Quantitative research methods in educational planning

Series editor: Kenneth N.Ross

Module 4

Richard M. Wolf

Judging educational research based on experiments and surveys

UNESCO International Institute for Educational Planning

Quantitative research methods in educational planning

These modules were prepared by IIEP staff and consultants to be used in training workshops presented for the National Research Coordinators who are responsible for the educational policy research programme conducted by the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ).

The publication is available from the following two Internet Websites:
http://www.sacmeq.org and **http://www.unesco.org/iiep**.

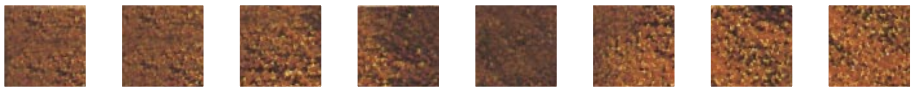
International Institute for Educational Planning/UNESCO
7-9 rue Eugène-Delacroix, 75116 Paris, France
Tel: (33 1) 45 03 77 00
Fax: (33 1) 40 72 83 66
e-mail: information@iiep.unesco.org
IIEP web site: http://www.unesco.org/iiep

September 2005 © UNESCO

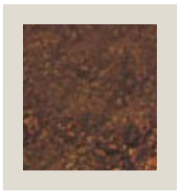
The designations employed and the presentation of material throughout the publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning its frontiers or boundaries.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means: electronic, magnetic tape, mechanical, photocopying, recording or otherwise, without permission in writing from UNESCO (International Institute for Educational Planning).

Graphic design: Sabine Lebeau
Typesetting: Sabine Lebeau
Printed in IIEP's printshop



Content



Introduction and purpose 1

Educational planners are continually being asked to participate in, and to provide information that can be used to guide administrative decisions. These decisions may range from developing a set of detailed procedures for some aspect of an educational enterprise, to a major reorganization of an entire education system. The quality of such decisions is the major determinant of successful administrative practice and, eventually, these decisions define the long-term nature of educational organizations.

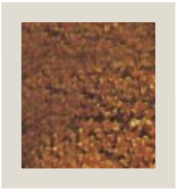
The quality of educational planning decisions depends, in turn, on the quality of the information upon which they are based. This information provides the best possible guidance for decision making when it is based on sound educational research combined with expertise derived from a comprehensive knowledge of the innermost 'workings' of the education system. The purpose of this module is to furnish educational planners with information on how to read and judge research reports in education so that they can use the information contained in them wisely.

While educational research reports provide an essential source of information for making decisions, there are other sources of information that educational planners need to consider. These are: costs, local customs and tradition, the views of various individuals who have a stake in an educational enterprise, governmental policies, laws, and the like. In making almost any decision, an educational administrator will need to consider research results alongside issues associated with some or all of these other sources of information.

Using the results of research in decision making is not an easy task. *First*, the educational planner needs to be able to distinguish good educational research from bad educational research. Currently, a great deal of research is being carried out all over the world. Some of it is of extremely high quality while some, unfortunately, is unquestionably poor. The educational planner needs, first of all, to be familiar with the key characteristics of research design and execution that will permit valid judgements to be made about research quality. Much of this module is directed towards developing an understanding of these characteristics. *Second*, just as most people have increasingly recognized the complexity of education, so have educational researchers. Accordingly, educational research has become more and more complex. This can easily be seen when one compares current research reports with those that were produced thirty or forty years ago. Contemporary educational research studies generally consider far more variables in a single study and employ more complex analytic procedures than their counterparts of a generation or two ago. This makes the task of reading and extracting information from research reports much more difficult for the reader. While one welcomes the more extensive understanding that has arisen from the increasing sophistication of much modern educational research, the problems that this sophistication creates for the reader must also be acknowledged. This module aims to discuss a number of issues in this area in an attempt to ameliorate such difficulties for the readers of research reports.

Having described the major focus of this module, it seems equally important to state what it is not intended to provide. It does *not* purport to be a substitute for specific training in the planning and conduct of research. Neither is it in any way a substitute for courses in statistics. That kind of training is best acquired in university courses and through gaining applied experience in those areas. This module is not intended to be technical in nature. Accordingly, discursive language has been used and specialized terminology

has been avoided wherever possible. When it has been necessary to introduce technical concepts, these have been explained as fully as possible and in as non-technical a way as ordinary language will permit. No special training should therefore be needed to read and understand the material presented throughout. A short list of references and additional readings is presented at the end of this module in association with references linked to research studies that have been cited in the text.



