THE

TERNATIONAL ENCYCLOPEDIA

OF

EDUCATION

Research and Studies

Volume 8

S

Editors-in-Chief

TORSTEN HUSEN

T. NEVILLE POSTLETHWAITE

THE

INTERNATIONAL ENCYCLOPEDIA OF EDUCATION

Research and Studies

Volume 8
S

Editors-in-Chief

TORSTEN HUSEN
University of Stockholm, Sweden

T. NEVILLE POSTLETHWAITE
University of Hamburg, FRG

PERGAMON PRESS

OXFORD · NEW YORK · TORONTO · SYDNEY · PARIS · FRANKFURT

U.K.	Pergamon Press Ltd., Headington Hill Hall, Oxford OX3 0BW, England
U.S.A.	Pergamon Press Inc., Maxwell House, Fairview Park, Elmsford, New York 10523, U.S.A.
CANADA	Pergamon Press Canada Ltd., Suite 104, 150 Consumers Rd., Willowdale, Ontario M2J 1P9, Canada
AUSTRALIA	Pergamon Press (Aust.) Pty. Ltd., P.O. Box 544, Potts Point, N.S.W. 2011, Australia
FRANCE	Pergamon Press SARL, 24 rue des Ecoles, 75240 Paris, Cedex 05, France
FEDERAL REPUBLIC OF GERMANY	Pergamon Press GmbH, Hammerweg 6, D-6242 Kronberg-Taunus, Federal Republic of Germany

First edition 1985

Library of Congress Cataloging in Publication Data
Main entry under title:
The International encyclopedia of education

Includes bibliographies.
Index: v. 10.
1. Education—Dictionaries. 2. Education—Research—Dictionaries. I. Husén, Torsten, 1916– II. Postlethwaite, T. Neville.
LB15.I569 1985 370'.3'21 84–20750

British Library Cataloguing in Publication Data
The International encyclopedia of education

1. Education—Dictionaries

I. Husén, Torsten II. Postlethwaite, T. Neville
370' 3'21 LB15

ISBN 0-08-028119-2

Computer data file designed and computer typeset by Page Bros (Norwich) Ltd.
Printed in Great Britain by A. Wheaton & Co. Ltd., Exeter

S

Safety Education

Safety education is the study of those human, machine, and environmental variables which interact to affect the probability of injury or illness to people or damage to property; it embraces a host of situations involving people, such as work, recreation, sport, transportation, home, and natural and human-created disasters; it encompasses not only the safe production of goods or delivery of services but the integrity of the products themselves in the consumers' environment.

The modern study of safety has taken much from the discipline of public health and the epidemiological techniques associated with it. This approach recognizes that accidents are one of the most serious of public health problems, that their causes are usually complex, that they are foreseeable, and that they are not caused solely by the acts of people because machines and environmental factors, in their broadest sense, are also usually involved.

One need for the formal study of accidents developed as an outgrowth of laws relating to working conditions and workers' compensation insurance programs in Europe and the United States. This movement of the late nineteenth and early twentieth century created an awareness of the human toll and costs of occupational injuries and provided the motivation to reduce accidents. The National Safety Council was formed in the United States in 1912 and provided educational services covering the extent and variety of safety problem areas. Educational programs were started in some schools in the United States in the 1920s, and the Center for Safety Education at New York University was founded in 1938, with Herbert J. Stack as its director, to conduct research and provide training of safety professionals. Amos E. Neyhart of Pennsylvania State University pioneered driver education for high-school students in the United States.

In 1970 the United States legislature passed the Occupational Safety and Health Act (OSHA) which resulted in the promulgation of many safety and health standards in the United States industries, upheld by inspection of work places by OSHA personnel. In addition to inspections initiated by request, usually by workers, OSHA personnel periodically inspect at random those industries (and now also government facilities) having the potential for health hazards or relatively high accidental injury reports. A key aim of the OSHA program is to educate workers and management in various aspects of health and safety in industry, mining, and construction. Comparable legislation exists in England (Health and Safety at Work Act 1974), France, the Federal Republic of Germany, and Sweden (Working Environment Act 1977). Each of these acts also requires education and training of workers in safe procedures.

The insurance companies that offer workers' compensation coverage have pioneered safety and health training for workers and management. They also provide consultation to reduce work accidents for the mutual benefit of insurer and policy holder.

Various professional organizations, such as the American Society of Safety Engineers, the American Industrial Hygiene Association, the Safety Systems Society, and England's Royal Society for the Prevention of Accidents, among others, serve a strong educational function by publishing technical journals, holding meetings, and offering short courses.

In the elementary-school curriculum, safety education is often woven into other subjects by examples of safe human behavior in such activities as crossing streets, riding bicycles, or using seatbelts in vehicles, and by fire drills. Safety education in United States high schools culminates in driver education, usually consisting of 30 hours in the classroom and 6 hours of driving. The latter is sometimes augmented by some hours in simulators and on ranges away from other traffic to learn how to control the car. In most countries, except the United States, driver education is done informally or by commercial schools (OECD 1976).

A number of research studies on the effects of driver education on accidents have been done in the United States (McGuire and Kersh 1971) and in England (Raymond et al. 1973). Those studies that have used appropriate control groups have not found consistent benefits of driver education on measures related to accidents or violations. There are also extensive programs now in motorcycle rider education and training for novice riders, using curricula based on a task-analysis approach (McKnight and Heywood 1974) in the United States; other courses are available on a more limited scale for experienced riders. Similar programs exist in the United Kingdom.

Some universities offer degree programs in safety. In the United States, the Board of Certified Safety Professionals (BCSP 1982) has recommended a curriculum for the baccalaureate degree in safety, which puts a heavy emphasis on the physical sciences, mathematics, communications skills, human factors/ergonomics, and basic concepts of industrial safety and

hygiene. This suggested curriculum reflects the growing technical complexity of the problems confronting safety professionals. The methods used to control hazards (e.g., nuclear power) are often technically complex and require sophisticated methods of analysis to forecast the risks.

A major aspect of safety education is to teach the methods of collecting data that provide indices of the level of safety and measures of the exposure of people, so that the risk associated with various situations can be quantified. Quantification of the level of safety is necessary to determine if a need exists for corrective action, to indicate the kinds of corrections that should be applied, and to evaluate their effectiveness. Thus, for example, the American National Standards Institute Z16.1 defines some aspects of industrial accidents in quantitative terms such as by frequency and severity of injury rates. These measures are augmented by analyses of costs and benefits to realize the most benefit for the financial investment and by cost–effectiveness analysis to choose the most effective corrective action for the cost invested. Systems analytic techniques, such as fault-tree analysis, are now a part of the education of safety professionals.

While much emphasis is still being placed upon the education of safety personnel in basic concepts related to the elimination of hazards—such as machine guarding, materials handling, and fire safety—the development of new technologies imposes increasing demands upon the education of the safety staff in sophisticated techniques for the control and reduction of injuries and illnesses caused by hazards in the human environment.

