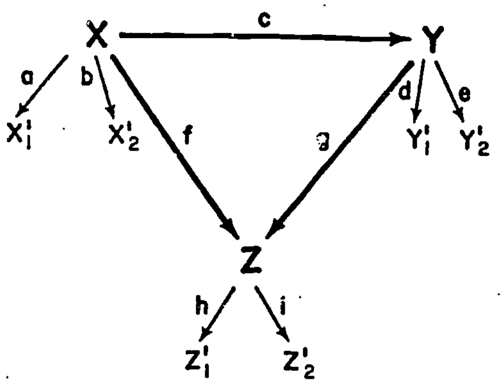


FIG. 1

Looking at the relationship between X and Z, the correlation between X_1' and Z_1' is the sum of two separate paths--one being directly from X_1' to Z_1' (or afh) and one being directly through Y (or acgh). Then, by factoring out a and h (which are the measurement-error paths) one obtains afh + acgh = ah(f + cg). For all other combinations of X_1' and Z_1', the following result.

$$r_{x_1'x_2'} = ab \qquad r_{x_1'z_2'} = ai(f+cg)$$

$$r_{z_1'z_2'} = hi \qquad r_{x_2'z_1'} = bh(f+cg)$$

$$r_{x_1'z_1'} = ah(f+cg) \qquad r_{x_2'z_2'} = bi(f+cg).$$

The compound path represented by f + cg appears in the pairings involving x_1 and z_1, and it can be shown to equal r_{xz}. In particular, the products $r_{x_1'z_1'}r_{x_2'z_2'}$ and $r_{x_1'z_2'}r_{x_2'z_1'}$ both equal abhi$(f + cg)^2$ = abhir_{xz}^2. Thus, we obtain a test for randomness of the measurement error and an estimate of $r_{x2}2$. The article extends the approach to a four-variable model, and the procedure could be extended to any number c recursively related variables.

Next, the problem of the use of a single measure to represent one or more of the variables is considered. In the case of the three-variable model, the author shows that a solution may be obtained only when the single indicator is linked with an intervening variable (Y) in a simple causal chain of the form X Y Z. (This means that one must make the a priori assumption that no direct link exists between X and Z.)

In the last section of the paper, two problems are discussed: (1) that of having a large number of unknown factors affecting the intercorrelations among the indicators, and (2) that of being unable to collect data to have multiple indicators of all variables. The suggestion is to combine the multiple-indicator approach with the instrumental-variable approach. For example, we assume that X causes Y (X Y) with X and Y measured with random error. If we find one or more instrumental variables Z_i that are assumed to cause X but not Y (except indirectly through X), then instrumental-variable estimators b^*_{yx} can be formed by taking the ratios of the covariances of Y and X with Z^i. The instrumental variable estimator provides an estimate of B_i used in the equations for the variables (e.g., B_i in the equation $y = a_i + B_1 x + \ _y$.

The bibliography includes nine references.

Bohrnstedt, George W. A quick method for determining the reliability and validity of multiple-item scales. American Sociological Review, 1969, 34, 542-548.

This paper demonstrates a method for obtaining reliability and validity estimates of composite measures using only information from the items comprising these measures. Estimates are provided for discriminant validity (whether various measures are discriminating among concepts), criterion-related validity (whether the measure correlates with some criterion), and internal consistency (whether the items provide independent measures of the same construct).

For discriminant validity, the correlation between two composites can be computed from the zero-order correlations among the items. For scales X and Y where

$$X = \sum_{i=1}^{m_1} V_i \text{ and } Y = \sum_{k=m_1+1}^{m_2} V_k$$

and V_i and V_k are the items, the correlation between the composites, r_{xy}, can be computed as follows:

$$r_{xy} = \frac{\sum_{i=1}^{m_1} \sum_{k=m_1+1}^{m_2} \sigma_1 \sigma_k r_{1k}}{\sqrt{\sum_{i=1}^{m_1} \sigma^2{}_1 + 2\sum_{i<j}^{m_1} \sum^{m_1} \sigma_1 \sigma_j r_{1j}} \sqrt{\sum_{j-m_1+1}^{m_2} \sigma^2{}_k + 2\sum_{k<1}^{m_2} \sum^{m_2} \sigma_k \sigma_1 r_{k1}}} \qquad \text{(Formula 1)}$$

This formula can be rewritten:

$$r_{xy} = \frac{e}{\sqrt{a+2b} \quad \sqrt{c+2d}} \qquad \text{(Formula 2)}$$

where a equals the sum of variances of Items 1 through m_1; b equals the sum of all covariances of Items 1 through m_1; c equals the sum of variances of Items m_1+1 through m_2; d equals the sum of all covariances of

Items m_1+1 through m_2; and e equals the sum of covariances of Items 1 through m_1 with Items m_1+1 through m_2.

For criterion-related validity, it is necessary to obtain the correlation of the composite measure with a criterion variable. Formula 2 can again be used by substituting the criterion variable for the second composite measure (i.e., Items m_1+1 through m_2). Where no external criterion exists, one can use the total score, with the item-to-total correlations usually called item analysis. Item analysis can be calculated using Formula 2, but substituting the total score for the first composite measure (i.e., Items 1 through m).

For determining internal consistency reliability, the author focuses on coefficient alpha (α), which can also be derived from the covariance matrix. Based on the first composite (i.e., Items 1 through M), and using the notation provided in Formula 2, the following formula was presented:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{a}{a+2b}\right]$$

The bibliography contains 14 references.

Bradburn, Norman M. Selecting the questions to be asked in surveys. Monthly Labor Review, 1970, 93, 27-29.

This article focused on the problems of selecting appropriate questions to test social-psychological variables. The author cited the following as major obstacles in developing adequate decision rules for selecting survey questions:

1. A lack of agreement concerning the appropriate variables that are relevant to particular social programs.
2. An inadequate conceptualization of the variables suggested for study.
3. A relative lack of interest in systematic methodological research and survey measurement.
4. The relative underdevelopment of measurement theory in survey work as compared with the sophistication of sampling theory.
5. The special historical and cultural problems affecting the phrasing of questions and causing response differences across different people at the same time and across the same people at different times.

He then briefly discussed the possible causes for each problem. Finally, he presented three "rules of thumb" in deciding upon survey questions:

1. If the research is unable to specify the theoretical relevance of the selected variables, then the effort to measure them should be abandoned.
2. Minimum standards for item construction, response reliability, and other psychometric properties should be specified and followed.
3. The researcher should identify the degree of heterogeneity in the survey population that might require different forms of questions for different segments of the population.

Bradburn, Norman, & Seedman, Seymour. Improving interview methods and questionnaire design: Response effects to threatening questions in survey research. San Francisco: Jossey-Bass, 1979.

This book focuses on methods to be used in a survey to minimize the underreporting of socially undesirable behaviors and the overreporting of socially desirable behaviors. The suggestions are based on three large samples of the general population interviewed by the field staff of the National Opinion Research Center.

1. A technique for determining whether a question is threatening is to ask respondents how "most people" would react to the questions. If 10% to 20% feel that most people would be very uneasy, then some threat problem exists. If 20% or more feel that people would be uneasy, then the response to the question will be distorted due to threat.

2. Question length is somewhat related to responding to threatening questions by allowing more time for remembering socially undesirable behaviors; thus, less understatement of these behaviors occurs.

3. Respondents acknowledged significantly more socially undesirable behavior with questions where response categories were not provided.

4. Although using the respondent's own terms for socially undesirable behavior yielded a slight improvement in response accuracy, the needed format is cumbersome and usually not worth the extra effort.

5. The authors found that 30% to 50% of responses were not valid due to underreporting of socially undesirable behavior (e.g., about 50% of those known to have been arrested for drunken driving failed to acknowledge it). They also found that 14% to 36% of responses were not valid due to overreporting of socially desirable behavior (e.g., 36% of those known to have failed to vote in the most recent election claimed that they had).

6. Asking about specific behaviors of particular friends produced some increase in reports of socially undesirable behavior beyond that indicated by the respondents for themselves.

7. The authors found that "vague" modifiers are in fact vague. The meaning of such words as "rarely," "hardly ever," or "often" varies according to the context and across individuals. Therefore, they recommend that such responses should not be quantified. Indeed, if respondents can give quantitative responses, they should be asked to give such answers.

8. Variations in the way that interviewers present questions do not have an impact on the validity of the data. On the other hand, interviewers' prior expectations did affect reporting of socially undesirable behaviors; however, in most cases, the effects were trivial. The authors recommend that (1) interviewers who expect a study to be difficult should not be hired; and (2) interviewers should not be trained to expect underreporting.

9. Third parties brought in by the interviewer, including a tape recorder, did not appear to influence responses. Third parties related to the respondent (e.g., parents in the presence of children, children in the presence of parents, husbands in the presence of wives) did affect responses.

10. The length of the introduction was unrelated to responses. Explicit assurance of confidentiality yielded a "small but significant and consistent effect on willingness to answer particular questions" (p. 132). Those who refused to sign following the interview were more likely to have refused to answer certain questions and to have provided poorer quality data. Thus, the authors suggest that more accurate estimates of population parameters could be obtained by excluding the responses of those refusing to sign after the interview.

11. Of four data collection methods--face-to-face interviews, telephone interviews, self-administered questionnaires, and randomized response techniques--none is superior in obtaining valid responses to all types of threatening questions. Self-administered questionnaires yielded the lowest cooperation rate, worked less well for adult respondents who had not completed high school, and were worse on socially undesirable behavior but slightly better on socially desirable behavior.

12. In the randomized response technique, a respondent is asked to answer one of two questions, one threatening and the other innocuous. Furthermore, the authors describe a procedure that can be used so that only the respondent knows which question is being answered. Since this technique requires large samples and makes multivariate analyses difficult, they advise using the approach for simple surveys about socially undesirable behavior.

13. For correcting underreporting due to respondent anxiety, the authors suggest that one distinguish those who report a question as being highly threatening. Then, one can assume that those who feel high threat engage in that behavior as much as those who feel moderate or low threat.

Brooks, C. A., & Bailar, B. A. A statistical policy working paper 3. An error profile: Employment as measured by the current population survey. Washington, D.C.: U.S. Department of Commerce, 1978.

The Subcommittee on Nonsampling Error of the Federal Committee on Statistical Methodology decided to illustrate the ways in which nonsampling error could affect survey statistics by constructing "error profiles." An error profile has as its objective the listing of the survey operations with the investigation. This error profile describes the potential sources of error in the Current Population Survey (CPS) as they affect the national employment statistics. The purposes of this document are as follows:

1. To illustrate how an error profile is created in an effort to encourage government statisticians to provide error profiles for the major recurrent survey statistics;

2. To compile in a single document the sources of error and the information that is available about these sources of error and their impact;

3. To illustrate the need for controlled experiments to measure nonsampling errors because of the lack of knowledge of the impact of these errors;

4. To stimulate development of a mathematical model that will reflect the ways in which the errors from different sources interact.

Campbell, Donald T., & Fiske, Donald W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

The authors argue for an approach to validation that uses a matrix of intercorrelations among tests representing at least two traits, with each measured by at least two methods. Thus, validity should involve both convergent and discriminant indicators; measures of the same trait using different methods should correlate higher than either those of different traits measured by the same method or those of different traits measured by different methods. In discussing convergence in validity, the authors emphasize a distinction between validity and reliability; namely, validity represents the convergence of independent methods measuring the same trait, while reliability represents the convergence of similar methods measuring the same trait. Discriminant validation emphasizes the problems that can arise when high correlations are obtained with tests purporting to measure different traits.

Illustrations taken from the literature reveal an excessive amount of method variance, which usually exceeds the amount of trait variance. In these cases, method or apparatus factors make very large contributions to the measurement variance.

In discussing this approach, the authors emphasize the importance of determining the adequacy of measure (through convergent and discriminant validation) before using the measure to test the adequacy of a construct as determined by theoretically predicted associations with measures of other constructs. Thus, the approach can be viewed as similar to construct validity and convergent operationism. Several difficulties with the approach were mentioned: (1) the methods selected for a matrix should be independent of each other; (2) although convergence (of distinct methods) can be clearly demonstrated, discriminative validity (of nonoverlap with other traits) is not so easily achieved; and (3) convergent validity may not be obtained because neither method adequately measures the trait, because one of the methods does not measure the trait or because trait is not a functional unity.

Conger, Anthony J., & others. Reliability and validity of National Longitudinal Study measures: An empirical reliability analysis of selected data and a review of the literature on the validity and reliability of survey research questionnaires (National Longitudinal Study of the High School Class of 1972). Research Triangle Park, N.C.: Center for Educational Research and Evaluation, Research Triangle Institute, 1976. (ERIC Document Reproduction Service No. ED 013 371)

This report has two major purposes: (1) to review the literature on the validity and reliability of survey data and (2) to analyze the reliability of selected questions in the Second Follow-Up Questionnaire of the National Longitudinal Survey of the High School Class of 1972 (NLS). The key part of the reliability study was an empirical analysis of selected NLS items on a sample of NLS respondents.

The report has four sections: (1) A Capsule View of the National Study of the High School Class of 1972, (2) A Review of Survey Data Validity and Reliability, (3) A Reliability Study of NLS data, and (4) Implications and Conclusions.

Findings in section four include:

- Contemporaneous, objective, factually oriented items were more reliable than subjective, temporally remote, or ambiguous items.
- Items with future or retrospective orientation were less reliable than contemporaneous items.
- Personally sensitive items were less reliable than other factually oriented items.
- Highly factually oriented items were more valid than less factually oriented items.
- The reliability and validity of attitudinal and psychological variables included in the sample were weak.
- Mail-in questionnaires produced more valid data than interviews when records could be consulted.
- High ability and high SES persons were less influenced by data collection procedures than low ability or low SES persons; the latter groups were more cooperative and produced more accurate or valid data in the interview procedure.

- High SES and high ability respondents were more reliable and to a lesser extent more accurate than lower ability or lower SES respondents.

- High SES respondents were more accurate than low SES respondents.

- Blacks provided less accurate information than whites; however, blacks provided more reliable information than whites.

The authors contend that based on these findings, generalizing results across populations with different respondent characteristics would be highly problematic. They suggest using demographic variables not only as control variables but also as moderator variables when survey data are used to construct structural or causal models. They further state that path analyses can be conducted with factually oriented data on high ability or high SES respondents but that structural modeling is required with error of measurement estimates based on a similar population. In general, they find the safest approach is to conduct reliability and validity pilot studies prior to the main survey.

Cronbach, Lee J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

The author begins with a historical resume regarding the work on measures of reliability. A formula (α) is presented and is the general form of the Kuder-Richardson Formula 20 ($r_{tt(KR20)}$):

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\Sigma_i V_i}{V_t}\right)$$ where V_t is the variance of test scores and V_i is the variance of item scores after weighting.