2 Research as a way of knowing

Research is a way of knowing the world and what happens in it. Its application is governed by various principles, canons, and rules. It can be contrasted with several other ways of knowing: authority, tradition, and personal experience. Each of these has its limitations and advantages.

Authority is a primary way of knowing used by many people. When an individual uses a reference work to obtain information, he or she is drawing on authority. Much of the education of the young is directed at teaching them how to make wise use of authority. The reason for this is very simple. It is too costly and inefficient for an individual to go out and obtain all the information that one needs to know through direct experience. It is also impossible. How is one to know anything about history except through reliance on authority in the form of published (written or oral) histories? Educational planners also routinely make use of authority when they draw on information in published works.

Tradition is a way of knowing in both developed and emerging societies. All societies are guided by accumulated knowledge of 'what is' and 'how things are to be done'. Such knowledge is often established through a process of trial and error. Methods of agriculture and manufacturing are but two examples of knowledge that is acquired over a long period of time and that serves as a guide to human endeavours. Sometimes tradition provides a useful guide to the conduct of human affairs and sometimes it does not. For example, methods of crop cultivation may, in fact, have adverse

long-term consequences such as soil depletion despite apparent short-term benefits. That is, the existence of traditional knowledge does not necessarily insure that it is also useful knowledge. However, traditions can often be strong and educational planners and policy-makers need to take them into account.

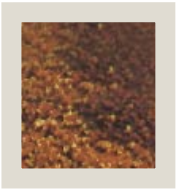
Personal experience has served as a guide to conduct throughout human history. Virtually everyone relies on his or her own experiences to make decisions about their own actions. While personal experience can be a useful guide to behaviour, this is not necessarily always the case. If one encounters truly novel situations, there may not have been any prior experiences that are available for guidance. Furthermore, personal experiences may be based on such limiting conditions that they become more of an impediment than an aid to conduct in new situations. Educational planners should therefore try to use care and caution when using personal experience to guide their planning decisions.

Research is a way of knowing based on systematic and reproducible procedures that aim to provide knowledge that people can depend on. It is, however, a somewhat expensive way of knowing since it demands that people who engage in research follow particular canons that usually require the use of special procedures, instruments, and methods of analysis. A major advantage of research as a way of knowing is that it is both deductive and inductive. It is self-correcting in that knowledge produced through research is public and subject to verification by others. Of the various ways of knowing, it probably produces the most dependable knowledge. This has been its major appeal in modern society and is probably most responsible for the high status that it is accorded – however it is not without problems in terms of application and interpretation. Throughout this module the reader will be continuously alerted to some of the problems that occur frequently with research.

In broad terms, research is generally concerned with the study of relationships among variables. A variable is a characteristic that can take on a number of values. Height, for example, is a variable that can take on a number of values, depending on the stature of the individual being measured. Achievement, attitudes, interests, and aspects of personality are all variables because they can take on a number of values, depending on the individual being measured. Variables do not refer only to characteristics of individuals. Variables can also refer to 'treatments' that might be applied to a group of individuals. For example, school subjects such as mathematics can be taught in very different ways to classes of students in schools. Thus, 'method of teaching mathematics' is a variable and each different way of teaching mathematics is a different value of this variable. 'Type of school organization' is also a variable since students can be grouped in many different ways for learning. Each way of organizing students would then represent a different value of 'type of school organization'. Many educational research studies are concerned with studying the relationship between a variable that describes a particular instructional intervention or method of organization and a student outcome variable, such as achievement, attitudes and behaviours developed, length of job search, etc.

At this stage, it would also seem important to indicate what research is not. Research does not provide fixed and immutable knowledge. That is, knowledge gained through research is relative and probabilistic in nature, not absolute and certain. For example, a researcher may have conducted a study and found a particular relationship between two quantities, say pressure and volume of certain gases – but this relationship is not necessarily static. It probably varies with temperature and, even with a known temperature, the relationship may not be exact because there is invariably some error in any research finding. The key message here is that knowledge produced through research can vary depending on the conditions under which it was obtained. Researchers,

accordingly, often attach probabilities to their findings. This point will be emphasized throughout this module. Suffice it to say for now that research knowledge is not absolute, but relative and with a particular likelihood (or probability) of occurrence.



3 Types of educational research

There are many types of educational research studies and there are also a number of ways in which they may be classified. Studies may be classified according to 'topic' whereby the particular phenomena being investigated are used to group the studies. Some examples of topics are: teaching methods, school administration, classroom environment, school finance, etc. Studies may also be classified according to whether they are 'exploratory' or 'confirmatory'. An exploratory study is undertaken in situations where there is a lack of theoretical understanding about the phenomena being investigated so that the main variables of interest, their relationships, and their (potential) causal linkages are the subject of conjecture. In contrast, a confirmatory study is employed when the researcher has generated a theoretical model (based on theory, previous research findings, or detailed observation) that needs to be 'tested' through the gathering and analysis of field data.