Bibliography

Board of Certified Safety Professionals (BCSP) 1981 *Curricula Guidelines for Baccalaureate Degree Programs in Safety*. BCSP, Champaign, Illinois

McGuire F L, Kersh R C 1971 *An Evaluation of Driver Education: A Study of History, Philosophy, Research Methodology, and Effectiveness in the Field of Driver Education*. University of California Press, Los Angeles, California

McKnight A J, Heywood H B 1974 *Motorcycle Task Analysis*. National Public Services Research Institute, Alexandria, Virginia

Organisation for Economic Co-operation and Development (OECD) 1976 *Driver Instruction*. OECD, Paris

Raymond S, Jolly K W, Risk A W, Shaoul J E 1973 *An Evaluation of the Effectiveness of Driver Education in Reducing Accidents to Young People*. University of Salford, Salford

R. Mortimer

Sampling

Social science research is aimed at developing useful generalizations about society and the ways in which individuals behave in society. However, due to practical constraints on research resources, the social scientist is usually limited to the study of a sample rather than a complete coverage of the population for which these generalizations are appropriate. Provided that scientific sampling procedures are employed the use of a sample often provides many advantages compared with a complete coverage: reduced costs associated with obtaining and analyzing the data, reduced requirements for specialized personnel to conduct the fieldwork, greater speed in most aspects of data manipulation and summarization, and greater accuracy due to the possibility of closer supervision of fieldwork and data preparation.

Kish (1965) has divided the social science research situations in which samples are used into three broad categories: (a) experiments—in which the treatment variables are deliberately introduced and all extraneous variables are either controlled or randomized; (b) surveys—in which all members of a defined population have a known nonzero probability of selection into the sample; and (c) investigations—in which data are collected without either the randomization of experiments or the probability sampling of surveys. Experiments are strong with respect to internal validity because they are concerned with the question of whether a true measure of the effect of a treatment variable has been obtained for the subjects in the experiment. In contrast, surveys are strong with respect to external validity because they are concerned with the question of whether the findings obtained for the subjects in the survey may be generalized to a wider population. Investigations are weak on both types of validity and their use is due frequently to convenience or low cost.

In educational research, the survey and experimental approaches have often been portrayed as quite separate methodologies. The perceived differences between these approaches have not been a consequence of statistical theory but rather would appear to be associated with the degree of control which the researcher may exert over the educational environment. Educational researchers have rarely been placed in the enviable situation of being able to introduce experimental treatments in an independent fashion, with appropriate allowances for extraneous variables, into randomly selected portions of a large and dispersed population. Consequently, the practical difficulties involved in the design of educational research experiments so as to investigate causal relationships within specific populations have often resulted in questions of sample design being largely ignored.

The following discussion of sample design for educational research has focused on some aspects of the survey approach and its application to large-scale educational studies. However, the issues which have been raised have direct bearing on the conduct of experimental studies because the distributions of relationships between characteristics in causal

systems, like the distributions of these characteristics taken alone, exist only with reference to particular populations.

1 Populations

The populations which are of interest to educational researchers are generally finite populations that may be defined jointly with the elements that they contain. A population in educational research is therefore, usually, the aggregate of a finite number of elements, and these elements are the basic units that comprise and define the population.

Kish (1965) stated that a population should be described in terms of (a) content, (b) units, (c) extent, and (d) time. For example, in a study of the characteristics of Australian secondary-school students, it may be desirable to specify the populations as: (a) all 14-year-old students, (b) in secondary schools, (c) in Australia, (d) in 1985.

In order to prepare a description of a population to be considered in an educational research study it is important to distinguish between the population for which the results are desired, the desired target population, and the population actually covered, the survey population. In an ideal situation these two populations would be the same. However, differences may arise due to noncoverage: for example, for the population described above, a list may be compiled of schools during early 1985 which accidentally omits some new schools which begin operating later in the year. Alternatively, differences may occur because of nonresponse at the data collection stage. For example, a number of schools having large concentrations of educationally retarded students might be unwilling to participate in the study (see *Data Analysis: Nonresponse*).

Strictly speaking, only the survey population is represented by the sample, but this population may be difficult to describe exactly and therefore it is often easier to write about the defined target population (Kish 1965). The defined target population description provides an operational definition which is used to guide the construction of a list of population elements, or sampling frame, from which the sample may be drawn. The elements that are excluded from the desired target population in order to form the defined target population are referred to as the *excluded population*.

For example, during a cross-national study of science achievement carried out in 1970 by the International Association for the Evaluation of Educational Achievement (IEA), one of the desired Australian target populations for the study was described as:

> All students aged 14.0–14.11 years at the time of testing. This was the last point in most of the school systems in IEA where 100 percent of an age group were still in compulsory schooling. (Comber and Keeves 1973 p. 10)

In Australia it was decided that, for certain administrative reasons, the study would be conducted only within six states of Australia and not within the smaller Australian territories. It was also decided that only students in those school grade levels which contained the majority of 14-year-old students would be tested.

The desired Australian target population was therefore reformulated in order to obtain the defined Australian target population:

> All students aged 14.0–14.11 years on 1 August 1970 in the following Australian states and secondary-school grades:
>
> | | |
> |---|---|
> | New South Wales | Forms I, II, and III |
> | Victoria | Forms I, II, III, and IV |
> | Queensland | Grades 8, 9, and 10 |
> | South Australia | 1st year, 2nd year, and 3rd year |
> | West Australia | Years 1, 2, and 3 |
> | Tasmania | Years I, II, III, and IV. |

The majority of students in the excluded population were 14-year-olds who were in grade levels which were outside the ranges specified in the description of the defined target population. The students in the "other territories" of Australia (Australian Capital Territory and Northern Territory) were excluded because of certain administrative and cost constraints which were placed on the study.

2. Sampling Frames

Before selecting the sample, the elements of the defined target population must be assembled into a sampling frame. The sampling frame usually takes the form of a physical list of the elements, and is the means by which the researcher is able to "take hold" of the defined target population. The entries in the sampling frame may refer to the individual elements (for example, students) or groups of these elements (for example, schools).

In practice, the sampling frame is more than just a list because the entries are normally arranged in an order which corresponds to their membership of certain strata. For example, in a series of large-scale studies of educational achievement carried out in 21 countries during the early 1970s (Peaker 1975), sampling frames were constructed which listed schools according to their size (number of students), type (for example, comprehensive or selective), region (for example, urban or rural), and sex composition (single sex or coeducational). The use of strata during the preparation of a sampling frame is often undertaken in order to ensure that data are obtained which will permit the researcher to study, and more accurately assess, the characteristics of both individual and combined strata.

3. Probability Samples and Nonprobability Samples

There are usually two main aims involved in the conduct of sample surveys in educational research: (a) the estimation of the values of population attributes (parameters) from the values of sample attributes (statistics), and (b) the testing of statistical hypotheses about population characteristics. These two aims require that the researcher has some knowledge of the accuracy of the values of the sample statistics as estimates of the population parameters. Knowledge of the accuracy of these estimates may generally be derived from statistical theory provided that probability sampling has been employed. Probability sampling requires that each member of the defined target population has a known, and nonzero, chance of being selected into the sample. The accuracy of samples selected without using probability sampling methods cannot be discovered from the internal evidence of a single sample.

Nonprobability sampling in educational research has mostly taken the form of judgment sampling in which expert choice is used to guide the selection of typical or representative samples. These samples may be better than probability samples, or they may not. Their quality cannot be determined without knowledge of the relevant population parameters and if these parameters were known then there would be no need to select a sample.