The remainder of the article consists of an examination of the properties of alpha. First, alpha is shown as the mean of all split-half coefficients resulting from different splittings of a test. Second, alpha provides an estimate of the correlation between two random samples of items from a pool of items like those in the given test. Third, alpha is a lower-bound for the coefficient of precision and for coefficients of equivalence obtained by the simultaneous administration of two tests with matched items. Fourth, it is a lower-bound to the proportion of test variance attributable to common factors among items; thus, it is an index of common-factor concentration and can serve as an index of homogeneity. Finally, alpha is an upper-bound to the proportion of variation due to the first factor. Tests having distinct subtests should be divided into these subtests before using the formula.

Another coefficient $\bar{r}_{ij}$ (or $\bar{\phi}_{ij}$) is provided as follows:

$$r_{ij(est)} \doteq \frac{\alpha}{n + (1-n)\alpha} \quad \text{or} \quad r_{ij(est)} = \frac{1}{n-1}\;\frac{V_t - \Sigma V_i}{\Sigma V_i}$$

It represents the correlation required, among items with equal variances and equal covariances, to obtain a test of length n having common-factor concentration α. As a measure of item interdependence, it indicates heterogeneity in both difficulty and content.

The author concludes with some implications for test design. Maximum interpretability of scores can be obtained by increasing the first-factor concentration in any separately-scored subtest and by avoiding substantial group-factor clusters with a subtest.

The reference section contains 39 citations.

Cronbach, Lee J. Test validation. In Robert L. Thorndike (Ed.), <u>Educational measurement</u>. Washington, D.C.: American Council on Education, 1971.

Validation is the process of examining the accuracy or soundness of predictions or inferences made from a particular measure or test score. It calls for the integration of many types of evidence. The following table provided in the article presents a summary of the types of validation, the kinds of questions asked, the kinds of data used, and the types of judgments made, and it gives an overview of the major concerns presented in the paper.

TABLE 14.1

Summary of Types of Validation

Focus of Investigation	Question Asked	Use Made of Student Response Data	Use Made of Judgment
Soundness of Descriptive Interpretations:			
Content validity	Do the observations truly sample the universe of tasks the developer intended to measure or the universe of situations in which he would like to observe?	Scores on test forms constructed independently may be compared.	To decide whether the tasks (situations) fit the content categories stated in the test specifications. To evaluate the process for content selection, as described in the manual.
Educational importance	Does the test measure an important educational outcome? Does the battery of measures neglect to observe any important outcome?	None.	To compare the test tasks with the educational objectives stated by responsible persons.
Construct validity	Does the test measure the attribute it is said to measure? More specifically: the description of the person in terms of the construct, together with other information about him and in various situations; are these implications true?	Scores are compared with measures of behavior in certain other situations. Or the test is modified experimentally and changes in score are noted.	To select hypotheses for testing. To integrate findings so as to decide whether the differences between persons with high and low scores are consistent with the proposed interpretation. To suggest alternative interpretations of the data.
Usefulness for Decision Making (Criterion-Oriented):			
Validity for selection	Do students selected by the test perform better than unscreened students?	Regression of outcome measure on test score is examined.	To decide whether the criterion fully represents the outcomes desired, including outcomes more distant in time. To decide whether a new situation is enough like the validation situation for the results to generalize.
Validity for placement	Is performance improved when students are allocated to treatments according to their test scores?	Regression slope relating outcome measure to test score for one treatment is compared with that for another treatment.	Same as above.

The author devotes an extensive section of this paper to a discussion of construct validity. The procedures used to examine construct validity include correlational, experimental, and logical studies. Correlational studies focus on the convergence of similar indicators and the discriminability from other constructs. In experimental studies, interventions are implemented deliberately to change scores on the measure. The judgmental or logical analysis involves careful scrutiny of the test and is one of the best sources of counterhypotheses. The author then discusses construct validation as the formal testing of nomological networks, and he provides some examples of this approach.

The final section of the paper focuses on predictions and decisions for selection and placement and provides suggestions for specific procedures.

The reference section includes 159 citations.

Cronbach, Lee J., Gleser, Goldine C., Nanda, Harinder, & Rajaratnam, Nageswari. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

This book provides a detailed expansion of the work of Cronbach, Rajaratnam, and Gleser (1963) and of Gleser, Cronbach, and Rajaratnam (1965). Measurement is used for making decisions or arriving at conclusions such as (1) absolute decisions about an individual, (2) comparisons between two courses of action for an individual, (3) comparisons between persons, and (4) conclusions about the relation between pairs of variables. In all of these cases, the researcher is not particularly interested in a specific score, rather he or she generalizes from the sampled score to the universe. So, "reliability" becomes a question of the accuracy of generalization.

A distinction is made between generalizability (G) studies and decision (D) studies. A G study collects data for making estimates of the components of variance in the measurements, while a D study collects data for making decisions or drawing conclusions. Such a distinction recognizes that certain studies are conducted for developing a measuring procedure (G studies) and that other studies actually use the developed procedure (D studies). Therefore, reporting the results of the G study is important so that new investigators can plan D studies and can estimate the error of generalization under that plan.

A behavioral measurement is a sample from a collection of measurements, or the universe. The measurements or observations encompassed in a G study represent the universe of admissible observations. A decisionmaker using the same measuring technique generalizes to some universe of conditions, referred to as the universe of generalization. The G study is useful to this decisionmaker only if its universe of admissible conditions is identical to or includes the proposed universe of generalization.

The statistical concepts of traditional reliability theory appear in altered form. True-score variance is viewed as a universe-score variance and is a property of the selected universe of generalization as well as the selected measure and the population. Observed-score variance is a property of the measurement procedure, the population, and the experimental design. The coefficient of generalizability is the ratio of universe-score variance to expected observed-score variance and is an intraclass correlation (similar to Rulon's split-half formula, Horst's formula for reliability with multiple observations, Kuder-Richardson formulas 20 and 21, and the Hoyt-Cronbach alpha coefficient).

In traditional measurement theory, the reliability coefficient represents the ratio of "true-score" variance to observed-score variance. The authors alter this theory, because (1) the various universes of generalization imply many variance ratios, and (2) each alternative D-study design results in a different variance of observed scores, altering the ratio. They distinguish different kinds of error associated with different kinds of universe-score estimates: (1) decision based directly on the observed score with error Δ, (2) decision based on the observed deviation from the sample mean with error δ, and (3) decision based on a regression estimate with error Σ. With these different kinds of error, the concepts in classical theory become ambigious. Error Δ equals the difference between the universe score and the observed score, while error δ has its variance equal to the difference between the observed-score variance and the universe-score variance. In classical theory, these errors are identical, but with weaker assumptions $\sigma^2(\Delta)>\sigma^2(\delta)$. The index $\sigma(\Delta)$ indicates how far the measures are likely to depart from their "true" values (or the person's universe score). On the other hand, the standard error $\sigma(\delta)$ indicates whether one of two persons tested under the same conditions has a significantly higher universe score than the other. Finally, error ε is associated with the regression estimate, and it will, on the average, be less than error δ. The value of $\sigma(\varepsilon)$ provides a rough indication of the precision with which the estimation procedure estimates the universe scores.

The analysis of a G study focuses on "components of variance" of an observed score. The following equation divides the observed score (X_{pi}) into components representing hypothesized effects:

$$\begin{aligned}
X_{pi} = {} & \mu && \text{(general mean)}\\
& + \mu_p - \mu && \text{(person effect)}\\
& + \mu_i - \mu && \text{(condition effect)}\\
& + X_{pi} - \mu_p - \mu_i + \mu && \text{(residual)}
\end{aligned}$$

The associated variance equals the sum of the variance components:

$$\begin{aligned}
\sigma^2(X_{pi}) = {} & \sigma^2(p) && \text{(person component)}\\
& + \sigma^2(i) && \text{(condition component)}\\
& + \sigma^2(pi,e) && \text{(residual component)}
\end{aligned}$$

If conditions are classified with respect to two facets (i and j), seven score components and seven variance components result:

		$X_{pij} = \mu$	$\sigma^2(X_{pij}) =$
persons	p	$+ \mu_p - \mu$	$+ \sigma^2(p)$
conditions	i	$+ \mu_p - \mu$	$+ \sigma^2(i)$
conditions	j	$+ \mu_j - \mu$	$+ \sigma^2(j)$
interactions			
	pi	$+ \mu_{pi} - \mu_p - \mu_i + \mu$	$+ \sigma^2(pi)$
	pj	$+ \mu_{pj} - \mu_p - \mu_j + \mu$	$+ \sigma^2(pj)$
	ij	$+ \mu_{ij} - \mu_i - \mu_j + \mu$	$+ \sigma^2(ij)$
residual	pij,e	$+ X_{pij} - {}_{pi} - {}_{pj} - {}_{ij}$ $+ \mu_p + \mu_i + \mu_j - \mu$	$+ \sigma^2(pij,e)$

A one-facet model considers the universe of generalization to be the set of X_{pi} for a person under all admissible conditions, and the mean to which one generalizes is μp. A two-facet model provides three types of universes of generalization: (1) the universe of admissible observations with universe score μp, (2) a restricted universe in which I is fixed, with universe score μp_I, and (3) a restricted universe in which J is fixed, with universe score μp_J. The authors reject the one-facet study, with μ_p and the concept of true score in classical theory, as being deceptively simple since it does not take into account the many facets of the generalization. (Facets refer to conditions of observation such as test forms, observers, and occasions.) Thus, the theory is related to the test theory developed by Lord and Novick (1968, Chapter 8) for "imperfectly parallel" measurements.

The bibliography contains 187 references.

Cronbach, Lee J., & Meehl, Paul E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

The authors introduce construct validity and elaborate on the interpretation as provided by the APA Committee on Psychological Tests. Construct validity is involved in the development of tests for which traditional forms of validity (i.e., predictive, concurrent, and content) are inappropriate because no adequate criterion exists. A construct is some hypothesized attribute (of people), and it is assumed to be reflected in test performance.

The authors further specify certain conditions affecting or governing construct validity. A construct is defined by a nomological network, or an interlocking system of laws. For a construct to be scientifically admissible and for construct validation to be possible, some of the laws or statements in the network must lead to predicted relations among observables. In addition, the network defining the construct and the derivation leading to the predicted observation must exhibit explicit steps of inference; this will lead to proper interpretetion of the validating evidence. Many types of evidence are relevant to construct validity including content validity, inter-item correlations, inter-test correlations, test "criterion" correlations, studies of stability over time, and studies of stability under experimental intervention. Thus, construct validity cannot be expressed in terms of a single, simple coefficient. When a predicted relation fails to occur, the problem may result from the test interpretation or from the network. Modifying the network in line with the observations redefines the construct, and the new interpretation must be validated by new data. It should be recognized that such investigations of construct validity are similar to scientific procedures used to develop and confirm theories.

The reference section includes 60 citations.

Cronbach, Lee J., Rajaratnam, Nageswari, & Gleser, Goldine. Theory of generalizability: A liberalization of reliability theory. The British Journal of Statistical Psychology, 1963, 16, 137-163.

This paper examines reliability theory and provides a rationale for combining and reinterpreting various approaches to reliability. Just as classical validity theory postulates a criterion against which a test is judged, classical reliability theory postulates two or more measures equivalent in content, in mean, in variance, and in intercorrelations. Measures of internal consistency, using half-tests of items, assumed the equivalency of the parts as well as the whole tests. Later work showed that the same formulas result under weaker assumptions about part-scores or that more general formulations result in the lack of a requirement for an equivalence assumption.

Concerns about the precision or reliability of a measure arise because the researcher wants to generalize from a specific observation to a class of observations. The investigator must specify a universe of conditions of observations over which the generalization is to be made. The following assumptions are made: (1) the universe is described unambiguously; (2) conditions are experimentally independent; (3) scores X_{pi} are on an interval scale; and (4) conditions and persons are randomly sampled from their respective populations. A generalizability study (G study) assesses the measuring technique by examining the relation between the observed score and the universe score, while a decision study (D study) provides data from which decisions are made about individuals. In addition, data may be matched (in which case persons are observed under the same conditions) or unmatched (in which case the conditions for each person are sampled independently).

For an unmatched G study, only a one-way analysis of variance yields information about unmatched decision data, such as estimates of the variance of universe scores, the expected variance of the difference between universe scores and observed scores, and the expected observed variance. The coefficient of generalizability P^2MX for unmatched D data can be obtained from the ratio of universe-score variance to observed variance, and it resembles the KR21 and the Horst generalized reliability coefficients. Although a specific coefficient P^2Mi can be estimated for any condition in a matched G study, the estimates will include large sampling errors unless n_i is large. For both matched and unmatched D data, a two-way analysis of variance provides the needed estimates. The intraclass correlation for matched D data is V_{MP}/EV_i, and it provides an estimate of alpha α (of which KR20 is a special case).

The following table, duplicated from the article, presents the most important formulas.

TABLE 1. CHIEF FORMULAS

Data in G Study (n_i observations)	Matched	Unmatched
Data in D Study (n_i' observations averaged)	Matched	Unmatched
Variance estimates for D study:		
Universe score	$\frac{1}{n_i}(MS_p - MS_r)$	†
Error	$\frac{1}{n_i'}MS_r$	†
Observed variance	$\frac{1}{n_i}MS_p + \left(\frac{1}{n_i'} - \frac{1}{n_i}\right)MS_r$	†
Intraclass correlation:		
n_i' variable	$\frac{n_i'(MS_p - MS_r)}{[n_i' MS_p + (n_i - n_i')MS_r]}$	†
$n_i' = n_i$	$\frac{MS_p - MS_r}{MS_r}$ ‡	† §
$n_i' = 1$	$\frac{MS_p - MS_r}{[MS_p + (n_i - 1)MS_r]}$	†

†Same as in first column except that MS_{wp} replaces MS_r.
‡KR20 or our (2) may be used as a computing form.
§KR21 may be used as a large-sample computing form.

It indicates that the design of the G and D studies influences the choice of the appropriate formula.

The bibliography contains 39 references.

Cureton, Edward E., Cook, Joseph A., Fischer, Raymond T., Laser, Stephen A., Rockwell, Norman J., & Simmons, Jack W., Jr. Length of test and standard error of measurement. *Educational and Psychological Measurement*, 1973, *33*, 63-68.

A formula is presented that supports Lord's argument that, under appropriate conditions, tests of the same length have equal standard errors of measurement, regardless of the function measured, the maturity of the examinees, the discriminating powers of the items (at least as long as these are all significantly positive), the range of ability in the group tested, and the number of alternatives per item. The appropriate conditions are:

1. The score is the number of right answers, with no correction for guessing.

2. Every examinee reacts to every item, and ideally answers every item, so that the test is not a speed test but a power test.

3. Every item is functional for the group tested; no items exist for which all examinees give the right answer.

Lord found a correlation of .996 between the standard error of measurement (SE) and the square root of the number of items for a sample of 50 ETS tests. The regression of SE on n was SE = $.432\sqrt{n}$.