A more widely applied way of classifying educational research studies is to define the various types of research according to the 'kinds of information' that they provide. Accordingly, studies may be classified as: (1) historical, (2) case, (3) longitudinal, (4) survey, and (5) experimental. Within each major type of study there are other types of studies. For example, case studies are often 'ethnographic' studies that focus on detailed investigations of an individual or group's socio-cultural activities and patterns. Historical studies deal with past events and depend heavily on the use of source documents. Case studies seek to study an individual or particular group of individuals and are therefore not always intended to lead to inferences that are generalizable to wider populations. Longitudinal studies are concerned with the

study of individuals over time in order to describe the process of development and the stability and change in various characteristics. Survey studies furnish a picture of a group of individuals or an organization at a particular point in time. They often contain a number of items of information for describing the individuals or organization. Finally, experimental studies assess the effects of particular kinds of interventions on various outcomes of individuals.

Each type of research study has its own particular canons, procedures, techniques, and methods of analysis. This module has been restricted to judging research reports that are produced from survey studies and experiments. These comprise the major portion of educational research and are of the greatest relevance to educational planners. This focus is not intended to demean *historical*, *case*, and *longitudinal* studies. These three types of studies are very important and may well enter into the decision-making process.

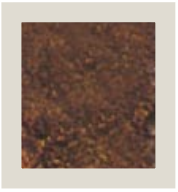
4 Association and causation

One of the most important distinctions that readers of research reports must be aware of in order to judge them properly is the distinction between association and causation. As described above, research involves the study of the relationships among variables. The relationship between two variables may be one of association, or of causation, or of both of these. The difference between the concepts of association and causation is critical to the understanding of research. An association between two variables states that there is a relationship. It does not necessarily mean that one variable causes another (or vice versa). Causation, on the other hand, means that one variable is the cause of another.

In many studies, investigators are simply able to establish an association between two variables, say, method of instruction and student achievement – which does not necessarily mean that one variable (method of instruction) has caused the other (student achievement). On the other hand, a study that establishes a causal relationship between two variables is stating that one variable is responsible for changes in another variable. Properly conducted experimental studies come closest to establishing the existence of causal relationships. Survey studies can only establish associations and 'suggest' causal linkages, but they cannot establish causal relationships.

Sometimes investigators who conduct survey studies attempt to claim that they have established causal relationships. They have not. No matter how elegant the methods of analysis that are used, survey or descriptive studies can not establish causal relationships. The most that can be claimed for survey or descriptive studies, no

matter how carefully planned and carried out, is a presumption of causation. The presumption may be quite strong, however, if there is considerable previous evidence and a strong theory favouring a certain conclusion. For example, the evidence on the relationship between cigarette smoking and lung cancer in human beings is based on associational studies (it would be highly unethical to conduct experimental studies in this area with human beings). However, the weight of evidence from many studies and a strong physiological theory makes the conclusions from such studies strongly presumptive of a causal relationship even though these studies do not definitively establish a causal relationship.



5 The main characteristics of experimental and survey studies

The main characteristics of experimental studies are: (1) active manipulation of treatment variables by the researcher, and (2) the use of random assignment of units (usually students) to each type of treatment. These characteristics constitute the essential controls exercised by a researcher to establish a causal relationship. For example, consider a situation where a researcher is interested in studying the effect of two methods of teaching multiplication of decimals on student achievement as measured by a test of multiplication of decimals. In a true experiment, the researcher selects the method of teaching to be studied, instructs two groups of teachers, (each in one of the selected methods), assigns students in a random fashion to one of the two types of classes, follows each class to see that it is following the prescribed method of instruction, and tests each student at the end of the period of instruction on a common test of multiplication of decimals. The resulting data are then analyzed and if the difference in the average level of performance between students in the two methods of instruction differs sufficiently, one comes closer to obtaining a causal relationship than in a situation where pre-existing conditions are merely compared. Such a study is experimental in nature because the researcher was able to exercise full control over the selection of methods to be studied, the random assignment of teachers to each method of instruction, and, finally, the random assignment of students to each method of instruction. Any study that does not

exercise this level of control cannot be termed an experimental study and any causal conclusions from it must be regarded as presumptive.