The use of judgment samples in educational research is sometimes carried out with the (usually implied) justification that the sample represents a hypothetical universe rather than a real population. This justification may lead to research results which are not meaningful if the gap between this hypothetical universe and any real population is too large. Since nonprobability samples are not appropriate for dealing objectively with the aims of estimation and hypothesis testing, they will not be examined in the following discussion.

4. Accuracy, Bias, and Precision

The sample estimate derived from any one sample is inaccurate to the extent that it differs from the population parameter. Generally, the value of the population parameter is not known and therefore the actual accuracy of an individual sample estimate cannot be assessed. Instead, through a knowledge of the behaviour of estimates derived from all possible samples which can be drawn from the population by using the same sample design it is sometimes possible to assess the probable accuracy of the obtained sample estimate.

For example, consider a random sample of n elements which is used to calculate the sample mean, $\bar{x}$, as an estimate of the value of the population mean, μ. If an infinite set of independent samples of size n were drawn from this population and the sample mean calculated for each sample then the average of the sampling distribution of sample means, the expected value, could be denoted by $E(\bar{x})$.

The accuracy of the sample statistic, $\bar{x}$, as an estimator of the population parameter, μ, may be summarized in terms of the mean square error (MSE). The MSE is defined as the average of the squares of the deviations of all possible sample estimates from the value being estimated (Hansen et al. 1953).

$$\begin{aligned} \text{MSE}[\bar{x}] &= E[\bar{x} - \mu]^2 \\ &= E[\bar{x} - E(\bar{x})]^2 + [E(\bar{x}) - \mu]^2 \\ &= \text{Variance of } \bar{x} + [\text{Bias of } \bar{x}]^2 \end{aligned} \tag{1}$$

A sample design is unbiased if $E(\bar{x}) = \mu$. It is important to remember that "bias" is not a property of a single sample, but of the entire sampling distribution, and that it belongs neither to the selection nor the estimation procedure alone, but to both jointly.

The reciprocal of the variance of a sample estimate is commonly referred to as the precision, whereas the reciprocal of the mean square error is referred to as the accuracy.

For most well-designed samples in educational survey research, the sampling bias is either zero or small—tending towards zero with increasing sample size. The accuracy of sample estimates is therefore generally assessed in terms of the sampling variation of the values of $\bar{x}$ around their expected value $E(\bar{x})$.

4.1 The Accuracy of Individual Sample Estimates

The educational researcher is usually dealing with a single sample of data and not with all possible samples from a population. The variance of a sample estimate as a measure of sampling accuracy cannot therefore be calculated exactly. Fortunately, statisticians have derived some formulas which provide estimates of the variance based on the internal evidence of a single sample of data.

For a simple random sample of n elements drawn without replacement from a population of N elements, the variance of the sample mean may be estimated from a single sample of data by using the following formula (Kish 1965 p. 41):

$$\text{var}(\bar{x}) = \frac{N - n}{N} \frac{s^2}{n} \tag{2}$$

where $s^2 = \Sigma(x_i - \bar{x})^2/(n - 1)$ is an unbiased estimate of the variance of the element values, x_i, in the population.

Note that for sufficiently large values of N, the variance of the sample mean may be estimated by s^2/n because the finite population correction, $(N - n)/N$, tends to unity.

In many practical survey research situations, the sampling distribution of the estimated mean is

approximately normally distributed. The approximation improves with increasing sample size even though the distribution of elements in the parent population may be far from normal. This characteristic of the sampling distribution of the sample mean is associated with the "central limit theorem" and it occurs not only for the mean but for most estimators commonly used to describe survey research results (Kish 1965).

From a knowledge of the properties of the normal distribution, it is possible to be "68 percent confident" that the range $\bar{x} \pm \sqrt{[V(\bar{x})]}$ includes the population mean, where $\bar{x}$ is the sample mean obtained from one sample from the population. The quantity $\sqrt{[V(\bar{x})]}$ is called the standard error, $\text{SE}(\bar{x})$, of the sample mean, $\bar{x}$. Similarly, it is known that the range $\bar{x} \pm 1.96\ \text{SE}(\bar{x})$ will include the population mean with 95 percent confidence. The calculation of confidence limits for estimates allows researchers to satisfy the estimation aim of survey research. Also, through the construction of difference scores $d = \bar{x}_1 - \bar{x}_2$, and using a knowledge of the standard errors $\text{SE}(\bar{x}_1)$ and $\text{SE}(\bar{x}_2)$, the statistical hypothesis aim may be satisfied.

It should be remembered that, although this discussion has focused on sample means, confidence limits could also be set up for many other population values, which, for example, are estimated by $\bar{v}$, in the form $\bar{v} \pm t\sqrt{[V(\bar{v})]}$. The quantity t represents an appropriate constant which is usually obtained from the normal distribution or under certain conditions from the t distribution. For most sample estimates encountered in practical survey research, assumptions of normality lead to errors that are small compared to other sources of inaccuracy.

5. *Multistage Sampling*

A population of elements can usually be described in terms of a hierarchy of sampling units of different sizes and types. For example, a population of school students may be seen as being composed of a number of classes each of which is composed of a number of students. Further, the classes may be grouped into a number of schools.

The hypothetical population of school students in Fig. 1 shows 18 students distributed among six classrooms (with three students per class) and three schools (with two classes per school).

From this population a multistage sample could be drawn by randomly selecting two schools at the first stage, followed by randomly selecting one classroom from each of the selected schools at the second stage, and then randomly selecting two students from each selected classroom at the third stage. This three-stage sample design would provide a sample of four students. It would also provide a sample which is an epsem sample (equal probability of selection method) (Hansen et al. 1953). That is, the probability of selecting any student in the population would be

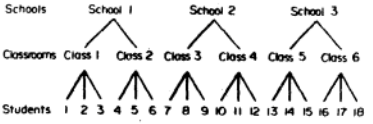

Figure 1
Hypothetical population of eighteen students grouped into six classrooms and three schools

the same for all students ($2/3 \times 1/2 \times 2/3 = 4/18$). Similarly, a simple random sample of four students from the population of 18 students would also provide an epsem sample in which the probability of selection would be the same for all students (4/18). Epsem sampling is widely used in survey research because it usually results in self-weighting samples. In these samples an unbiased estimate of the population mean may be obtained by taking the simple average of the sample cases.

It is important to remember that the use of probability sampling does not automatically lead to an epsem sample. Probability sampling requires that each element in the population has a known and nonzero chance of selection which may or may not be equal for all elements. There are many examples in the literature which demonstrate that educational researchers often overlook this point. For example, one popular sample design in educational research has been to select a simple random sample of, say, a schools from a list of A schools, and then select a simple random sample of b students from each selected school.

The probability of selecting a student by using this design is ab/AB_i, where B_i is the size of the ith school in the population. Consequently, students from large schools have less chance of selection and the simple average of sample cases may result in biased estimates of the population mean—especially if the magnitudes of the B_i values vary a great deal and the survey variable is correlated with school size.

6. *Stratification*

The technique of stratification is often employed in the preparation of sample designs for educational survey research because it generally provides increased precision in sample estimates without leading to substantial increases in costs. Stratification does not imply any departure from probability sampling—it simply requires that the population be divided into subpopulations called strata and that the random sampling be conducted independently within each of these strata. The sample estimates of population parameters are then obtained by combining the information from each stratum.