The authors present evidence on six generally well regarded tests that are based on reliability coefficients. The mean values were all found to be a little lower than Lord's .432, and the standard deviations and ranges suggest that the stability of estimates of SE are limited to two decimals at best. Even so, the uniformity of the results is impressive, considering the range of grades and tests on which they are based. The formula presented for estimating reliability is:

$$r_{11} = 1 - .043n\left(\frac{N}{\Sigma X_1 - \Sigma X_2}\right)^2$$

Duckworth, Pauline A. Construction of questionnaires: A technical study. Washington, D.C.: Civil Service Commission, Personnel Measurement Research and Development Center, July 1973.

This pamphlet was developed to assist government staff in planning and designing questionnaires. The first chapter discusses the characteristics of a good questionnaire: (1) validity, (2) reliability, (3) the method for administration and scoring, and (4) face validity. The second chapter deals with the plans and objectives of a questionnaire. To develop a clear definition of the purpose, the survey researcher should:

- describe the overall problem to be investigated,
- decide whether a questionnaire is the most appropriate approach to study the problem,
- define the needed information,
- decide on the persons/institutions to receive the questionnaire,
- decide on the method of administration,
- decide on the form of the questionnaire and the method for data analysis,
- consider the purpose of survey in terms of the above factors, and
- plan for continuing documentation of the process and eventual follow-up.

The third chapter focuses on the process of defining the questionnaire content. The first step involves the development of general areas of information followed by the identification of specific areas; this may take the form of an outline of the questionnaire content. The fourth chapter considers the problem of selecting appropriate item types. It presents examples of the basic item types (i.e., free-answer or open-end items, short-answer items, dichotomous or two-response items, multiple-choice items, ranking items, and descriptive items) and discusses the advantages and disadvantages of each type. The fifth chapter concerns the procedures for coding and extracting the data and those procedures that must be considered when deciding on the types of items, their arrangement, and the questionnaire layout. The sixth chapter covers the problems in writing questionnaire items. The author recommends that the item writer (1) gain some knowledge of the topic area, (2) develop the items with the help of subject-matter specialists, (3) devise more than one item dealing with a specific subject matter topic, and (4) maintain a careful record of all reviews of the item. The following are characteristics of a good questionnaire item:

- It is limited to a single topic.
- It is appropriate to the selected population and administrative method.
- It can be scored or rated according to the specified method.
- It is written in precise, understandable, and unbiased language.
- It is arranged in a logical position within the questionnaire.
- It yields reliable and valid information.

The seventh chapter deals with item review and tryout. A reviewer should evaluate each item in terms of subject-matter coverage, basic measurement principles, and choice of language. The first tryout may be conducted by the writer personally and is focused on determining (1) whether the subject matter is adequately covered, (2) whether responses in the "Other--Specify" category require a revision of alternative responses, (3) how much time is needed to complete the form, and (4) whether the coding system is effective. The eighth chapter concerns designing the questionnaire form. The general instructions describe the purpose of the survey, the sponsoring agency or office, information that motivates the respondent to complete the form carefully, and general directions on the method of completing the items. The designer of the questionnaire should recognize that (1) the appearance of the form influences response, (2) the ease of response and the length of the questionnaire are important considerations, and (3) provisions for coding should be included. Based on a pretest of the questionnaire, revisions in the test items, their ordering, and the instructions can be made before the questionnaire is developed in final form.

The bibliography includes 15 references.

Erdos, Paul L. High response rates in mail surveys. Paper presented at the annual meeting of the American Psychological Association, New York, September 1979.

The author begins with a brief review of some of the major nonresponse problems (e.g., response refusals, illness, death). Then an example of distorted results is presented, leading to the following rule: "No survey can be considered reliable unless it has a high percentage of return (completions) or unless some kind of verification proves that he response actually received is representative of the entire sample." The next question is, of course, "What is a 'high' percentage of response?" The answer is dependent on the population, the sample, the questionnaire, and the use that will be made of the results. For example, the Advertising Research Foundation, which oversees media and advertising research, recommends an 80% or better response. Although the author claims that it is impossible to generalize about an adequate return rate, he does indicate a minimum standard of 50% response or some verification that nonrespondents are similar to respondents.

The discussion then turns to consideration of the factors affecting the response rate. These include:

1. Prestige of organization doing survey.
2. Amount of interest respondents have in survey topic.
3. Ease of reading questionnaire in terms of copy and layout.
4. Length of questionnaire.
5. Short personal letter that
    a. convinces respondents that they are important to survey and that survey is important,
    b. tells why survey will benefit recipient, and
    c. assures respondent of anonymity.
6. Offer of copy of report.
7. Stamped and processed reply envelope.

8. Postcards to inform respondent that the questionnaires will be mailed and to remind them to return the questionnaires.

9. Several mailings.

10. Incentive (with the following considerations)

    a. Provide something desirable.

    b. Be sure not to introduce bias.

    c. Use something small and light-weight.

    d. Be sure that it is not too expensive.

After discussing these points, the author provides an example from a 1977 survey by The American Banker.

Erdos, Paul L. Professional mail surveys. San Francisco: McGraw-Hill, 1970.

This book describes professional mail survey techniques in enough detail to enable readers to perform mail surveys on their own. The author also attempts to enable the reader to evaluate the validity of other mail surveys.

Chapter 1 contains a synopsis of the history of mail surveys. Return rates are shown to have improved dramatically in recent years because of improved research techniques.

Chapter 2 describes the advantages and limitations of mail surveys. The major advantages are wider distribution; less distribution bias in connection with the type of family; less distribution bias in connection with the individual; no interview bias; better chance of a truthful reply; better chance of a thoughtful reply; time saving (under certain circumstances); centralized control; and cost-saving. The limitations include: no mailing list is available; mailing list is incomplete; mailing list is biased; subject requires specially trained interviewers; the questionnaire cannot be structured; the questionnaire is too long; the questionnaire is too difficult; the information required is confidential; the respondent is not the addressee; the budget is inadequate; and the time available is insufficient.

Chapter 3 describes the various types of surveys: industrial and consumer surveys, media studies and surveys on advertising, multiple surveys and panels, and "hybrid" mail surveys.

Chapter 4 discusses the necessary steps to get from the research question to the survey design. The following steps are proposed as logical and essential parts of survey design:

1. Outline the problem.

2. Define the research objectives.

3. Investigate existing research on the same problem or with the same objectives.

4. Define the universe.

5. Decide on the degree of reliability aimed at within a realistic budget.

6. Define the sample and scope.

7. Decide on the survey method.

8. Decide who will conduct the survey.

9. Establish the techniques that will be needed to achieve the research objectives.

10. Outline the type of tabulation, analysis, and report desired.

11. Make up a time schedule.

12. Make up a cost estimate.

Each step is discussed to familiarize the reader with the concepts. Common problems and errors are described, and the importance of defining specific research objectives is stressed. If questions cannot be formulated into attainable research objectives, the project cannot be carried out. How the survey results will be used must also be considered before beginning the project because the use of the results will influence most of the survey design decisions. The pilot study is a useful aid in making study design decisions (see Chapter 9).

Chapter 5 discusses mailing lists and sampling. Sources of mailing lists are provided. Examples of common errors in sampling are given. The difference between frame bias and sampling bias is discussed. The two basic rules of survey sampling include (1) some acceptable procedures should be used in the selection, and (2) every unit of the universe should have a known chance to be selected.

Chapter 6 describes the main considerations in questionnaire construction. They are:

1. Include questions on all subjects essential to the project.

2. The questionnaire should be brief and "easy to complete."

3. The reader must feel that he or she is participating in an important and interesting project.

4. The form should not contain any questions that could bias the answers.

5. The questionnaire should be designed to elicit clear and precise answers to all questions.

6. Phrasing, structure, and layout must be designed with problems of tabulating in mind.

The author elaborates on each of these considerations. The composition of a brief and "easy-to-complete" questionnaire is described in detail. The ideal questionnaire would be monarch size (7 x 10 inches), one page in length, and printed in Times Roman type face on light color stock. The layout of questions should be uncomplicated. Check questions have a slightly better response rate than open-ended questions.

Chapter 7 is concerned with the quality of the questions, and the importance of the title and introductory questions is stressed. Recommendations include avoiding jargon in the phrasing of questions, avoiding the introduction of bias through the content and position of the questions, and providing clear and precise responses. Ambiguity in the responses provided can render the findings of a survey useless. The final consideration in questionnaire construction is the use of data processing operations. Changes in layout or wording that would make data processing operations more accurate and less costly without impairing the quality of the questionnaire should be made.

Chapter 8 discusses the three considerations that should be given in estimating survey costs:

1. The cost of each item should be figured as closely as possible.

2. No item should be omitted simply because of the cost.

3. Timing, insofar as it affects cost, should be considered carefully.

The first step in estimating survey costs should be listing items that will cost money; the item most often forgotten is "executive time." Examples of an estimate sheet and a job schedule are provided. Factors to be considered in filling out the estimate sheet include timing (when the survey is to be done, time available for completing the survey) and scheduling (number of hours needed for each operation, specific starting and completion dates).

Chapter 9 describes pilot studies. They can be conducted for a number of purposes: to test the quality of a mailing list; to check percentage of returns; to check the occurrence of bias resulting from the wording of cards, letters, and questionnaires; to check how well questions are understood and answered; to check the usefulness of information received; and to check or even establish a cost estimate. The basic principles essential to pilot studies are:

1. The study must be conducted among a random sample of the universe surveyed.
2. The number of questionnaires mailed should be large enough to provide meaningful results.
3. The pilot study should only measure one variable at a time.

Chapter 10 discusses the use of advance notices to increase returns. Types of advance notice include postcards, letters, telephone introduction, and announcement of the survey in a publication.

Chapter 11 is about the use of financial incentives.

Chapter 12 discusses the importance of the accompanying letter in obtaining high percentages of returned questionnaires. A listing of what good letters should convey is provided, and the importance of creating a feeling of personal communication between researcher and respondent is stressed. The entire mailing piece can end up in the waste basket without ever being read if the transmittal letter fails to elicit a favorable response from the participant.

Chapter 13 describes mailing procedures. Mail surveys must not look like junk mail. Questionnaires should be correctly addressed, and addressing should be done by typewriter. The name of the sender should be printed in the corner of the envelope, and first class postage should be affixed. Reply envelopes should be stamped to discourage participants from throwing them away, and they should be preaddressed to the person who signed the transmittal letter. Results from previous surveys have shown that the highest response is achieved from an easy-to-look-at questionnaire mailed in a personal manner with the promise to keep the information confidential. Questionnaires should be keyed for identification, and a possible plan for preparing key mailings is presented.

Chapter 14 discusses follow-up mailings. Ninety percent of all survey returns are received within two weeks after the first mailing. Follow-up mailing to nonrespondents three to four weeks after the main mailing can increase the total number of responses. Reminder postcards and follow-up letters can also be effective in increasing response percentages.

Chapter 15 is concerned with the problem of nonresponse. In the author's opinion, "No mail survey can be considered reliable unless it has a minimum of 50% response, or unless it demonstrates with some form of verification that the nonrespondents are similar to the respondents." The Advertising Research Foundation recommends an 80% or better response on mail surveys. If a survey is well planned and well executed by following the suggestions made in this book, the percentage of nonresponse will be small. If a survey has a large number of nonrespondents, their combined effect must be accounted for.

Chapter 16 is titled Checking in Returns. Daily tallys of returns, undeliverables, and refusals should be kept. The origins of each questionnaires should be marked. Examples of cards for follow-up mailings are provided.

Chapter 17 concerns data processing. The essential part of data processing is tabulating. This chapter discusses in detail tabulations by hand or by machine.

Chapter 18 is a detailed discussion of the punch card and its use.

Chapter 19 describes the editing operation necessary to bring maximum usefulness to the results of the questionnaires. The use of test tabulations to establish codes is described. Coding and coding procedures are described in detail.

Chapter 20 is a discussion of the use of financial incentives in mail surveys.

Chapter 21 discusses recaps of tabulations. The purpose of recaps is to present tabulations in the clearest and most useful form. The operations necessary for recaps are discussed.

Chapter 22 discusses final checking and presentation. Helpful final "checks" are described. The plausible contents of a final report are given as well as a recommended format for the final report.

Chapter 23 is a discussion of recognition studies and corporate-image surveys. The pitfalls of "before and after" surveys are discussed. A short discussion of case histories is also presented. Examples of questionnaires for both types of surveys are provided.

Chapter 24 discusses the difficulties of conducting international mail surveys (e.g., language and customs, postage, response rate and language, and content of questions).

Chapter 25 contains a checklist for evaluating surveys. The checklist can be used by researchers when planning and executing surveys to make sure that no details have been overlooked. The user of the findings of a survey or the reader of a research report may want to use the list to evaluate a survey.

Chapter 26 discusses ethical standards in mail research and describes the obligations a researcher has to the sponsors of the survey, the respondents, the readers of the report, and organizations and individuals that potentially could be harmed by improper use of the findings. A number of associations have adopted codes of ethics, including the American Association for Public Opinion Research and the World Association for Public Opinion Research.

The appendix contains statistics on questionnaires sent out over a four-year period by the author's market research company. In general, small, one-page questionnaires had the highest return rates. Large, two or more page questionnaires had the lowest returns. Second wave mailings raised rates of return 10% to 15%. The speed of response analyses showed that 72% of all returns are received within seven days after the mailing date; 94% of all returns are received within two weeks after the initial mailing.

A glossary of terms is provided along with a section citing books on related subjects.

Erickson, Jeanne M., McDonald, Blaír W., & Gunderson, E. K. Eric. Reliability of demographic and job-related information. Journal of Psychology, 1971, 79, 237-241.

The purpose of this study was to determine the reliability of biographical and job-situation items using a test-retest procedure. Eighteen items, with completion or multiple-choice response formats, were administered to 562 men aboard an attack carricr at the beginning of a cruise and en route to home port six months later. Bivariate distributions of matched items resulted in the identification of three groups of items: (1) those with invariant responses (e.g., five factual questions regarding marital, parental, and sibling status; region of residence; and assigned duty division); (2) those with lower stability but generally consistent responses (e.g., eight items predictably affected by the time interval including rate group, pay grade, duty time, age, number worked closely with, number supervised, and religion); and (3) those with low stability and with incompleteness in response (e.g., five items including father's occupation, father's education, number of dependents, education, and General Classification Test score). Together, these data showed a high level of reliability, with an overall response consistency of 94%. The authors concluded that almost all of the items possessed a sufficiently high reliability to justify their use in predictive studies.

Ferber, Robert. The reliability of consumer surveys of financial holding: Demand deposits. Journal of the American Statistical Association, 1966, 61, 91-103.

The results of two studies that examined response and nonresponse errors in consumer reports of demand deposits are discussed. The two studies involved farm owners or operators and residents of a large metropolitan area. The responses of the sample members as to holdings in demand deposits were verified by examining the records available from savings institutions. Thus, for each study, it was possible to determine the actual and the reported size of the account for those who responded or who refused to respond.