Stratification may be used in survey research for

reasons other than obtaining gains in sampling precision. Strata may be formed in order to employ different sample designs within strata, or because the subpopulations defined by the strata are designated as separate domains of study. Some typical variables used to stratify populations in educational research are: school location (metropolitan/rural), type of school (government/nongovernment), school size (large/medium/small), and sex of pupils in school (males only/females only/coeducational).

Stratification does not necessarily require that the same sampling fraction is used within each stratum. If a uniform sampling fraction is used then the sample design is known as a proportionate stratified sample because the sample size from any stratum is proportional to the population size of the stratum. If the sampling fractions vary between strata then the obtained sample is a disproportionate stratified sample. The simple random sample design is called a self-weighting design because each element has the same probability of selection equal to n/N. For this design, each element has a weight of $1/n$ in the mean, 1 in the sample total, and $F = 1/f$ in the population total, where $f = n/N$ is the uniform sampling rate for all population elements (Kish 1965 p. 424).

In a stratified sample design of elements, different sampling fractions may be employed in the defined strata of the population. The chance of an element appearing in the sample is specified by the sampling fraction associated with the stratum in which that element is located. The reciprocals of the sampling fractions, which are sometimes called the raising factors, describe how many elements in the population are represented by an element in the sample. At the data analysis stage either the raising factors, or any set of numbers proportional to them, may be used to assign weights to the elements. The constant of proportionality makes no difference to the sample estimates. However, in order to avoid confusion for the readers of survey research reports, the constant is usually selected so that the sum of the weights is equal to the sample size.

For example, consider a stratified sample design of n elements which is applied to a population of N elements by selecting a simple random sample of n_h elements from the hth stratum containing N_h elements. In the hth stratum the probability of selecting an element is n_h/N_h, and therefore the raising factor for this stratum is N_h/n_h. That is, each selected element represents N_h/n_h elements in the population.

The sum of the raising factors over all n sample elements is equal to the population size. If there are two strata for the sample design then:

$$\left(\frac{N_1}{n_1} + \frac{N_1}{n_1} + \ldots \text{ for } n_1 \text{ elements}\right) + \left(\frac{N_2}{n_2} + \frac{N_2}{n_2} + \ldots \text{ for } n_2 \text{ elements}\right) = N \quad (3)$$

In order to make the sum of the weights equal to the sample size, n, both sides of the above equation will have to be multiplied by a constant factor of n/N. That is:

$$\left(\frac{N_1}{n_1} \cdot \frac{n}{N} + \ldots \text{ for } n_1 \text{ elements}\right) + \left(\frac{N_2}{n_2} \cdot \frac{n}{N} + \ldots \text{ for } n_2 \text{ elements}\right) = n \quad (4)$$

Therefore the weight for an element in the hth stratum is $N_h/n_h \cdot n/N$.

An estimate of the variance of the sample mean, $\bar{x}_{st}$, for the stratified random sample design described above may be obtained from the following formula (Kish 1965 p. 81):

$$\text{var}(\bar{x}_{st}) = \sum_h \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h} \quad (5)$$

where

$$s_h^2 = \sum_i (x_{hi} - \bar{x}_h)^2/(n_h - 1)$$

is the variance of the simple random sample of n_h elements in the hth stratum.

Note that for fixed values of n, n_h, N, and N_h, the precision depends upon the sum of the s_h^2 values across strata. If the stratification procedures are extremely successful then element values within strata will be very similar and consequently the magnitude of $\text{var}(\bar{x}_{st})$ will be small. For the special case of proportionate stratified random sampling of elements, the values of n_h/N_h are equal to n/N for all strata. The element weight in this special case is 1 for all sample elements.

Kish (1965 p 88) has listed several aspects of a research study which benefit from using proportionate random sampling of elements from the strata: (a) sampling precision—the variance of the sample estimate of the mean cannot be greater than for an unstratified sample of the same size; (b) administration—proportionate allocation can typically be done simply and easily; and (c) analysis—proportionate allocation generally leads to self-weighting designs.

7. The Comparison of Sample Designs

In a previous section it was shown that, for the hypothetical population in Fig. 1, either a three-stage sample design or a simple random sample design could be used to select epsem samples of the same size. However, equality of selection probabilities in the two designs provides no guarantee that the variances of sample estimates obtained from each design will be the same.

Fisher (1922) suggested that sample designs could be described and compared in terms of their efficiency. For example, one sample design, denoted i,

may be compared to another sample design, denoted j, by considering the inverse of the variance of sample estimates for the same sample size. Using E to represent the efficiency of a sample design for the sample mean, and n to represent the sample size, the efficiency of these two sample designs can be compared by constructing the following ratio:

$$\frac{E_i}{E_j} = \frac{\text{Var}(\bar{x}_j)}{\text{Var}(\bar{x}_i)} \quad (n_i = n_j) \tag{6}$$

More recently, Kish (1965) recommended that the simple random sample design should be used as a standard for quantifying the efficiency of other types of more complex sample designs. Kish introduced the term Deff (design effect) to describe the ratio of the variance of the sample mean for a complex sample design, denoted c, to the variance of a simple random sample, denoted *srs*, of the same size:

$$\text{Deff} = \frac{\text{Var}(\bar{x}_c)}{\text{Var}(\bar{x}_{srs})} \quad (n_c = n_{srs}) \tag{7}$$

The values of Deff for sample means and multivariate statistics, such as correlation coefficients and regression coefficients, have been found to be greater than unity for many sample designs which are commonly used in educational survey research (Peaker 1975, Ross 1978).

8. The Effective Sample Size

Complex sample designs may also be compared to simple random sample designs by calculating the value of the effective sample size (Kish 1965 p. 259) or the simple equivalent sample (Peaker 1967 p. 149). For a given complex sample, the effective sample size is, for the variable under consideration, the size of the simple random sample which would have the same variance as the complex sample. For example, consider a population of N students. If a complex sample design is used to select an epsem sample of n_c students, then the variance of the sample mean, $\text{Var}(\bar{x}_c)$ may be written as:

$$\text{Var}(\bar{x}_c) = \text{Deff} \cdot \text{Var}(\bar{x}_{srs}) \quad (n_c = n_{srs}) \tag{8}$$

Or, alternatively, since $n_c = n_{srs}$, this expression may be written in the form presented by Kish (1965 p. 258):

$$\text{Var}(\bar{x}_c) = \text{Deff} \cdot \frac{N - n_c}{N} \cdot \frac{S^2}{n_c} \tag{9}$$

where S^2 is the population variance.

Now consider a simple random sample design which is used to select a sample of n^* elements from the same population of students. Let the variance of the sample mean for this sample, $\text{Var}^*(x_{srs})$, be equal to the variance of the sample mean for the complex sample design, $\text{Var}(\bar{x}_c)$. That is, $\text{Var}(\bar{x}_c) = \text{Var}^*(\bar{x}_{srs})$.

Substituting on both sides gives the following:

$$\text{Deff} \frac{N - n_c}{N} \cdot \frac{S^2}{n_c} = \frac{N - n^*}{N} \cdot \frac{S^2}{n^*} \tag{10}$$

If N is large compared to n_c or n^*, then $n^* = n_c/\text{Deff}$ is the effective sample size for the complex sample design.

It is important to recognize that in complex sample designs the sampling precision is a function of the whole sample design and not just the total sample size. In order to make meaningful comparisons of the sampling precision of complex sample designs, the design effects must be compared in association with the total sizes of the complex samples.

9. Simple Two-stage Cluster Sampling

In educational research, a complex sample design is often employed rather than a simple random sample design because of cost constraints. For example, a two-stage sample consisting of the selection of 10 schools followed by the selection of clusters of 20 students within each of these schools would generally lead to smaller data collection costs compared with a simple random sample of 200 students. The reduced costs occur because the simple random sample may require the researcher to collect data from as many as 200 schools. However, the reduction in costs associated with the complex sample design must be balanced against the potential for an increase in the variance of sample estimates. The selection of groups of students at the first stage in a two-stage sample design is referred to as cluster sampling. Cluster sampling involves the division of the population into clusters which serve as the initial units of selection.

The variance of the sample mean for the simple two-stage cluster sample design depends, for a given number of clusters and a given ultimate cluster size, on the value of the intraclass correlation coefficient. This coefficient is a measure of the degree of homogeneity within clusters. In educational research, student characteristics are generally more homogeneous within schools than would be the case if students were grouped at random. The homogeneity of individuals within sampling units may be due to common selective factors, or to joint exposure to the same influence, or to mutual interaction, or to some combination of these. It is important to remember that the coefficient of intraclass correlation may take different values for different populations, different clustering units, and different variables.

Consider a population of elements divided into equal-sized clusters. Firstly, a simple random sample can be drawn of size n from the population. Secondly, a two-stage sample of the same size can be drawn from the population by using simple random sampling to select m clusters, and then for each of the selected clusters by using simple random sampling to

select $\bar{n}$ elements, so that the total sample size n is given by: $n = m \cdot \bar{n}$. The relationship between the variances of the sampling distributions of sample means for these two designs is (Kish 1965 p. 162):

$$\text{Var}(\bar{x}_c) = \text{Var}(\bar{x}_{sr}) [1 + (\bar{n} - 1) \cdot roh] \qquad (11)$$

where $\text{Var}(\bar{x}_c)$ is the variance of the sampling distribution of sample means for the simple two-stage cluster design; $\text{Var}(\bar{x}_{sr})$ is the variance of the sampling distribution of sample means for the simple random sample design; $\bar{n}$ is the ultimate cluster size; and *roh* is the coefficient of intraclass correlation.

By transposing the above equation, the value of the design effect for the simple two-stage cluster sample design may be written as a function of the ultimate cluster size and the coefficient of intraclass correlation:

$$\text{Deff} = \frac{\text{Var}(\bar{x}_c)}{\text{Var}(\bar{x}_{sr})} = 1 + (\bar{n} - 1)roh \qquad (12)$$

Since *roh* is generally positive (for students within schools and students within classrooms) the precision of the simple two-stage cluster sample design (which uses either schools or classrooms as primary sampling units) will generally result in sample means which have larger variance than for a simple random sample design of the same size. The losses in sampling precision associated with the two-stage design must therefore be weighed against the "gains" associated with reduced costs due to the selection and measurement of smaller numbers of primary sampling units.

Experience gained from large-scale evaluation studies carried out in many countries (Peaker 1967, 1975) has shown that *roh* values of around 0.2 provide reasonably accurate estimates of student homogeneity for achievement variables within schools. Higher values of *roh* for achievement variables have been noted in Australia when considering student homogeneity within classrooms (Ross 1978). These higher values for students within classrooms are sometimes due to administrative arrangements in school systems. For example, students could be allocated to classrooms by using ability streaming within schools, or there may be substantial differences between classroom learning environments within schools.

10. Estimation of the Coefficient of Intraclass Correlation

The coefficient of intraclass correlation was developed in connection with studies carried out to estimate degrees of fraternal resemblance, as in the calculation of the correlation between the heights of brothers. To establish this correlation there is generally no reason for ordering pairs of measurements obtained from any two brothers. The initial approach to this problem was the calculation of a product-moment correlation coefficient from a symmetrical table of measures consisting of two interchanged entries for each pair of measures. This method is suitable for small numbers of entries—however the number of entries in the table rises rapidly as the number of pairs increases.

Some computationally simpler methods for calculating estimates of this coefficient have been described by Haggard (1958). The most commonly used method appears to have been based on using one-way analysis of variance where the clusters which define the first-stage sampling units, for example schools or classrooms, are regarded as the "treatments". The between clusters mean square, BCMS, and the within clusters mean square, WCMS, are then combined with the number of elements per cluster, $\bar{n}$, to obtain the estimate of *roh*:

$$\text{estimated } roh = \frac{\text{BCMS} - \text{WCMS}}{\text{BCMS} + (\bar{n} - 1)\,\text{WCMS}} \qquad (13)$$

An alternative formula, which is based upon variance estimates for elements and cluster means has been presented by Ross (1983):

$$\text{estimated } roh = \frac{\bar{n}s_c^2 - s^2}{(\bar{n} - 1)s^2} \qquad (14)$$

where s_c^2 is the variance of the cluster means; s^2 is the variance of the elements; and $\bar{n}$ is the ultimate cluster size.

Both of these formulas assume that the data have been collected by using simple two-stage cluster sampling, and also that both the number of elements and the number of clusters in the population are large.

11. Sample Design Tables for Simple Two-stage Cluster Sample Designs

The two-stage cluster sample design is probably the most often used sample design in educational research. Generally this design is employed by selecting either schools or classes at the first stage of sampling, followed by the selection of either students within schools or students within classes at the second stage. In many research situations these sample designs will be less expensive than simple random sample designs of the same size. Also, they offer an opportunity for the researcher to conduct analyses at higher levels of data aggregation. For example, the selection of clusters of students according to their membership of classes would allow the researcher, provided there were sufficient numbers of classes and sufficient numbers of students per class in the sample, to create a data file based on class mean scores and then to conduct analyses at the "between-class" level (see *Units of Analysis*).

The previous discussion showed that the precision of the simple two-stage cluster design relative to a simple random sample design of the same size was a function of $\bar{n}$, the ultimate cluster size, and *roh*, the

coefficient of intraclass correlation. With a knowledge of both of these statistics, in combination with the required level of sampling precision, it is possible to establish a planning equation which may be used to guide decisions concerning the appropriate numbers of first- and second-stage sampling units.

For example, consider an educational research study in which test items are administered to a sample of students with the aim of estimating the item difficulty values and mean test scores for the population. If a simple random sample of n^* students is selected from the population in order to calculate the proportion p who have obtained the correct answer on an item, the variance of p as an estimate of the population difficulty value may be estimated from the following formula (Kish 1965 p. 46):

$$\text{var}(p) = \frac{p(1-p)}{n^* - 1} \tag{15}$$

This formula ignores the finite population correction factor because it is assumed that the population is large compared to the sample size.

If it is specified that the standard error of p, expressed as a percentage, should not exceed 2.5 percent, then by assuming normality this would give $p \pm 5$ percent as 95 percent confidence limits for the population value. The maximum value of $p(1-p)$ occurs for $p = 50$. Therefore in order to ensure that these error requirements could be satisfied for all items, it is necessary to require that

$$(2.5)^2 \geq \frac{50(100-50)}{n^* - 1} \tag{16}$$

That is, n^* would have to be greater than or (approximately) equal to 400 in order to obtain 95 percent confidence limits of $p \pm 5$ percent.