In the farm study, for example, the average balance was underestimated by 24.2%. This discrepancy resulted partly because the average actual balances of the nonrespondents exceeded those of the respondents and partly because the respondents underestimated their balances. A cross-tabulation of the reported balance size by the actual balance size showed that very small accounts tended to be overestimated and very large accounts tended to be underestimated. In reports on changes in the balance over the preceding three-month period, an average decrease of $158 was reported, whereas the average balance actually increased by nearly $300. Cross-tabulations of the change data revealed that (1) respondents tended to report no change when a change actually occurred, (2) large decreases (of $500 or more) tended to be overstated, and (3) large increases tended to be understated.

The effect of these nonsampling errors on estimates of the reliability of the mean was assessed. The apparent standard error of the mean was calculated as $317 while the true standard error of the mean was $403. However, the extent to which confidence intervals remained valid in the presence of nonsampling errors (K) indicated a favorable result.

(K = $\sqrt{\frac{MSE_{\bar{y}}}{\sigma_y^2} - 1}$ where MSE_y is the mean square error and σ_y^2 is the sample variance.) In this example, K equalled .78, which means that the 95% confidence interval applied to the sample observations had a probability of .87 of containing the true parameter.

Recognizing that the response error may have been due to the check-float (i.e., the interval that elapses between the time a check is written and the time it is presented to the issuing institution), the author attempted to estimate the extent of the check-float affect. This was accomplished by dividing the samples into two groups: (1) those for whom observed discrepancies were clearly attributed to check-float, and (2) all others in the sample. By segmenting the variance into the float error and the response error, the author found that about 25% of the variance in the discrepancy in checking account balances was accounted for by check-float and about 75% by response error.

The author concludes that, compared with time deposits and personal debt, reports on demand deposit holdings are fairly reliable. However, changes in these holdings appeared to be inaccurate as reported changes in other assets. The bibliography contains 18 references.

Foster, Penny D., & Neal, Phillip. Graduate science education student support and postdoctorals. Washington, D.C.: National Science Foundation, Division of Science Resources Studies, 1975. (ERIC No. ED 114 282)

The National Science Foundation conducted two separate data collection efforts:

(1) Applications by department chairmen for NSF graduate traineeships from 1967-1971, and

(2) Statistical surveys in 1972 and 1973 by NSF to continue series with broader range coverage of departments.

A method was devised to examine the responses from departments that reported consistently for three or four years. This "matching" process enabled NSF to examine short-term trends, which then became the basis for construction of an index, using 1967 as the base year. This report primarily discusses the second data collection effort, particularly the 1973 survey findings.

National statistics are presented on federal aid to graduate students in the sciences and in engineering for fall 1973. Data were provided by every institution with a doctoral program in clinicial and medical sciences and in engineering as listed in the 1973-1974 Directory of American Medical Education of the Association of Medical Colleges. In 1973, responses were received from 6,559 master's and doctorate departments in 339 institutions (876 master's and 5,683 doctorate departments). The response rate was 100%.

Characteristics of graduate enrollment examined in this report are:

- enrollment status (full- and part-time)
- distribution among fields of science
- level of study (first year or beyond)
- citizenship (U.S. and foreign)
- control of institution (public or private)

- sex of students

- data on type and source of major support (available for full-time students only)

Graduate Enrollment and Sources of Support

Highlights of the results included the following findings:

- Doctorate granting institutions enrolled almost 218,000 full- and part-time students in the sciences and in engineering in 1973, a drop of 1% from 1972. This seemed to be a continuing trend since 1970. Every area of science was affected by the decrease except life science and psychology, both of which increased by 2%.

- Full-time enrollment was down almost 3% from 1972 to 1973. Part-time enrollment, on the other hand, went up 4% during this period. This shift to part-time enrollment indicated the growing need and dependence of graduate students on employment to complete their graduate education.

- Over the six-year span (1967 through 1973), full-time enrollment showed an overall decline of 5% from the 1967 base, with the number of students dependent on federal support declining by 40% during this period. While federal assistance was declining, both institutional support and self-support were increasing.

- The number of students dependent on federal fellowships and traineeships declined 22% from 1972 to 1973, while institutionally supported programs increased by 15%. Research assistantships also declined by 2%, offsetting an increase of 8% in institutional support and 10% in other support areas.

- Foreign graduate student enrollment continued its downward trend noted in previous years. Psychology was the only area of science to show an increase in foreign students between 1972 and 1973.

- Postdoctoral utilization by field of science was examined in terms of type and source of support and year of the Ph.D. Science and engineering graduate departments utilized 16,400 postdoctoral appointees in 1973, 69% of whom received some form of federal support. In terms of change since 1967, the number of postdoctorals rose 31% by 1972, but dropped 6% between 1972 and 1973. This may have been influenced by lower unemployment rates for doctoral scientists and engineers in 1973, since postdoctoral appointments are considered temporary, short-term employment for recent Ph.D. graduates.

Appendices to this report include notes on general methodology, classification of institutions in the survey, detailed statistical tables, and a reliability and validity assessment of the 1973 survey.

Sample Selection Methodology for the Reliability and Validity Check

The sample included 30 institutions selected with a stratified random design from the 235 graduate and 104 medical schools surveyed in 1973. Since medical schools were one-third of the total, ten were selected.

Within each of the two sets of institutions, schools were selected systematically with probabilities proportional to the estimated number of graduate students in science departments plus the number of postdoctoral appointees in these departments in 1973. (A full explanation of procedures involved in reliability and validity checks is presented on page 73.) Also included in the appendices is coverage of data comparability between the NSF study of graduate student support and other surveys of graduate students (p. 80).

Frankel, Lester R. Survey design in anticipation of nonresponse and imputation. Paper presented at the Symposium on Incomplete Data of the National Research Council, Washington, D.C., August 1979.

The problem of nonresponse can be greatly reduced by minimizing the extent of nonresponse during data collection and by minimizing the impact of nonresponse through imputation techniques. Of three sources of error--statistical sampling error, response error (nonresponse and inadequate response), and error in specification of objectives--the author examines the third source as a means of controlling the other two sources. The objectives of the survey include a specification of the variables and the population. As an example of the role of these objectives in controlling nonresponse, the author cites a situation that might arise in marketing research. Because ghetto areas have a high nonresponse rate and low purchasing power, the market researcher may want to eliminate such areas from the sample.

The design of the questionnaire is a critical factor. The first concern is to ease the burden of the respondent and to reduce the sensitivity of the questions (e.g., ask for an indication of a range rather than a specific age or income level). The second concern is to include questions asking for key information that can be used in the imputation procedures to estimate missing observations.

In terms of interviewing procedure, the author recommends using a replacement procedure as compared with a "block equalization" technique. The latter approach involves imputing missing observations from a block (i.e., the primary sampling unit) by weighting data from respondents in the same block. He claims that using an actual observation is statistically more efficient than taking an average of other observed values already included in the sample and that "as a result, no weighting is required and the design effect of this latter procedure is 1.00."

The last section of the paper focuses on a model for repeated interviewing attempts. Three distributions of the latent response function (i.e., a density distribution of response probabilities) are contrasted for repeated calls in terms of the following results: (1) proportion of sample reached, (2) average response probability of those reached, and (3) estimates of population means. The major point is that, in determining the number of calls to be made to respondents, it is necessary to consider the percentage of the designated sample reached, the nature of the relationship of the variable being measured, the response rate, and the latent response function.

The bibliography contains seven references.

Frankel, Lester, & Dutca, Solomon. The role of respondent resistance in the census of population. Marketing Review, 1979, 34.

The article begins with a description of activities being conducted in anticipation of the 1980 Census. Concern about undercounting in the Census did not arise until the 1970s. Then, by using birth registration and Medicare records, it was determined that the undercount was not uniform. In particular, the undercount rate was much higher among blacks than among whites.

After discussing the procedures used by the Census and the indications of respondent reluctance, the author discusses the following hypotheses related to respondent reluctance: (1) burden hypothesis, (2) frustration hypothesis (from a complicated and confusing format), (3) fear of disclosure of some form of nonlegal activity, (4) hypothesis of overall reluctance and noncooperation, (5) hypothesis relating to feeling of lack of confidentiality, and (6) hypothesis relating to superstition against giving information about one's self. Empirical studies of nonresponse and nonresponse bias are then discussed.

Gentry, James W., & Milliken, George A. A comment on Wayne E. Hensley's Increasing response rate by choice of postage stamps. The Public Opinion Quarterly, 1975, 3(39), 367-372.

This short article disputes W. E. Hensley's contention that he found significant differences in response rates among different combinations of types of postage. The authors contend that according to their findings the response rates could have been randomly generated. When the authors tested the null hypothesis that all of the nine combinations would have equal response rates, they found they could not reject the null hypothesis. The test for independence of Hensley's 3x3 summary table again found that the null hypothesis could not be rejected.

The reply by W. E. Hensley defends the summary table on the grounds that even though the sample used may not have been large enough to produce statistically significant chi-square results, it is reasonable to expect the observed trends to continue since the sample was randomly assigned to conditions. Hensley defends his findings that the use of a commemorative postage stamp on a mailed survey stimulates a high response rate with the final contention that his findings are supported and validated by the findings of past research. According to Hensley, the cheapest and most efficacious combination is to have one envelope metered and the other bearing a commemorative stamp.

Havighurst, Robert J. The reliability of rating scales used in analyzing interviews with parents, students, teachers, and community leaders (The National Study of American Indian Education, Series IV, No. 9, Final Report). Chicago: University of Chicago, 1970. (ERIC No. ED 046 600)

This paper is part of the final report of the National Study of American Indian Education and reports on the reliability of rating scales used in analyzing the interviews conducted during the study. The rating scales were used (1) to evaluate a particular school or school system in a particular community, (2) to compare schools and communities singly and in various combinations, and (3) to compare perceptions and attitudes of parents with students, parents with teachers, teachers with students, etc.

A number of sources of error or disagreement that can occur in the use of rating scales are described. These include the halo effect, leniency, logical error, and clustering ratings near the center of the scale. The reliability measures used to determine the level of reliability of the ratings on the scales were:

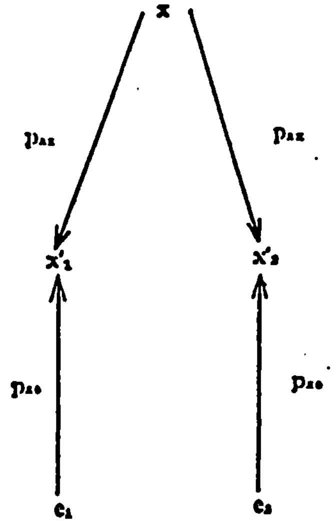
1. product-moment correlation coefficient,
2. Spearman rank order coefficient (rho),
3. Kendall's coefficient of concordance (w) (the author felt this measure superior to the first two since it contains a correction procedure for tied ranks), and
4. intra-class correlation coefficient (most useful to authors because the procedure is based on analysis of variance and gives reliability coefficients for an average judge in a group of two or more judges).

Relatively high reliability of the data for the rating scales was found (<.70). The consistency of the rating and the reliability of the ratings from the various field centers were high enough to permit useful comparisons between various schools or communities and between various types of respondents to the interviews.

Heise, David R. Separating reliability and stability in test-retest correlation. American Sociological Review, 1969, 34, 93-101.

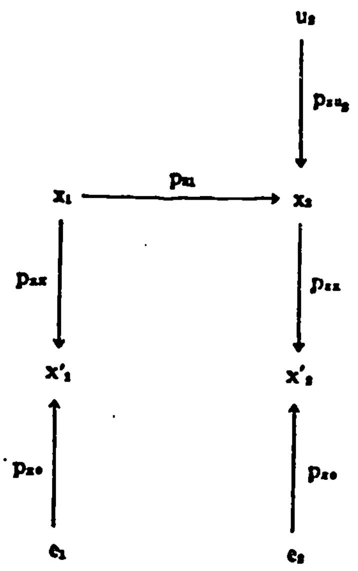
Techniques developed by psychometricians to assess reliability may pose problems for the sociologist since, in sociological research, key variables may be measured by a single question. Although the test-retest correlation may be used in such situations, it may not measure true reliability because of temporal instability. To separate measurement errors and true-score instability, one can gather data at three different times rather than two and use a path analytic approach.

The author first discusses path analysis as applied to reliability coefficients based on parallel forms. The following path diagram shows the relationship between the conceptual variable (X), the observed score variables (X_1 and X_2), and the random error variables (e_1 and e_2).



The relationship between p_{xx} and the reliability coefficient, r_{12}, is $r_{12} = p_{xx}^2$, indicating that the reliability coefficient in the path analysis (p_{xx}) equals the square root of the traditional reliability coefficient.

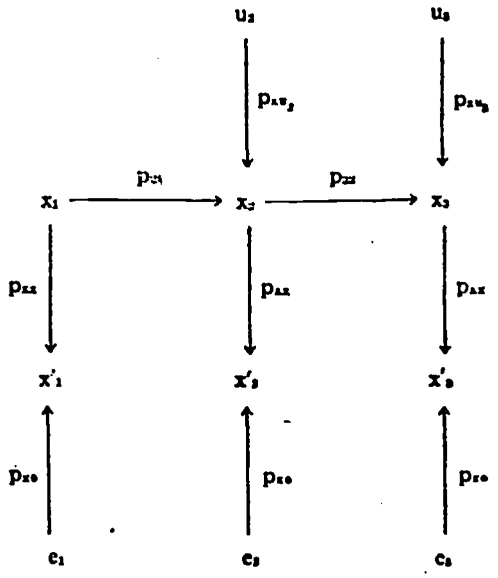
In the test-retest situation, a new variable, U_2, is introduced, and it represents the aggregation of variables affecting X during the interval from the administration of X_1 to the administration of X_2. The path analysis for this situation is as follows:



The path coefficient p_{21} is a measure of the stability of X over time. Following the rules of path analysis, the test-retest correlation is

$r_{12} = p_{xx}p_{21}p_{xx} = p_{xx}^2 p_{21}$ where r_{12} indicates the true correlation between the measured variables x'_1 and x'_2.