The variance of a sample mean obtained from a simple random sample which is greater than or equal to 400 in size would be less than or equal to $s^2/400$. Also, the standard error of the sample mean would be less than or equal to $s/20$. Assuming normality, this would give a 95 percent confidence band of ± 10 percent of a student standard deviation score when the sample mean is used as an estimate of the population mean.

Now consider the size of a simple two-stage sample design which would provide equivalent sampling accuracy to a simple random sample of 400 students. That is, it is necessary to discover the numbers of primary sampling units (for example, schools or classes) and the numbers of secondary sampling units (students) which would be required in order to obtain 95 percent confidence bands of ± 5 percent for item difficulty estimates, and ± 10 percent of a student standard deviation score for test mean estimates.

From previous discussion, the relationship between the size of a complex sample, n_c, which has the same accuracy as a simple random sample of size $n^* = 400$ may be written as:

$$n^* = \frac{n_c}{Deff} = 400 \tag{17}$$

Since the complex sample is a simple two-stage cluster sample design, the value of Deff may be replaced by $1 + (\bar{n} - 1)roh$ in the above expression to obtain the planning equation:

$$n_c = 400[1 + (\bar{n} - 1)roh] = m\bar{n} \tag{18}$$

where *roh* is the coefficient of intraclass correlation for the student measure which is being considered; m is the number of primary selections; and $\bar{n}$ is the number of secondary selections within each primary selection.

It is important to remember that the planning equation is derived with the assumption that the two-stage sample design fits the model of a simple two-stage cluster sample design. In practical educational research studies sample designs may depart from this model by incorporating such complexities as the use of stratification prior to sample selection, and/or the use of varying probabilities of selection at each of the two stages of sampling. Consequently the planning equation must be seen as a tool which assists with the selection of a sample design, rather than a precise technique for predicting sampling errors. The actual sampling accuracy of a sample design must be determined after the sample data become available for analysis.

As an example, consider $roh = 0.2$ and $\bar{n} = 10$. Then,

$$\begin{aligned} m &= \frac{400}{\bar{n}}[1 + (\bar{n} - 1)roh] \\ &= \frac{400}{10}[1 + (10 - 1)0.2\} \\ &= 112 \end{aligned} \tag{19}$$

That is, for $roh = 0.2$, a simple two-stage cluster design of 1,120 students consisting of 112 primary selections followed by the selection of 10 students per primary selection would be required to obtain accuracy which is equivalent to a simple random sample of 400 students.

In Table 1, the planning equation has been employed to list sets of values for $\bar{n}$, m, and n_c which describe a group of simple two-stage cluster sample designs that have equivalent sampling accuracy to a simple random sample of 400 students. Two sets of sample designs have been listed in the table corresponding to *roh* values of 0.2 and 0.4.

The most striking feature of Table 1 is the rapidly diminishing effect that increasing $\bar{n}$, the cluster size, has on m, the number of clusters which must be selected. This is particularly noticeable for both values of *roh* when the cluster size reaches 10 to 15

Table 1
Sample design table for simple two-stage cluster samples having an equivalent sample size of 400[a]

Students per cluster	*roh* = 0.2			*roh* = 0.4		
$\bar{n}$	Deff	n_c	m	Deff	n_c	m
1 (srs)	1.0	400	400	1.0	400	400
2	1.2	480	240	1.4	560	280
5	1.8	720	144	2.6	1,040	208
10	2.8	1,120	112	4.6	1,840	184
15	3.8	1,520	102	6.6	2,640	176
20	4.8	1,920	96	8.6	3,440	172
30	6.8	2,720	91	12.6	5,040	168
40	8.8	3,520	88	16.6	6,640	166
50	10.8	4,320	87	20.6	8,240	165

a Note: The values of m, the number of clusters selected, have been rounded upwards to the nearest integer value

students. For example, when *roh* = 0.4, the selection of 15 students per cluster from 176 clusters would have equivalent sampling accuracy to a design in which 50 students per cluster were selected from 165 clusters. The total sample size in these two cases differs by a factor of over three—from 2,640 to 8,240.

The selection of an appropriate cluster size for an educational research study usually requires the researcher to reconcile the demands of a set of often competing requirements. A number of authors (for example, Hansen et al. 1953, Kish 1965, Sudman 1976) have presented descriptions of the use of cost functions to calculate the optimal or most economical cluster size for certain fixed costs associated with various aspects of sampling and data collection. These approaches provide useful guidelines but they must be considered in combination with the need for high validity in the collection of data. For example, achievement tests which are to be administered in schools should preferably be given at one point of time in order to prevent the possibility of those students who have completed the test being able to discuss the answers with students who will be given the test at some later time. Educational researchers generally cope with this problem by limiting the cluster size to the number of students who can be tested under standardized conditions in one test administration. In most education systems this would represent cluster sizes of around 20 to 30 students when tests can be given by group administration. Much smaller cluster sizes may be necessary for tests which require individualized administration unless a large number of test administrators can be assigned at the same time to a particular school.

A further constraint on the choice of the cluster size may occur when analyses are planned for the between-student level of analysis and also at some higher level of data aggregation—for example, at the between-school level of analysis. In order to conduct analyses at the between-school level, data from students are usually aggregated to obtain data files consisting of school records based on student mean scores. If the number of students selected per school is too small then estimates of school characteristics may be subject to large within-school sampling errors.

12. PPS Two-stage Cluster Sample Designs

The preceding discussion of the simple two-stage cluster sample design was based on the assumption that the primary sampling units were of equal size. In educational research the most commonly used primary sampling units, schools and classes, are rarely equal in size. If the sizes of the primary sampling units vary a great deal then problems often arise in controlling the total sample size when the researcher aims to select a two-stage epsem sample.

For example, consider a two-stage sample design in which a schools are selected from a list of A schools, and then a fixed fraction of students, say $1/k$, is selected from each of the a schools. This design would provide an epsem sample of students because the probability of selecting a student is a/Ak which is constant for all students in the population. However, the actual size of the sample would depend directly upon the size of the schools which were selected into the sample.

One method of obtaining greater control over the sample size would be to stratify the schools according to size and then select samples of schools within each stratum. A more widely applied alternative is to employ probability proportional to size (PPS) sampling of the primary sampling units followed by simple random sampling of a fixed number of elements within these units. An exact execution of the PPS method provides complete control over the sample size and yet ensures epsem sampling of elements.

For example, consider a sample of m schools selected with PPS from a population of M schools followed by the selection of a simple random sample of $\bar{n}$ students from each of the m schools. Consider student i who attends school j which has n_j members from the total of N students in the defined target population.

The probability of selecting student i, p_{ij}, into this sample may be expressed as:

$$p_{ij} = m \times \frac{n_j}{N} \times \frac{\bar{n}}{n_j} = \frac{m\bar{n}}{N} \qquad (20)$$

Since m, $\bar{n}$, and N are constants then all students in the defined target population have the same chance of selection. That is, this PPS sample design would lead to epsem sampling, and at the same time fix the total sample size as $m\bar{n}$ students.