Unfortunately, it is impossiblé to solve for pxx because p_{xx} and p_{21} are unknown. To obtain the needed information for the solution, a third testing wave must be given. The situation can be diagrammed as follows:



From the equations for this path model, one finds that:

$$p_{xx}^{2} = r_{xx} = \frac{r_{12}r_{23}}{r_{13}}$$

Since the square of p_{xx} corresponds to the usual reliability coefficient, this last formula provides a new measure of reliability based on test-retest data and free of the effects of temporal change. Furthermore, stability coefficients can be obtained as follows:

$$s_{12} = \frac{r_{13}}{r_{23}}$$

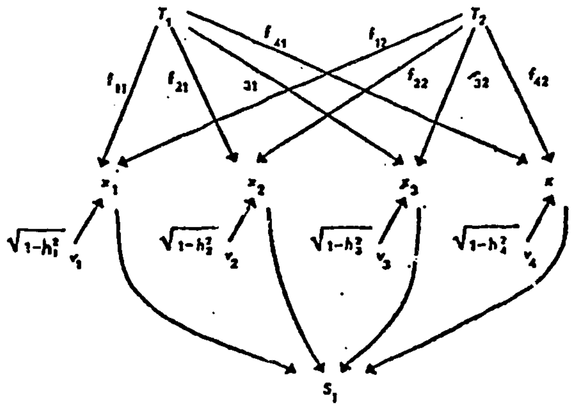
$$s_{23} = \frac{r_{13}}{r_{12}}$$

$$s_{13} = \frac{r_{13}^{2}}{r_{12}r_{23}}$$

The following assumptions must be met in applying this analysis: (1) determination of the index by the underlying variable is constant over time; (2) the rate of instability in the underlying variable is constant between adjacent measurement times; (3) measurement errors are uncorrelated with true scores; (4) measurement errors at different times are uncorrelated with each other; and (5) disturbances at times 2 and 3 are uncorrelated with each other or with the true scores at time 1. The problems of correlated errors and correlated disturbances can be assessed by conducting a fourth wave of measurements to determine whether there is any effect on the reliability and stability estimates.

Heise, David R., & Bohrnstedt, George W. Validity, invalidity, and reliability. In E. F. Borgalla (Ed.), <u>Sociological Methodology, 1970</u>. San Francisco: Jossey-Bass, 1970.

The authors discuss in detail an approach to making parametric estimates from fallible cross-sectional data. The basic procedure involves obtaining multiple measurements on the same underlying "true" variable and using the correlations among these measurements to estimate the values of the parameters. Using path analysis, the authors develop the following latent-trait model that assumes that the inter-item correlations are entirely a function of the latent traits.



Where T_i = the latent traits

X_i = the individual items

f_{iK} = the factor loading of item i on factor K and the validity coefficient for that empirical item $(f_{ik} = p_{ik}\sqrt{p_{ii}1})$

h_i2 = communality for that item

v_i = combined effects of the unique sources of variance for the items and the degree of influence from those unique sources

S_1 = total score

Using this diagram, it is possible to define the correlation between a latent trait and a composite score. For example, the correlation between latent trait T_1 and the score S_1 is the validity of the composite:

$$p_{T_1S_1} = (f_{11}B_{S,1}) + (f_{21}B_{S,2}) + (f_{31}B_{S,3}) + (f_{41}B_{S,4})$$

Generalizing this result, the authors show that

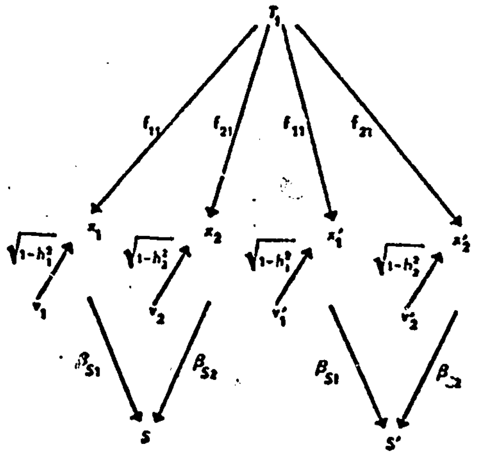
$$p_{T_kS_k} = \frac{\sum_{i=1}^{n} W_{ik}f_{ik}}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ik}W_{jk}P_{ij}}}$$

where W_{ik} = weight assigned to item i to measure factor T_K.

They also show that the following provides a measure of invalidity or the proportion of variance in S_K that is associated with latent variables other than the one of interest.

$$\psi_{S_k}^{2} = 1 - p_{T_kS_k} + \frac{\sum_{i=1}^{n} W_{ik}h_i^{2} - \sum_{i=1}^{n} W_{ik}^{2}}{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ik}W_{jk}p_{ij}}$$

The reliability of a measure can be defined as the correlation between two equivalent forms of a test. The following diagram shows this relationship in terms of the latent trait model, and it indicates that the correlation between the two forms depends only on their mutual dependence on T_1.



The authors then derive the formula for the reliability of composite K as:

$$\Omega = p_{S_k S_k'} = 1 - \frac{\sum_{i=1}^{n} W_{ik}^{2} - \sum_{i=1}^{n} W_{ik}^{2} h_{i}^{2}}{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ik} W_{jk} p_{ij}}$$

After transforming the formulas into the original metric of the items (by substituting the item standard deviations for the weights, w) , they show that

$$\psi_{S_k}^{2} = \Omega - p_{T_k S_k}^{2} .$$

An example of this approach is provided. They find that the Ω estimates exceeded Cronbach's α (since α is a lower bound to reliability) and that reliability can be increased by adding items from a content domain other than the one being measured which has the effect of decreasing the validity.

Herriot, Robert E. Survey research methods. In Robert L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed.). London: MacMillan, 1969.

The article begins with a distinction made between descriptive survey research and explanatory or analytic survey research. In the former, the sample is selected to describe a well-defined population in terms of its characteristics, attitudes, or behaviors; in the latter, the sample is selected to test theoretical relationships and processes within the population. In both types of surveys, plausible inferences can be made through sophisticated reasoning, sampling, instrumentation, data collection, data reduction, and data analysis. The remainder of the article focuses on these five major phases.

In the reasoning phase, the social scientist outlines the central concern; explores previous related research; derives hypotheses, prediction, or explanations; and identifies or constructs indicators to represent concepts being investigated. The sampling phase involves the development of a sample design, including a selection process and an estimation process. The author briefly defines the concepts of population element, element sampling, cluster sampling, and multistage cluster sampling. During the instrumentation and data collection phases, the research objectives become even more concrete. Basic dimensions along which data collection procedures differ are (1) the degree of structure in the questions (e.g., closed versus open-ended questions) and (2) the amount of researcher contact with the respondent (e.g., individual interviews, telephone interviews, group-administered questionnaires, and self-administered questionnaires). The author recommends that pilot studies be conducted before actually using any data collection procedure. The data reduction phase may involve the complex problem of creating reliable and valid indexes of specific concepts. Guttman scaling and principal-components analysis provide two empirical methods for separating "good" indicators from "bad" indicators and for weighting the good indicators in terms of their goodness. The data analysis phase in survey research is critical for the quality of the final product of the research. In addition to examining the effects of the independent or antecedent variable on the dependent or consequent variable, survey research allows an examination of the effects of the control or intervening variables on the independent-dependent variable or the primary relationship. The use of the third variables can be (1) to specify the conditions that differentiate a primary relationship, (2) to explain theoretically why a primary relationship exists or does not exist, and (3) to reveal that a primary relationship is being hidden by the effects of a third variable.

The final sections of this article provide some examples of the survey research method. The bibliography contains 68 references.

How to re-warm your public's support of its schools--and of you. American School Board Journal, October 1973, 160, 20-23.

This article explains how a school district public opinion poll can be of real use to a school board and school administration. The following subjects are discussed in the article: how to determine the necessary factors in a survey, how to design the survey instrument your district will use, and how to conduct the survey. The article stresses that what a good public opinion poll should measure is not the public's opinion of its schools or the public's knowledge of its schools but the exact level of public understanding. Carefully designed and conducted public opinion polls can measure all three factors--public opinion, knowledge, and understanding--with considerable accuracy. The idea is to design a public opinion poll that not only measures what the public thinks but also indicates what the public's thinking means. To accomplish this, a poll should include limited and free-response questions, those that give respondents a series of possible replies from which to select the one most in accord with his or her thinking as well as those that encourage a one or two sentence comment. Examples of questions that meet these qualifications are given.

Howard, George S., Ralph, Kenneth J., Gulanick, Nancy A., Maxwell, Scott E., Nance, Don W., & Gerber, Sterling K. Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement*, 1979, *3*, 1-23.

True experimental designs are thought to control for all sources of internal invalidity. Unfortunately, the possibility of confounding the instrumentation with the experimental treatment exists when self-report instruments are used. This paper presents the results of five studies using self-report instruments to evaluate treatment interventions.

In Study 1, pretest-posttest difference scores revealed an *increase* in dogmatism following a workshop designed to reduce dogmatism. Discussions with workshop participants indicated that the workshop experience changed their perception of their initial level of dogmatism (or a response shift).

This result is examined in Study 2 using the retrospective pretest-posttest design. The results using this latter design were exactly opposite to those obtained using the pretest-posttest difference method. In Studies 3, 4, and 5, the two methods were compared using more objective measures of change. These studies supported the proposition that when self-report measures are used in a pre/post design, the results may be confounded by a response shift. The implications of the findings for evaluation research and the strengths and limitations of retrospective measures are discussed.

The reference section includes 27 citations.

Kaufman, Alan S. Should short-form validity coefficients be corrected? Journal of Consulting and Clinical Psychology, 1977, 45(6), 1159-1161.

This article begins with a short history of the controversy between McNemar and Silverstein regarding the correction of short-form part-whole coefficients for correlated error variance. Kaufman presents a case for correcting the coefficients in some instances but for using the uncorrected correlations in other instances, depending on how the short form is to be used by the practitioner.

Kaufmann stresses that part-whole short-form coefficients, when obtained from a single administration, have correlated errors of measurement that spuriously raise the obtained coefficients. The question to be answered is not whether correction is technically correct but rather which coefficient--the one actually obtained or the one corrected for overlapping errors--is most meaningful in a given situation.

The author contends that in a clinical or school psychological setting the uncorrected correlation is often the appropriate validity coefficient, since it most closely reflects the actual relationship between the short form and full scale. Uncorrected coefficients are most relevant whenever the possibility exists that a child who is tested on the short form will be given the rest of the battery. Corrected correlations are appropriate when the short form is to be used as a replacement for the full scale and the battery is never meant to be given. Corrected correlations are appropriate in such instances because the relevant relationship is the degree to which the short form and full scale are measures of a single construct. The way in which the short form is to be used should determine whether the corrected or uncorrected correlations are used to pick the tests.

One instance in which the corrected coefficients are always pertinent is when defining the psychometric properties of the already selected short form. Regardless of usage of the short form, the correlations with the full scale provide evidence of the criterion-related validity of the abbreviated version. Since the correlated error variance spuriously inflates the pertinent relationship, it is always desirable to correct the coefficients.

A short comment by Silverstein follows the article in which Silverstein supports Kaufman's contentions. McNemar's formula,

$$r_{ts} = \frac{K+\Sigma\Sigma r_{hj}}{\sqrt{n+2\Sigma r_{ij}}\ \sqrt{K+2\Sigma r_{gh}}},$$

is presented along with Silverstein's corrected formula,

$$r' = \frac{\Sigma r_{hh}+\Sigma\Sigma r_{hj}}{\sqrt{n+2\Sigma r_{ij}}\ \sqrt{K+2\Sigma r_{gh}}}.$$

Silverstein states that when the short form is to be used as part of the full scale, it seems reasonable that its validity should approach unity as the number of subtests increases. When the short form is used as a replacement for the full scale, it makes sense that the upper limit to its validity should be the correlation of the full scale with a "comparable form."

Kosa, J., Alpert, J., & Haggarty, R. On the reliability of family health information: A comparative study of mother's reports on illness and related behavior. Social Science and Medicine, 1967, 1(2), 165-181.

Data were collected in a longitudinal study on the health care of 500 low income urban families. The sample was drawn on the basis of a randomized research design from users of the Medical Emergency Clinic of the Children's Hospital Medical Center. Families selected had at least one child in the age group requiring pediatric services and had no private physician regularly providing such services. They had a low income and corresponding education. The information presented in this article was based on a randomly picked 15% subsample of the respondents. Data on the health of the families were collected from medical records and from information furnished by the mother. Three instruments were developed: (1) a questionnaire on the utilization of health facilities; (2) a Child Health Index; and (3) a Family Health Calendar. These instruments were constructed on the basis of previous studies and were designed to collect various kinds of health-related data. Fourteen items were chosen in an attempt to find a single comprehensive and reliable measurement of the health of the family unit, and the Pearson product-moment correlation coefficient was computed for these items. The general lack of consistency between any two instruments suggested that none of the measurements should be regarded as comprehensive and reliable indicators of the health of the family as a unit. Further effort was given to investigate whether the lack of consistency could be attributed to (1) errors of recall or past events, (2) errors of recording daily events in the Health Calendar, or (3) independence of the sets of data referring to the various aspects of family health.

Mothers' reports on clinic visits were checked against clinical records and less than one-fourth of the mothers were found to report the number of visits correctly. An examination of the Child Health Index suggested that norms of current medical relevance influenced mothers in completing reports of symptoms and conditions experienced by their children. Different symptoms and illness were found for each calendar according to the season it represented. Illnesses tended to be of longer duration in the winter than in the summer. Certain norms of censorship were also found to operate in the completion of the Health Calendars.

Lipset, Seymour Martin. The wavering polls. The Public Interest, 1976, 43, 70-89.

The lack of reliability of most survey research results is discussed at length. Examples from Gallup, Harris, the California Poll, and others show the differing response rates to similar questions. Evidence is presented to show the effect of wording on positive and negative response. The author also discusses the lack of specificity in reporting most survey results. Examples are provided of the influential effect of survey results on the decisions of policymakers. In short, the author stresses caution and skepticism in evaluating survey results. He feels that by reemphasizing the instability of many attitudes and preferences of the public polls, the role of judgment and active leadership in decisionmaking may be restored.

Lord, Frederic M. The relative efficiency of two tests as a function of ability level. *Psychometrika*, 1974, *39*, 351-358.

The purpose of the paper is to present the derivation of a formula for determining the relative efficiency of two unidimensional tests measuring the same trait. It expresses relative efficiency solely in terms of the standard errors of measurement and the frequency distributions of true scores. The author shows that the relative efficiency equals:

$$\text{R.E. } \{y,x\} = \frac{\text{Var } (x/\varepsilon)p^2(\varepsilon)}{\text{Var } (y/\eta)q^2(\eta)}$$

where $\eta \equiv \eta(\varepsilon)$ is the equipercentile equivalent of ε.

If x is the number-right score, the range of ε is 0 to n. The author then shows that it is possible to rewrite the above formula in terms of $\dot{s} \equiv \varepsilon/n_x$, $z \equiv x/n_x$, $\omega \equiv \eta/n_y$, and $w = y/n_y$ as follows:

$$\text{R.E. } \{y,x\} = \frac{\text{Var } (z/\dot{s})g^2(\dot{s})}{\text{Var } (w/\omega)h^2(\omega)}$$

where g and h are the density functions of $\dot{s}$ and ω, $0 \leq \dot{s},\ \omega \leq 1$.

The paper ends with an example in which the relative efficiency was computed using the item parameters and using the method described in the paper. Comparison of the two approaches shows the new method provides a good estimate of the relative efficiency of two tests as a function of ability level. The paper includes 12 references.

McMorris, Robert F. Evidence on the quality of several approximations for commonly used measurement statiatics. Journal of Educational Measurement, 1972, 9(2), 113-121.

The author attempts to answer the question, "What is the extent of the error involved with approximations for the standard deviation, internal consistency reliability, and the standard error of measurement?" Approximations for each of the three statistics were computed and compared with corresponding standard statistics. The one-sixth method for the standard deviation provided essentially the same results as did the standard statistic. The Mason-Odeh variation was slightly less biased; with N used in the standard statistic, the bias would probably be reduced. Even the range divided by a consonant provided an estimate that would often be suitable.

The reliability approximations using either the variance or a one-sixth approximation for estimating the variance provided an estimate for KR20 having less bias than was found for KR21, although their relationships with KR20 were slightly lower. For tests of medium difficulty, the errors were small, even when a one-sixth approximation was used to supply the variance.

The standard error was quite accurately estimated by using the number of items as the independent variable. Rounding Lord's constant to .4 resulted in a slightly less biased and less conservative estimate for these data, but to generalize to other sets of data would be hazardous. The results for the reliability and standard error approximations imply that the Saupe and Lord approximations may be slightly short cut to make them even more suitable for non-computationally oriented individuals and for scurrying statisticians.

Novick, Melvin R., & Lewis, Charles. Coefficient alpha and the reliability of composite measures. Psychometrika, 1967, 32, 1-13.

The focus of this paper is to provide a general mathematical treatment of the derivation of the alpha coefficient (α), specifying the necessary and sufficient conditions under which α is equal to the reliability of the test. Adopting Guttman's approach, coefficient α is rederived as a lower-bound on the reliability of a test. By extending this work, a necessary and sufficient condition, under which α is equal to the reliability, is determined. This condition is that the true score random variables T_g and T_h of any pair of components are linearly related such that $T_g = a_{gh} + T_h$ where a_{gh} is a constant. This condition is shown to be closely related to Novick's concept of parallel measurements, and the Kuder and Richardson's unit rank assumption used in deriving formula 20, and it is equivalent to assumptions made by Jackson and Ferguson and by Gulliksen. In examining the property of coefficient α as the "mean of the split-half reliabilities"' the authors pointed out that this refers to the Rulon stepped-up split-half reliability or coefficient α for the test divided into two components and that it does not refer to the Spearman-Brown formula. Finally, one limitation in interpreting any function of α as a measure of internal consistency is that α is sensitive to changes in the scale of the components but not to changes in location.

The bibliography contains 13 references.

Nunnally, Jum C. Psychometric theory. New York: McGraw-Hill, 1967.

This book provides a comprehensive review of psychometric theory including chapters on scaling models, validity, variance and covariance, multivariate correlational analysis, theory of measurement error, assessment of reliability, test construction, fundamentals of factor analysis, and multidimensional scaling, plus chapters on the specific content areas of the measurement of abilities, the measurement of personality traits, the measurement of sentiments, and contingent variables. The remainder of this review will focus on the topics of validity and reliability.

A measuring instrument is valid if it does what it is intended to do. In examining psychological measures, the author identifies three types of validity. Predictive validity is assessed by the degree of correspondence (or correlation) between an instrument and some criterion variable occurring before, during, or after the period when the instrument is applied. Content validity refers to the adequacy with which the particular domain of a content area is sampled. In this case, one is not so much concerned with testing the validity of the measure as with ensuring validity by the plan and procedures of construction. However, evidence for content validity includes (1) a moderate level of internal consistency among the test items, (2) the comparison of performance on a test before and after some training, and (3) the correlation of scores on tests purporting to measure similar things. Construct validity involves a hypothesis that certain behaviors or measures will correlate or that they will be similarly affected by experimental treatments. Construct validity is the extent to which results obtained from two different measures will be the same. Construct measures are developed and validiated by (1) specifying the domain of observables that define the construct, (2) determining to what extent the observables of the construct correlate with each other or are affected alike by experimental treatments, and (3) determining whether or not one, some, or all measures of such variables act as though they measure the construct. If this process is followed, the result should be a construct that is well-defined in terms of a variety of observables, for which there are one or more variables that will represent the domain of observables and that eventually proves to relate strongly with other constructs of interest. In essence, construct validity concerns a hypothesized relationship between a supposed measure of a construct and a particular, observable variable. For example, tests of intelligence should correlate positively with grades in schools, ratings of intelligence by teachers, and level of professional accomplishment.

The author discusses reliability as the extent to which the results of a measurement are repeatable by the same person using different measures of the same attribute or by different persons using the same measure of an attribute. High reliability does not necessarily mean high validity. Any random influence that tends to make measurements differ from occasion to occasion is a source of measurement error. The author proposes the domain sampling model as the most useful model for discussing measurement error. The domain sampling model considers that any particular measure is composed of a random sample of items from a hypothetical domain of items. Measurement error is considered to be present only to the extent that samples are limited in size. Basic to the model is the concept of an infinitely large correlation matrix showing all correlations among items in the domain. The average correlation in the matrix, r_{ij}, would indicate the extent to which some common core existed in the items. If the average correlation among items is assumed to be positive, a very long test would always be a highly reliable test. The degree of reliability estimated by

$$r_{kk} = \frac{k\bar{r}_{ij}}{1 + (k-1)\bar{r}_{ij}}$$

when r_{kk} is the square of the correlation of scores on a collection of items with true scores. Error because of the sampling of items is entirely predictable from the average correlation. Coefficient alpha

$$r_{kk} = \frac{k}{k-1} \left(1 - \frac{\Sigma\sigma i^2}{\sigma y^2}\right)$$

would be the correct measure of reliability for any type of item. The special version of that formula, KR - 20

$$r_{kk} = \frac{k}{k-1}\left(1- \frac{\Sigma pq}{\sigma_y 2}\right)$$

would be used with dichotomous items.

The split-half approach, in which items within a test are divided in half and scores on the two half-tests are correlated, is another estimate of reliability. The difficulty with the split-half method is that the correlation between halves will vary somewhat depending on how the items are divided. Other estimates of reliability are the alternative form method and the retest method, both of which have obvious weaknesses. The alternative forms must not differ systematically in content. The major defect in the retest method is that experience in the first testing will influence responses in the second testing.

Steps that can be taken to prevent measurement error from occurring include writing items clearly, making test instructions easy to understand, and adhering closely to the prescribed conditions for administering and scoring an instrument. The most basic way to make tests more reliable is to make them longer. If the reliability is know for a test with any particular number of items. the following formula can be used to estimate how much the reliability would increase if the number of items were increased by any factor k:

$$r_{kk} = \frac{kr_u}{1 + (k-1)r_{11}}$$

Standards of reliability are stated to be .80 for most tests, but for test scores that are to be used to determine important decisions, a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 is desirable.

Payne, Stanley L. The art of asking questions. Princeton: Princeton University Press, 1951.

The focus of this book is on the use of words in survey questions and on the importance of asking the right question in the right way. The author's writing style in dealing with this complex and often ignored issue is easy and light; however, the subject matter requires careful reading. The author does not attempt to provide a set of definite rules or explicit directions, but instead provides a collection of possible considerations for wording questions.

An example of the importance of wording is the differing response rate to the same question when the words "should," "could," and "might" were interchanged. "Should" found 82% of the respondents agreeing, "could" found 77% agreeing, and "might" found 63% agreeing. Numerous examples are given of the consequences of different wordings in the decennial Census surveys; a change in wording in one survey brought an increase of 1,400,000 responses. The author states that "the most critical need for attention to wording is to make sure that the particular issue that the questioner has in mind is the particular issue on which the respondent gives his answers." The pitfalls of questions that presume too much are discussed in detail, as are free answer questions and their merits, the two-way question and its duplicities, and multiple-choice questions and this misconstruction. Descriptions of special types of questions and their special faults are also discussed.

One chapter is devoted to the care and treatment of respondents. Points to keep in mind include the right of the respondent to refuse to answer, the importance of not "talking down" to the respondent, the use of common sentence construction in questions, the avoidance of overelaboration and double negatives, and the avoidance of trick questions.

Other chapters cover the virtues of brevity and simplicity, "good" words (single in meaning and generally understood), problem words, the characteristics of a loaded question, and the importance of punctuation, emphasis, position of alternatives, easily understood words, and abbreviations of questions.

A visual demonstration of the development of a passable question is presented in Chapter 13. The author begins with a statement of the issue and then point-by-point constructs passable questions that will elicit answers that reflect the issue. A concise checklist of 100 considerations to be used in wording a question is contained in the final chapter; the considerations are enumerated for quick reference.

Schneider, Benjamin. Person/situation selection research: The problem of identifying salient situational dimensions (Research Report No. 13). College Park: University of Maryland, February 1977.

This article deals with the problem of identifying the relevant dimensions or features of the situation in conducting person/situation research. The author argues that too many surveys have been designed from the standpoint of the researcher. He describes a systematic procedure for identifying dimensions with which employees "naively" characterize their work organization. Although the study focuses on relevant dimensions of organizational life, the approach may be useful in developing questionnaire surveys in other areas. We can commend the author's concern about mapping the dimensions of an area, but we would strongly recommend the use of the critical incident technique for this purpose.

Personal interviews were conducted with 67 employees at different levels within a large utility company. The focus of the interview was on finding out "what kinds of things each person thinks about when he or she thinks about the job and the company." These interviews were open-ended and were concerned with identifying dimensions of the situation rather than evaluating these dimensions. Based on the interviews, a list of topics was generated, and these topics were sorted and distilled into 15 major categories. In addition, the dimensions were scored for frequency of mention, importance within each interview, and affect value (positive, negative, or neutral mention). The results can be used to generate more specific questions focusing on the situational dimension.

The bibliography contains 33 references.

Scott, Christopher. Research on mail surveys. Journal of the Royal Statistical Society, 1961, 124, 143-205 (Series H).

The information presented in this article was based on five mail surveys carried out by the Government Social Survey in England. Experimental features introduced into these studies allowed measurement of nonresponse bias, of early/late response bias, of response by non-addressees, of the influences of a variety of factors on the response rate, and of response reliability and validity. An attempt was also made to evaluate all published research on each topic. All five surveys were based on probability samples, three of the general adult population, one of motorcycle owners, and one of residential telephone subscribers.

Only one finding was reported on nonresponse bias: Often enough to be regarded as established, nonrespondents have lower mean educational level. Nonresponse bias in sex, marital status, age, parenthood, occupational status, income, or region of the respondent was tested with no conclusive results. A slight response bias in favor of interest in the subject of the survey was found.

Information on speed of response showed that an interest in the subject of the survey may lead to speedier response. Also, questionnaires that demand more work may not be returned as quickly. A possible rule for predicting speed of response was presented: Taking the total returns to a given mailing after 14 days as the base, and counting the days beginning with the day of receipt of the first batch of returns, about 50% are returned by the end of the third day, about 75% by the end of the fifth day, and about 90% by the end of the tenth day.

An analysis of early versus late responses found that repondents to whom the survey was more relevant (and probably more interesting) tended to reply later. This was explained by the fact that the particular questionnaire required a good deal of work to complete thoroughly. Many research studies have reported a tendency toward earlier response by persons with an interest in the subject of the survey.

Several interesting findings were noted about factors affecting the response rate. Factors found to favor response were:

1. Follow ups
2. Presence or absence of certain questions
3. Offical sponsorship
4. Stamped envelope
5. Special delivery or air mail
6. Handwritten postscript to the covering letter urging reply
7. Letter on back of questionnaire as opposed to separate letter
8. A premium of at least 25¢; larger sums appear to bring no further increases

Some additional findings based on the results of the five surveys showed that the mail survey does not necessarily appear to be less efficient than the interview as a means of collecting information and opinions from the public, unless the questions or their interrelation are complex. Mail survey response rates and validities can be as high as those achieved by interviewers.

A bibliography is included. Appendices and a discussion of Mr. Scott's paper are also provided.

Siegel, Paul M., & Hodge, Robert W. A causal approach to the study of measurement error. In H. M. Blalock, Jr., & A. B. Blalock (Eds.), Methodology in social research. New York: McGraw-Hill, 1968.

The effects of measurement errors on statistics derived from bivariate frequency distributions, such as the correlation coefficient, were examined; the approach involved the application of causal assumptions to models of the relations between measured and true values of variables. The investigation used Census data on years of school completed, occupational SES, and personal income. In explaining the variances in the first two variables, nonrandom errors and a negative correlation exist between the Post Enumeration Survey (PES) or the Current Population Survey (CPS) reports and the errors in census responses (when the PES or CPS are taken as true scores). Furthermore, both floor and ceiling effects appear in SES variables.

The authors first consider the analysis of bivariate statistics when true scores are known. They present a series of path models including (1) one having random errors with the CPS and the PES as a standard of accuracy, (2) one allowing correlated errors but with independence of errors between variables and using the CPS-PES as a standard of accuracy, (3) one allowing correlated errors with partial independence of errors between variables and using the CPS-PES as a standard of accuracy, and (4) one allowing correlated errors with the CPS-PES as a standard of accuracy. These models show that, when the CPS or PES reports are taken as true scores, errors of measurement have only a slight net effect on the correlation between education and income as observed in the census.

The authors then focus on the analysis of bivariate statistics when true scores are unknown; in this case, the CPS-PES data are considered to be imperfect sources of data. In the first model, the true scores (E_t - true level of education and I_t = true level of income) are shown as explicit causes of the census reports (E_c and I_c) and CPS reports (E_1 and I_1). The census and CPS report, both having random errors, are independent measures of the corresponding true scores.

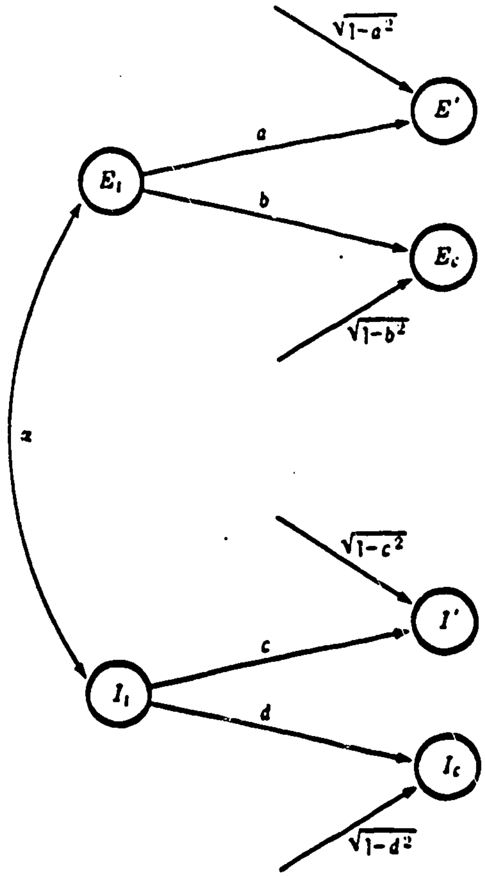


Fig. 2.5. ,Correction for Attenuation with Independent Measures of Different Quality.