An estimate of the variance of the sample mean,

$\bar{x}_{pps}$, obtained from the PPS sample design described above may be obtained from the following formula (Yamane 1967 p. 255):

$$\text{var}(\bar{x}_{pps}) = \frac{1}{m(m-1)} \sum_{j} (\bar{x}_j - \bar{x}_{pps})^2 \qquad (21)$$

where $\bar{x}_j = 1/\hat{n} \Sigma_i x_{ij}$, is the mean score for students in the jth ultimate cluster, and $\bar{x}_{pps} = 1/m\hat{n} \Sigma_j \Sigma_i x_{ij} = 1/m \Sigma_j \bar{x}_j$, is the mean score for students in the total sample.

This formula emphaiszes two important points which emerged from the discussion of the simple two-stage cluster sample design. First, the variance of the sample mean may be reduced (for a given population of clusters) by increasing the number of primary selections. Second, the variance of the sample mean may be reduced (for a given number of primary selections) by allocating the elements to clusters in a fashion which reduces the variation between cluster means.

When accurate information concerning the size of each primary sampling unit is not available, then PPS sampling is often conducted by using "measures of size" rather than true sizes. That is, at the first stage of sampling, the clusters are selected with probability proportional to their measure of size. The difference between the actual size of a cluster and its measure of size is compensated for at the second state of sampling in order to achieve an epsem sample design Kish (1965 pp. 222–23) has presented formulas which demonstrate how to calculate the appropriate second-stage sampling fractions for these situations.

12.1 The Lottery Method of PPS Selection

An often-used technique for selecting a PPS sample of, say, schools from a sampling frame is to employ a lottery method of sampling. Each school is allocated a number of tickets which is equal to the number of students in the defined target population.

For example, consider the hypothetical population described in Table 2. Only the first seven and final three schools have been listed. However the total number of schools and students are assumed to be 26 and 4,000, respectively. Each school is allocated a number of tickets equal to the number of students in the defined target population in the school.

If five schools are to be selected then five winning tickets are required. The ratio of number of tickets to the number of winning tickets is 4,000/5 = 800. That is, each ticket should have a 1 in 800 chance of being drawn as a winning ticket.

The winning tickets are selected by using a random start-constant interval procedure. A random number in the interval 1 to 800 is selected from a table of random numbers and a list of five winning ticket numbers is created by adding increments of 800. For example with a random start of 520 the winning ticket numbers would be 520, 1320, 2120, 2920, and 3720. The schools which are selected into the sample have been marked in Table 2. School D corresponds to winning ticket number 520, and so on to school X which corresponds to winning ticket number 3,720. The chance of selecting a particular school is proportional to the number of tickets associated with that school. Consequently each of the five schools is selected with probability proportional to the number of students in the defined target population.

13. The Problem of Nonresponse

In most educational research studies there is usually some loss of data due, for example, to the nonparticipation of schools, or the nonresponse of sample members within selected schools. The resulting

Table 2
Hypothetical population of schools and students

School	Number of students in target population	Cumulative tally of students	Ticket numbers
A	50	50	1–50
B	200	250	51–250
C	50	300	251–300
D[a]	300	600	301–600
E	150	750	601–750
F	450	1,200	751–1,200
G[a]	250	1,450	1,201–1,450
.	.	.	.
.	.	.	.
.	.	.	.
X[a]	100	3,750	3,651–3,750
Y	50	3,800	3,751–3,800
Z	200	4,000	3,801–4,000

a Schools selected into final sample

missing data give rise to differences between the designed sample and the achieved sample.

One of the most frequently asked questions in educational research is: "How much missing data can be accepted before there is a danger of bias in the sample estimates?" The only realistic answer is that there is no general rule which defines a safe limit for nonresponse. The nonresponse bias may be large even if there are small amounts of missing data, and vice versa.

There are two broad categories of nonresponse: total nonresponse and item nonresponse. Total nonresponse refers to a complete loss of data for particular sample members, and is often dealt with by employing weights in order to adjust for differential loss of data among certain important subgroups of the sample (see *Data Analysis: Nonresponse*). Item nonresponse refers to the loss of a few items of information for particular sample members, and is usually dealt with by the assignment of values which replace the missing data.

It is important to remember that the level of bias in sample estimates which may occur through nonresponse generally cannot be overcome by increasing the sample size. The common approach of using random replacement of nonresponders usually provides additional sample members who resemble responders rather than nonresponders. The level of bias which actually occurs in these situations depends upon the variables which are being examined and their relationships with the nature of the nonresponding subgroup of the defined target population.

The problem of nonresponse in educational research appears to have received limited research attention. This is unfortunate because even doing nothing about nonresponse is an implicit adjustment scheme in itself which is based upon the assumption that loss of data is equivalent to random loss of data.

In studies where only a few items of information are missing for a small number of sample members, the procedure of value assignment is often used. This approach requires that the researcher, working from information which is available from the achieved sample, provides the values which replace the missing data. Lansing and Morgan (1971) recommend that the use of assignment procedures should be restricted either to situations where there are very few missing values associated with an important explanatory variable, or to situations where a limited amount of missing data appears for a variable that forms one component of a variable made up of many components.

The simplest form of assignment, sometimes referred to as matching, involves the direct copying of items of information from another sample member who is matched with the nonrespondent on the basis of their similarity across a set of key variables. An extended form of this approach occurs when the nonrespondent is assigned a value which is equal to the mean or the median for a group of respondents having similar characteristics.

The hot deck assignment procedure, developed by the United States Bureau of the Census, also employs a form of matching. Initially the data available from the sample members are partitioned into homogeneous subgroups based on a set of key variables. A cold deck of information derived from past survey data is then stored in the computer. If the first record to be processed has complete information then it replaces the cold deck; if information is missing from this record then the cold deck data is assigned. The process continues with the hot deck being continually updated to reflect the most recently processed sample cases. All sample records, after the first record with computer information, for which information is missing are consequently assigned the values recorded for the last record processed in the subgroup. It is thus possible for the same record to be used to assign values to many different records in which data are missing.

Assignment may also be carried out by using regression estimates of missing data. This approach capitalizes on the correlational associations between items for the responders. For example, a group of student home background variables may be used to prepare a regression equation with family income as the dependent variable. The sample members who do not respond to the family income question are then assigned a predicted value obtained from a regression equation estimate based on the home background information which they have provided.

See also: Sampling Errors; Interviews in Sample Surveys; Experimental Design: Survey Research Methods; Statistical Analysis in Educational Research; Research Methodology: Behavioral Sciences

Bibliography

Comber L C, Keeves J P 1973 *Science Education in Nineteen Countries: An Empirical Study*. Wiley, New York

Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society Series A* 222: 309–68

Haggard E A 1958 *Intraclass Correlation and the Analysis of Variance*. Dryden, New York

Hansen M H, Hurwitz W N, Madow W G 1953 *Sample Survey Methods and Theory*. Vols 1 and 2. Wiley, New York

Kish L 1957 Confidence intervals for clustered samples. *Am. Sociol. Rev.* 22: 154–65

Kish L 1965 *Survey Sampling*. Wiley, New York

Lansing J B, Morgan J N 1971 *Economic Survey Methods*. Institute for Social Research, Ann Arbor, Michigan

Peaker G F 1967 Sampling. In: Husen T (ed.) 1967 *International Study of Achievement in Mathematics: A Comparison of Twelve Countries*. Vol. 1. Wiley, New York; pp. 147–62

Peaker G F 1975 *An Empirical Study of Education in Twenty-one Countries: A Technical Report*. Wiley, New York
Ross K N 1978 Sample design for educational survey research. *Eval. Educ.* 2: 105–95
Ross K N 1983 *Social Area Indicators of Educational Need*. Australian Council for Educational Research, Hawthorn, Victoria
Sudman S 1976 *Applied Sampling*. Academic Press, New York
Yamane T 1967 *Elementary Sampling Theory*. Prentice-Hall, Englewood Cliffs, New Jersey

K. N. Ross

Sampling Errors

The difference between a particular sample estimate, and the population parameter obtained from a complete analysis of all members of the defined target population is called the sampling error for that sample. In most practical situations the value of the population parameter is unknown and therefore it is not possible to calculate the sampling error for a particular sample. Instead, through a knowledge of the behaviour of estimates derived from all possible samples, it is sometimes possible to estimate the average, or expected, sampling error even though the value of the population parameter is unknown.

The notion of an average, or expected, sampling error is usually summarized in terms of the mean square error. The mean square error, MSE, is the expected value of the squared difference between a sample value, for example the sample mean $\bar{x}$, and the population parameter, μ, taken over all possible samples. Denoting the expected value of the sampling distribution of sample means by $E(\bar{x})$, the mean square error may be written as:

$$\begin{aligned} \text{MSE}(\bar{x}) &= E(\bar{x} - \mu)^2 \\ &= E[\bar{x} - E(\bar{x})]^2 + [E(\bar{x}) - \mu]^2 \\ &= \text{Variance of } \bar{x} + (\text{Bias of } \bar{x})^2 \end{aligned} \quad (1)$$

In most well-designed samples the bias of a sample estimate is either zero or small, tending towards zero with increasing sample size. Therefore, the average sampling error is usually described in terms of the variance.

1. Estimation of Sampling Errors

1.1 Simple Random Samples

The educational researcher is usually dealing with a single random sample of data rather than all possible samples from a population. The variance of a sample estimate therefore cannot be calculated exactly. Instead, by using formulas derived by statisticians, estimates are made of the variance from the internal evidence of a single sample of data.

For a simple random sample of n elements drawn without replacement from a large population, the variance of the sample mean may be estimated from a single sample by using the following formula (Kish 1965 p. 41):

$$\text{var}(\bar{x}) = \frac{s^2}{n} \quad (2)$$

where $\text{var}(\bar{x})$ refers to the sample estimate of the variance of $\bar{x}$, and s^2 is the unbiased estimate of the variance of the element values. A factor called the finite population correction has been left out of the formula because the population is assumed to be large.

In many practical survey research situations the sampling distribution of the sample mean, and many other sample estimators, is approximately normally distributed. The approximation improves with sample size even though the population of element values is far from normal (Kish 1965). Consequently, by taking the square root of the estimated variance it is possible to obtain an estimate of the standard error of the sampling distribution of these sample estimators and thereby calculate confidence limits for the corresponding parameter.

In the case of the sample mean, the estimate of the standard error would be:

$$\text{SE}(\bar{x}) = \sqrt{\text{var}(\bar{x})} = \frac{s}{\sqrt{n}} \quad (3)$$

Although there is general agreement among statistical authors concerning the formula for estimating the standard error of the sample mean for a simple random sample of elements, there are sometimes differences of opinion about the appropriate formulas for calculating the variance of more complex statistics. These differences generally become insignificant for the typically large population and sample sizes which are associated with educational survey research. In Table 1 the formulas for calculating the standard error of several commonly used statistics have been listed. The formulas were selected from one source (Guilford and Fruchter 1978).

The formulas in Table 1 are based on a simple random sample of n elements which are measured on m variables. The symbol s refers to the standard deviation and the symbol $R_{i\,jkl}$ refers to the multiple correlation coefficient associated with a regression equation which uses variable i as the criterion and variables j, k, and l as predictors (see *Regression Analysis*; *Correlational Procedures*).

1.2 Complex Samples

Educational research is generally conducted by using data obtained from complex sample designs which employ techniques such as stratification, clustering, and varying probabilities of selection. Computational formulas are available to provide estimates of the

Table 1
Formulas for estimation of sampling error when data are gathered by using a simple random sample design

Sample statistic	Estimate of standard error
Mean	$\frac{s}{\sqrt{n}}$
Correlation coefficient	$\frac{1}{\sqrt{n}}$
Standardized regression coefficient	$\sqrt{\left[\frac{1 - R^2 1.234 \ldots m}{(1 - R^2 2.34 \ldots m)(n - m)}\right]}$
Multiple correlation coefficient	$\frac{1}{\sqrt{n - m}}$

standard errors of descriptive statistics such as sample means for a wide range of these sample designs. Unfortunately, the computational formulas required for estimating the standard errors for analytical statistics such as correlation coefficients, standardized regression coefficients, and multiple correlation coefficients are not readily available for sample designs which depart from the model of simple random sampling. These formulas are either enormously complicated or, ultimately, they prove to be resistant to mathematical analysis (Frankel 1971).

Due to the lack of suitable sampling error formulas for analytical statistics estimated from complex sample designs, researchers have tended to accept estimates based on formulas which assume that data have been gathered by using simple random sample assumptions. While overestimates of sampling errors may lead to errors of a conservative kind, underestimates have the potential to misrepresent the stability of sample statistics in a fashion which might lead to erroneous conclusions concerning the importance of research findings.

The research evidence which is available concerning the magnitude of sampling errors for statistics such as means, correlation coefficients, regression coefficients, and multiple correlation coefficients suggests that the use of formulas based on the assumption of simple random sampling often results in gross underestimation of sampling errors for many sample designs which are commonly used in educational research (Peaker 1975, Ross 1978). The degree of underestimation may be summarized by the "design effect" or "Deff" value. In Table 2 some values of $\sqrt{\text{Deff}}$ have been presented for a two-stage sample design employed in seven countries during a cross-national research study carried out by the International Association for the Evaluation of Educational Achievement (IEA) (Peaker 1975). For each country the number of schools selected at the first stage, m, and the number of students selected within the sample schools, $\bar{n}$, has been presented.

The value of $\sqrt{\text{Deff}}$ represents the factor by which sampling errors, obtained from formulas based on simple random sampling assumptions, must be multiplied in order to obtain estimates of the actual value of the sampling error for the complex sample design. For example, from the data presented in Table 2, the standard error of a correlation coefficient for Australia based on the complex two-stage sample of $n_c = m\bar{n}$ elements would be:

$$\text{SE}(r_c) = \sqrt{\text{Deff}}\ \text{SE}(r_{srs}) = \frac{1.7}{\sqrt{n_c}} \qquad (4)$$

Table 2
Mean values of $\sqrt{\text{Deff}}$ Deff obtained for seven countries participating in the IEA science project at the 14-year-old level

Country	Schools m	Cluster size $\bar{n}$	Value of $\sqrt{\text{Deff}}$		
			Means	Correln. coeff.	Regn. coeff.
Australia	225	24	2.4	1.7	1.3
Chile	103	13	2.6	1.6	1.6
Finland	77	30	2.2	1.7	1.3
Hungary	210	33	3.2	1.9	1.5
New Zealand	74	27	1.9	1.4	1.4
Scotland	70	28	2.4	1.5	1.2
Sweden	95	26	1.8	1.2	1.3
Mean $\sqrt{\text{Deff}}$	—	—	2.3	1.6	1.4