In the special case of Figure 2.5 where a = b and c = d, then

$$r_{E_t I_t} = \frac{r_{E'I'}}{\left(r_{I'I_c}\right)^{1/2} \left(r_{E'E_c}\right)^{1/2}}$$

which is the correction for attenuation. By introducing an elaboration of the correction for attenuation, such that errors are allowed to be correlated with the true and measured scores for different variables or among themselves, the causal model becomes impossible to solve and inconsistent with the data.

The authors conclude that several avenues exist for further research. Although more researchers have been concerned with the reliability of a single item or test, of greater importance is an evaluation of the effects of errors in measurement on the observed association between different items. A second area for future research involves a determination of the relative contributions of the various sources of error (e.g., interviewer questioning, respondent misreporting, interviewer misrecording, improper coding or keypunching, and computer mishandling) and the evaluation of causal models showing the interrelationships of these errors. A third area concerns the development of causal models to simultaneously handle questions of reliability and validity. A final research focus involves the causal representation of alternative strategies of measuring the reliability of tests.

The bibliography includes 33 references.

Smedley, Rande H., & Olson, George H. Graduate follow-up studies: How useful are they? Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975. (ERIC No. 109 431)

Two general approaches to follow-up studies were discussed. The first approach, studying graduates of the previous year, was criticized as being inferior to the second approach, identifying a current class of seniors and following them through successive years following graduation.

Kerlinger (1974) distinguished survey research from status surveys. Status surveys were defined as routine fact and data collections. Significant biases on status surveys were found through the use of a reliability and validity "resurvey" technique on a subsample of participants in a National Longitudinal Study (Eihternacht, 1974). Survey research, on the other hand, is focused on uncovering relationships among important variables and lives up to rigorous scientific standards.

Two follow-up studies were presented and compared. The first, the Texas Educational Product Study (TEPS), studied graduates of the previous year, and the second, Project TALENT, identified a current class of seniors and followed them through successive years following graduation. TEPS is called a follow-up technique and Project TALENT a follow-through technique. Several biases were cited in follow-up studies, including (1) lack of control over independent variables (post hoc fallacy), (2) lack of item validity and reliability, (3) sampling biases (non-observation bias), and (4) observation bias (Heisenberg principle).

Because the follow-up (TEPS) strategy collects all data at one time, data are limited to descriptive accounts of status. In contrast, the follow-through strategy provides a longitudinal data base that can produce a baseline on important variables. In these respects, follow-through surveys come much closer to producing information relevant to relationships of concern to decisionmakers and therefore prove to be more worthy of the time and effort put into them.

Stanley, Julian C. Reliability. In Robert L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Reliability refers to the tendency toward consistency from one set of measurements to another, while unreliability results because repeated sets of measurements never exactly duplicate each other. The following formula indicates that the reliability coefficient equals the proportion of variance in the measure that is due to true differences within that particular population.

$$P_{ff'} = \frac{\sigma_T^{\,2}}{\sigma_X^{\,2}} = 1 - \frac{\sigma_e^{\,2}}{\sigma_X^{\,2}}$$

However, the magnitude of reliability coefficients is dependent on the dispersion of true ability in the group being measured; so, restriction of range lowers reliability. The table shown below identifies possible sources of variation in the measure.

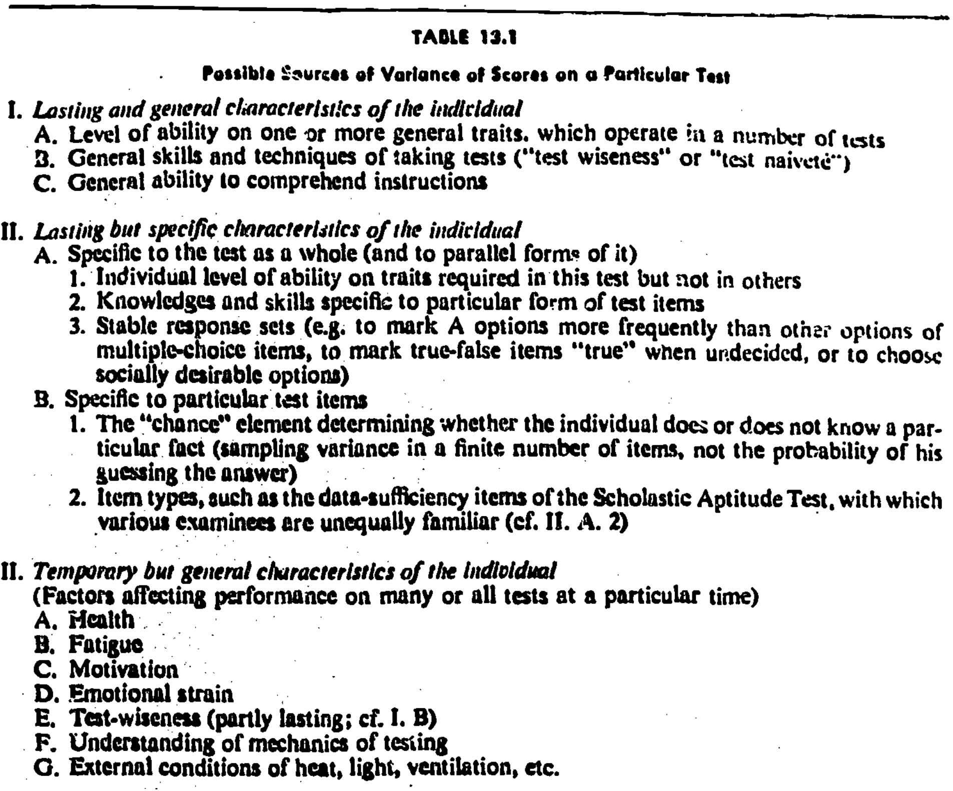


TABLE 13.1

Possible Sources of Variance of Scores on a Particular Test

I. *Lasting and general characteristics of the individual*
   A. Level of ability on one or more general traits, which operate in a number of tests
   B. General skills and techniques of taking tests ("test wiseness" or "test naiveté")
   C. General ability to comprehend instructions

II. *Lasting but specific characteristics of the individual*
   A. Specific to the test as a whole (and to parallel forms of it)
      1. Individual level of ability on traits required in this test but not in others
      2. Knowledges and skills specific to particular form of test items
      3. Stable response sets (e.g. to mark A options more frequently than other options of multiple-choice items, to mark true-false items "true" when undecided, or to choose socially desirable options)
   B. Specific to particular test items
      1. The "chance" element determining whether the individual does or does not know a particular fact (sampling variance in a finite number of items, not the probability of his guessing the answer)
      2. Item types, such as the data-sufficiency items of the Scholastic Aptitude Test, with which various examinees are unequally familiar (cf. II. A. 2)

III. *Temporary but general characteristics of the individual*
   (Factors affecting performance on many or all tests at a particular time)
   A. Health
   B. Fatigue
   C. Motivation
   D. Emotional strain
   E. Test-wiseness (partly lasting; cf. I. B)
   F. Understanding of mechanics of testing
   G. External conditions of heat, light, ventilation, etc.

IV. *Temporary and specific characteristics of the individual*
   A. Specific to a test as a whole
      1. Comprehension of the specific test task (insofar as this is distinct from I. B)
      2. Specific tricks or techniques of dealing with the particular test materials (insofar as distinct from II. A. 2)
      3. Level of practice on the specific skills involved (especially in psychomotor tests)
      4. Momentary "set" for a particular test
   B. Specific to particular test items
      1. Fluctuations and idiosyncrasies of human memory
      2. Unpredictable fluctuations in attention or accuracy, superimposed upon the general level of performance characteristic of the individual

V. *Systematic or chance factors affecting the administration of the test or the appraisal of test performance*
   A. Conditions of testing—adherence to time limits, freedom from distractions, clarity of instructions, etc.
   B. Interaction of personality, sex, or race of examiner with that of examinee to facilitate or inhibit performance
   C. Unreliability or bias in grading or rating performance

VI. *Variance not otherwise accounted for (chance)*
   A. Luck in selection of answers by sheer guessing
   B. Momentary distraction

In the latter portion of the paper, the author discusses four major procedures for obtaining the reliability coefficient:

1. Correlation of the resulting scores from the administration of two parallel forms under specified conditions.

2. Correlation of the resulting scores from repeated administration of the same test form or testing procedure.

3. Correlation of the resulting two scores from subdivision of a single test into two parallel groups of items (halfforms a and b). If the reliability coefficient of this half test is r_{ab}, the reliability coefficient of the whole test is $2r_{ab}/(1 + r_{ab})$.

4. Analysis of the covariance among individual items and determination of the true-score and error variance therefrom.

The author proceeds to provide the formula for each of these approaches (almost 70 formula altogether) and the derivation of those formula.

The bibliography contains 167 references.

Subkoviak, Michael J., & Levin, Joel R. Fallibility of measurement and the power of a statistical test. Journal of Educational Measurement, 1977, 14, 47-52.

This paper is concerned with the effects of measurement error in the dependent variable. In particular, a description is presented on the use of the reliability coefficient in estimating values of required sample size for planning a study or of available power for evaluating a completed study, and methods are suggested for estimating the pooled within-treatment reliability coefficient P_{yy}'.

In previous work, the authors presented the following formula for ψ_σ, which represented any linear contrast involving K population means that the researcher specified as important to detect.

$$\psi_\sigma = \frac{\sum_{k=1}^{K} a_K u_K}{\sigma_w(T)}$$

where K is the number of treatments, u_k are the populations means, $\sigma_w(T)$ is the common within-treatment standard deviation of true scores, and a_k are coefficients chosen such that $\sum_{k=1}^{K} a_k = 0$.

Using the value obtained ψ_σ, one can calculate ϕ, which is a parameter in the Pearson-Hartley power charts (appearing in most experimental design testbooks).

$$\phi = \sqrt{\frac{P_{yy}' n \psi_\sigma^2}{(V_1+1) \sum_{k=1}^{K} a_k^2}}$$

where P_{yy}' is the reliability within each treatment group, $K=V_1+1$ (the number of treatment groups), and n = subjects per treatment group.

In an example, the authors show that as the reliability within eac reatment group, P_{yy}', decreases, the effect is to reduce ϕ and, likewise, power. Given a specified desired power, the process can be reversed to determine the required sample size. Again the sample shows that, as P_{yy}' decreases, the sample size needed for a given level of power increases.

To estimate P_{yy}', one can use data from previous research using the same instrument on a similar group of subjects, or one can conduct a pilot study. For a single treatment group, the within-treatment reliability can be estimated using Kuder-Richardson Formula 20, KR20 (for a single administration) or the Pearson product-moment formula (for dual administration). When there is more than one treatment, the same formulas can be used with estimates of within-treatment variance and covariance pooled from the various groups. For the pooled single administration estimate of P_{yy}', the following analog of KR20 is appropriate:

$$\hat{p}_{yy'} = \frac{I}{I - 1} \left[\frac{\sum_{k=1}^{1} \hat{\sigma}_p^2(Y_1)}{\hat{\sigma}_p^2(Y)} \right],$$

where:

$\hat{\sigma}_p^2(Y_i)$ = the estimated within-treatment variance of scores Y_i on the i^{th} item pooled across the K treatments

$$= \frac{\sum_{k=1}^{K} n_k \hat{\sigma}_k^2(Y_i)}{\sum_{k=1}^{K} n_k - K}$$

and $\hat{\sigma}_p^2(Y)$ = the estimated within-treatment variance of total scores Y pooled across the K treatments

$$= \frac{\sum_{k=1}^{K} n_k \hat{\sigma}_k^2(Y)}{\sum_{k=1}^{K} n_k - K}.$$

For the pooled dual-administration estimate of $\rho_{YY'}$, the following is appropriate:

$$\hat{\rho}_{YY'} = \frac{\sum_{k=1}^{K} n_k \hat{\sigma}_k(Y,Y')}{\sqrt{\left[\sum_{k=1}^{K} n_k \hat{\sigma}_k^2(Y)\right]\left[\sum_{k=1}^{K} n_k \hat{\sigma}_k^2(Y')\right]}}$$

where $\hat{\sigma}_k(Y,Y')$, $\hat{\sigma}_k^2(Y)$, and $\hat{\sigma}_k^2(Y')$ are again biased estimates of the covariance and variances of total scores Y and Y', based on the n_k subjects in condition k.

The reference section includes 12 citations.

Taylor, Elaine N., and others. Procedures for surveying school problems: Some individual, group, and system indicators. A manual. Alexandria, Va.: Human Resources Research Organization, 1974. (ERIC No. 106 375)

This manual presents information on the preparation and conduct of a survey of mental health problems in schools. The instruments used were two questionnaires and a guide for an interview by a mental health consultant with a school principal.

Part I of this document discusses the instruments and their application. The statistical procedures used in analyzing the data were as follows:

1. Tabulate the responses to each item.
2. Find the mean rating for each item.
3. Find the overall mean rating and the overall variance of the responses (for the questionnaire as a whole).
4. Standardize the mean item scores by transforming them into T-scores with a mean of 50 and a standard deviation of 10.
5. Find a mean T-score for each problem area.
6. Order the areas from high to low on the basis of their mean T-Scores.
7. Order the items from high to low on the basis of their T-scores.

Each of the steps is shown in detail in Appendix A.

Two steps were used to predict the most potentially relevant problems: (1) identification of the most salient problems, as determined by participant responses, and (2) establishment of priorities for these problems, which were determined by the extent to which a problem disrupted the educational goals of a school.

Part II of the document deals with the reliability and validity of measures. The authors state that the reliability of an instrument is concerned primarily with two questions:

1. Are the scores telling me anything? (internal consistency)

2. Can I rely over time on what the scores tell me? (test-retest)

Measures of internal consistency were confined to data collected within a school. The reliabilities obtained were .95 for the staff questionnaire and .82 for the student questionnaire. The instruments were most reliable if there was something to discriminate and if the number of respondents was large. In short, more reliable data were obtained in schools with severe problems while reliability measures tended to be low in schools with no problems.

The authors' main concern with the validity of the instruments was to determine their "utility for some purpose." Data can be checked or validated against some external criterion: against a less fallible source of information or some ultimate measure of utility. Cross-validation was useful in schools with severe problems. Schools with few or no problems showed little cross-correlation, presumably because the problems did not affect everyone.

The report contains four appendices: Directions for Hand Tabulation and Analysis of Questionnaire Data, Notes on Punching Cards and Running Programs for Analysis of Staff and Student Data, FORTRAN Programs for Analyzing Data from Staff and Student Questionnaires, and Notes on Development of the Instruments.

Toole, Patrick F., Campbell, Paul B., & Beers, Joan S. Educational quality phase II findings assessment: Reliability and validity. Harrisburg: Pennsylvania Department of Education, August 1970.

This article contains a general discussion on reliability and validity as well as a more detailed discussion of the adequacy of the educational quality assessment (EQA) instruments. Reliability is defined as a measure of consistency or the degree to which the same results can be obtained again. Reliability can be measured using one of the following techniques:

1. Split-half coefficients in which scores from one-half of a test are correlated with scores from the other half;
2. Test-retest coefficients in which correlations are obtained for the test scores of a sample taken at two different times; and
3. Internal consistency coefficients that can be calculated by the split-half method, the item to total score correlation, or analysis of variance.

To improve the reliability of the EQA battery, each section of the battery went through three stages of reliability computations. Based on the results of the previous stage, items were reworded, deleted, or added. These decisions were based on item to total score correlations, proportions of students omitting an item, discrimination indices, and factor analysis results.

Validity indicates whether the item really measures what it is supposed to measure, and it is assessed by three methods. Content validity relates to the fact that the items must reflect the test's purposes, and it is assessed by observational analysis and content analysis as undertaken by a panel of independent experts. Criterion-related validity is measured by the correlation between the test under scrutiny and an independent test measuring the same variable. Construct validity is inferred from a predicted network of relationships and can be assessed using factor analysis.

Seven references are included in the bibliography.

Trow, Martin. Survey research and education. In Charles Y. Glock (Ed.), Survey research in the social sciences. New York: Russell Sage Foundation, 1967.

The author discusses the natural appeal of survey research to educational sociology; survey research in texts on methods in educational research; descriptive, explanatory, and "pseudo" surveys; problems and potentialities of survey research in education; problems of survey research among teachers and administrators; when not to ask "why?"; potential areas for survey research in education; and the relevance of survey research to educational practice.

Survey research strongly recommends itself to those interested in educational research because structured questionnaires can be administered to a convenient, captive, compliant, and literate population. Unfortunately, this ease of access sometimes results in other techniques and procedures being neglected or ignored. The author contends that the logic of survey research--problems of design, of analytical strategies, or of the interdependence of the elements in research--receives little consideration. Explanations of descriptive, explanatory, and pseudo surveys are provided. Pseudo surveys are defined as studies in which the authors are neither interested in saying something about the specific population from which the subject sample was drawn nor in uncovering social relationships and processes. The author feels that more studies of this nature are found in educational research than in other fields. He goes on to discuss the importance of interpreting findings and not merely reporting them. The findings of several educational research surveys are discussed, and potential areas for survey research are presented. In conclusion, the author states the purpose of this review of survey research in education "is to contribute to sound knowledge and fruitful theory about the institutions and processes of education." Suggestions to improve educational survey research include becoming familiar with advances in survey research methodology, analysis of survey data, sample design, and questionnaire construction; maintaining a sociological perspective; and being prepared to be wrong, thereby encouraging the researcher to take risks.

West, Leonard J. Design and conduct of educational surveys and experiments (Service Bulletin No. 2). St. Peter, Minn.: Delta Pi Epsilon, January 1977.

This article, written for the novice researcher, reviews some of the major problem areas to be addressed when designing and conducting educational surveys and experiments. It is divided into four major sections: (1) considerations for survey research, (2) considerations for experiments, (3) requirements for research reporting, and (4) impact of research on practice.

The first section begins with a discussion of the dimensions concerning questions amenable to survey methods. Survey questions should not solicit opinions on issues that are not matters of opinion; they should point to outcomes that will result in some practical action; they should supply findings that are generalizable; and they should indicate the incidence, distribution, and interrelationships of variables in the population. All data collection instruments should be pretested to identify ambiguous wording. Structured questions (e.g., multiple-choice items) facilitate response interpretation, coding, and analysis; however, on opinion questions addressed to sophisticated audiences, it may be best to use open-ended questions. At the very least, appropriate responses could be determined from a pilot test, and an open-ended response (e.g., "Other, please describe") could be provided. As a check on the clarity of questions and the reliability of responses, the researcher can build redundancy into the survey form; comparisons can then be made on the items of greatest interest, and the respondent can be contacted to explain the discrepancy. The next decision to be made concerns whether the survey will include the entire population (a census) or some portion of the population (a sample); unfortunately, the author does not provide any guidelines for making this decision. Having decided to draw a sample from the population, the researcher must then be concerned with (1) accessibility, (2) representativeness, (3) stratification, and (4) sample size. Nonresponse is an important factor affecting the reliability of any large-scale survey; at the very least, the researcher should determine whether any significant differences exist between the respondents and the nonrespondents. One approach that can be used when responses have been received over a long period of time involves a comparison of early respondents and late respondents; if no significant differences are found, then it may be assumed that differences do not exist between respondents and nonrespondents. Another approach involves determining whether the nonresponse related directly to the purposes of the survey. If the sample was drawn from a known population, the sample characteristics must be compared with the population characteristics.

The second section deals with methodological concerns in experiments. The topics discussed in this section include (1) the definition, purposes, and criteria of experiments; (2) weakness found in past business education experiments (excessive length, failure to find significant differences, and one-teacher experiments); (3) critical features of experiments (sampling, wording of hypotheses, adequate controls, valid and reliable criterion measures, statistical requirements); and (4) replication as a basis for external validity.

The third section on research reporting is very brief and merely indicates that such reports should (1) provide explicit details on the procedures, (2) discuss the findings in relation to previous studies, and (3) avoid excessive speculation beyond one's findings. The final section, also very short, makes a plea for the use of research findings in the development of teacher education.

The bibliography contains 21 references.

Wiseman, Frederick. Measurement problems in the calculation of rates of response and nonresponse. Paper presented at the annual meeting of the American Psychological Association, New York, September 1979.

The author discussed a current research study in which the ultimate objective is the standardization of response and nonresponse terminology. Examples of definitions currently being used for response rates are given. One example of the problem in computing a response rate for a mail survey cited the possibility of three denominators that could be used: (1) total number of potential respondents, (2) total number of questionnaires mailed, and (3) total number of questionnaires returned because of an incorrect address. The use of the latter denominator in a survey will yield a higher response rate than if either denominator (1) or (2) is used. The author offers no solution to this problem, but he does cite the need for standardization of the response rate definition.

Withey, Stephen B. Reliability of recall of income. Public Opinion Quarterly, 1954, 18, 197-204.

The author begins by discussing the factors affecting the validity and reliability of reports on income: (1) the view that such information is personal, (2) the fact that people receive income from several sources, and (3) the fact that incomes are shared. Using data from interviews with 655 adults living in urban areas of the United States, he investigated whether annual income was reliably recalled over a one-year period and whether direction of errors was systematic. In comparing 1948 and 1949 reports of 1947 income, the regression slope was less than 1.00 (i.e., .94) and the intercept was greater than 0 (i.e., $182). This evidence indicated that the recall was neither accurate nor random. Furthermore, the correlation between the magnitude of income change (indicated by the difference between current reports from two different years) and the magnitude of unreliability (indicated by the difference between the first current report and the recall report) was 0.58; thus, as the income change increased, the unrelibility in the recall report increased.

[illegible] *Quality of statistical data.* Rome: Food and Agriculture Organization of the United Nations, 1966.

This book is comprised of material developed by the author for use in seminars and training sessions organized by the Food and Agriculture Organization (FAO). The purpose of the seminars and the book was to increase awareness of the quality problem of statistical data. Toward this end, the author began by describing various kinds of errors such as (1) errors resulting from inadequate preparations (including biased procedures and biased tools), (2) errors committed during data collection (including listing error, missing data, and response or observational errors), and (3) processing errors (including errors in editing, coding, punching, and tabulation). The discussion then turned to a consideration of unbiased and biased estimates. The author concluded that, all else being equal, the estimator (or the method of arriving at the estimator) should be the one having the smallest mean square error.

To reduce errors and to examine the quality of the data, the author described two different approaches: (1) the use of post hoc techniques and (2) the use of sampling methods. In the first approach, the researcher can compare the results with data from independent sources, can compare the results with some generally accepted knowledge about the characteristics or their relationships (i.e., consistency checks), can examine the internal consistency or the degree to which estimates of different characteristics describe the same phenomenon in the same way, and can determine the expected cohort survival. Unfortunately, post hoc techniques have several limitations: (1) data from independent sources may not be accurate; (2) previously collected data on the same topic may not be available; (3) application of such techniques results in impressions on quality rather than in numerical measures; and (4) techniques refer to final survey results, and thus, individual errors are lost. Sampling methods, or quality checking, can overcome many of these problems: (1) they can be applied whether or not accurate data have been collected previously on the topic; (2) they provide numerical estimates; and (3) they can provide estimates of the quality for any part of the population and for any item in the survey. In conducting quality checks, the researcher should be examining the data to determine the existence of the following properties: (1) zero bias, (2) stability, (3) internal consistency, (4) compatibility with existing knowledge, (5) use of correct procedures, and (6) use of correct tools.

The next two chapters dealt with errors resulting from inadequate preparations. In terms of procedures, bias may enter in the measurement, in the selection of the sample, or in the estimation. As for the tools, bias may enter in the use of random numbers that are not actually random (i.e., digits do not occur with equal frequency; digit series do not occur with equal frequency; digit series do not appear with a certain expectation; and gaps of a specified length between the same two digits do not occur with a certain expectation), the preparation of inadequate questionnaires (through improper or unclear wording; use of problem words, vague concepts, or abbreviations; use of loaded or loading questions; development of an overly lengthy questionnaire; and use of inadequate publicity), the development of inaccurate sampling frames (through out-of-date or inaccurate lists, careless work), the inadequate use of the sampling frames (because of confusion of units, confusion of populations, inaccurate information about accurately listed units, and wrong assumptions about the structure of the population), and the development of inadequate instructions (ambiguous, incomplete, too long, or too short).

The next six chapters focused on the problems arising from errors committed during data collection. As for listing errors, the author was concerned with the magnitude of net errors rather than that of gross errors. To check on listing errors, a compact cluster sampling technique can be used in which specific units from randomly selected clusters are relisted. To determine the net error, the "erroneously included" units are added and the "omissions" are subtracted. For larger surveys, it may be necessary to calculate the net error associated with the listing of clusters as well as the net error associated with the listing of units. Measures for improving the quality of the listing included (1) preparing or improving the mapping material, (2) carefully selecting and training the enumerators, (3) using publicity to gain cooperation, (4) employing intensive supervision of enumerators, (5) using a short period for the enumeration, (6) using sample checks, (7) studying listing errors, and (8) providing monetary incentives. As for missing data, nonresponse, or incomplete samples, the survey results lose precision and may be biased. To overcome nonresponse problems, the author suggested (1) recontacting all nonrespondents, (2) recontacting a sample of nonrespondents, and using a multiphase approach of mail questionnaires and personal interviews. Whenever refusals occur, however, all available information on those persons must be collected for later use in data imputation. Factors affecting the quality of the data that are obtained included the respondent's knowledge background, social background, emotional background, and memory, and the length of the period of reference. In a self-enumeration survey, response error can be determined by checking the responses to certain questions against data available from other sources or gathered at another time. In a survey using enumerators or interviewers, the enumerator/interviewer error can be assessed using an analysis of variance paradigm to determine the "between-enumerators" and "within enumerators" between respondents "effects." To improve the enumerator's work, the author recommended that careful selection procedures be used (taking into account the size of the survey, sponsorship, available pay, and availability of qualified candidates), that appropriate, efficient training be provided (including the purpose of the survey, details of the work, role-playing of the procedures, and potential problems), and that intensive supervision be employed. Quality checking of the enumerator's work can be accomplished through a "reinterview" conducted either before or after the major survey.

The next chapter centered on checking the quality of the processing. Post hoc checking is concerned with determining the quality level that has been achieved, while process checking is concerned with controlling the quality of data processing while the work is in progress. The author argued for using process checking as an integral part of planning for data processing.

The final chapter emphasized the importance of rational survey design. Such survey designing includes (1) analysis of the survey into constituent elements (when needed for separate study and planning); (2) careful study of the factors affecting each constituent element; and (3) integration of the elements into a coherent, efficient, and convenient design. From this viewpoint, the conduct of a pilot study can provide needed facts in terms of either quantitative or qualitative estimates. An extensive bibliography of relevant material is provided.

Zimmerman, Donald W. Error and reliability in stochastic processes and psychological measurement. Psychological Reports, 1972, 31, 131-140.

The author argues that the concepts of random error and reliability of measurement can be represented using probability theory. By doing so, one can avoid using the notions of "true values" and "errors" inherent in traditional theories. Using a probability model, formulas relating observable events are derived from probability axioms and from primitive terms that refer to observable events. The benefits of such an approach are that (1) the model is more economical in language and formalism, (2) it is more general than the traditional approach, (3) it applies to stochastic processes in which joint distributions of many dependent random variables are of interest, and (4) it clarifies problems involving the "experimental independence" of measurements and the relation of sampling of individual to sampling of measurements.

The bibliography contains five citations.

Author Index

DATA COLLECTION METHODOLOGY

Erdos, 1970, 1979
Foster & Neal, 1975
Gentry & Milliken, 1975
Scott, 1961
Trow, 1967
Zarkovich, 1966

NONRESPONSE

Frankel, 1979
Frankel & Dutca, 1979
Gentry & Milliken, 1975
Wiseman, 1979

QUESTIONNAIRE DESIGN

Berdie & Anderson, 1974
Bradburn, 1970
Bradburn & Seedman, 1979
Duckworth, 1973
How to Re-warm Your Public's Support of Its Schools--and of You, 1973
Lipset, 1976
Payne, 1951

RELIABILITY

American Psychological Association, 1974
Armor, 1974
Bell, Lin, & Warheit, 1977
Blalock, 1969
Bohrnstedt, 1969
Brooks & Bailer, 1978
Conger & others, 1976
Cronbach, 1951
Cronback, Gleser, Nanda, & Rayaratnam, 1972
Cronbach, Rajaratnam, & Gleser, 1963
Cureton, Cook. Fischer, Laser, Rockwell, & Simmons, 1973
Erickson, McDonald, & Gunderson, 1971
Ferber, 1966
Foster & Neal, 1975
Havighurst, 1970
Heise & Bohrnstedt, 1970
Heise, 1969
Kosa, Alpert, & Haggarty, 1967
Lipset, 1976
Lord, 1974
McMorris, 1972
Novick & Lewis, 1967
Nunnally, 1967
Siegel & Hodge, 1968
Stanley, 1971
Subkoviak & Levin, 1977
Taylor & others, 1974
Toole, Campbell, & Beers, 1970
Withey, 1954
Zimmerman, 1972

RESEARCH DESIGN

Barber, 1973

Bell, Lin, & Warheit, 1977

Frankel, 1979

Frankel & Dutca, 1979

Herriot, 1969

How to Re-warm Your Public's Support of Its Schools--and of You, 1973

Smedley & Olson, 1975

West, 1977

Zarkovich, 1966

VALIDITY

American Psychological Association, 1974

Bell, Lin, & Warheit, 1977

Blalock, 1969

Bohrnstedt, 1969

Brooks & Bailer, 1978

Campbell & Fiske, 1959

Conger & others, 1976

Cronbach, 1971

Cronbach & Meehl, 1955

Foster & Neal, 1975

Heise & Bohrnstedt, 1970

Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979

Kaufman, 1971

Nunnally, 1967

Siegel & Hodge, 1968

Taylor and others, 1974

Toole, Campbell, & Beers, 1970