Survey studies also typically report relations among variables. These relationships are associational and not causal. For example, Coleman, Hoffer and Kilgore (1987) report the results of a large-scale study comparing the academic performance of students in public (government) and private (non-government) schools. Coleman, Hoffer and Kilgore found that students in private schools outperformed students in public schools in various tests of school achievement. Clearly, there is a relationship between type of school (public versus private) and school achievement. Is the relationship a causal one? The answer to the causal question is not known since Coleman, Hoffer and Kilgore were not able to assign students at random to the two types of schools. In fact, an examination of the backgrounds of the students shows that they were quite different at their time of entrance into their type of school. Students who attended private schools came from more affluent families, had higher levels of material resources in the home, were higher in achievement when they entered their private school, and held higher levels of expectation about what they would achieve in secondary school than students who attended public schools. The fact that the two groups of students differed so greatly at the start of the study and that it was not possible to equalize the groups attending the two types of schools through randomization makes it impossible to establish a causal relationship between type of school attended and academic performance. Causal inferences in this case would be, at best, presumptive.

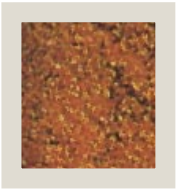
Survey studies are characterized by the study of relationships among variables in already existing units. No attempt is made to randomly assign individuals to groups and groups to treatments. This restricts such investigations to studies of associations among

variables. However, this does not mean that the possibility of causal relationships cannot be explored. They often are. In recent years, a number of highly technical statistical procedures have been developed that are used to explore possible causal relationships among variables. The general term for such procedures is causal modelling. Briefly, such procedures allow one to set up a network of hypothesized causal relationships among variables and to test the tenability of these relationships.

These models and the associated statistical procedures that are used are often quite complex. They are based on a rather simple notion, however. Whereas the existence of an association among two variables does not mean that the two variables are causally related, the researcher can examine the information collected in a study to see if an association exists between the two variables in question. If there is no association, then the researcher's theory is disconfirmed. However, if the two variables are associated, then the possible causal relationship between the variables remains tenable. This is not to say that a causal relationship has been established. It is only that the existence of a causal relationship remains a possibility. A great deal of rather complex analytic work goes on in the area of causal modelling and readers of research reports often have difficulty in following it. The basic goal of causal modelling should be clearly kept in mind though. Causal modelling, at best, tests the tenability of causal relationships; it does not establish them.

There is a third class of studies that lie somewhere between experimental and survey studies. These studies are called quasi-experimental. These are studies in which the researcher does not exercise full control over the selection and scheduling of treatments and the assignment (random) of students to groups for purposes of study. The researcher exercises only partial control over the study. For example, the researcher may not have the power to assign students to groups, but may be able to schedule which groups receive particular treatments. Strictly speaking, such studies are

not experimental studies because of the lack of full control over the research situation. However, it is clearly a better state of affairs than exists in a survey study. Causal conclusions drawn from the results of such studies are still presumptive, but the presumptions are often fairly strong. It is up to the reader of the reports of quasi-experimental studies to decide whether the causal speculations of the researcher are warranted. Unfortunately, there is no simple formula for judging whether they are causal in nature.



6 Qualitative and quantitative studies

Experimental, quasi-experimental and survey studies are regarded as quantitative studies because of the collection of information that is quantifiable and subjected to statistical analysis. Many research studies are quantitative in nature. They are designed to expose relationships among variables. In contrast, there are studies that are basically qualitative in nature. The information collected in such studies is usually not quantified and not subjected to statistical analysis. Usually, case studies and historical studies are qualitative in nature although, on certain occasions, they may employ quantitative procedures. Some years ago, for example, a quantitative study was undertaken to resolve the question of authorship of some important historical papers in the USA. The researchers in that study had samples of the writings of two individuals who were quite prominent in their time and, based on the characteristics of the writings of each author, the researchers were able to use statistical procedures to resolve the issue of authorship. Such examples of the use of quantitative methods in historical studies are rare.

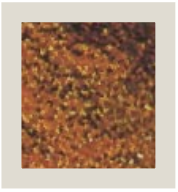
It is often difficult to distinguish between qualitative and quantitative studies at the level of research technique or data collection procedures. Both qualitative and quantitative studies may use the same techniques or procedures. Thus, for example, interviews and direct observation can and are used in both kinds of studies, often with excellent results. The information that is obtained, however, is treated quite differently in the two kinds of studies. In quantitative studies, information obtained through the use of interviews or direct observation is typically subjected to

statistical analysis while in qualitative studies, such information is not subjected to these procedures.

Currently, there is considerable debate regarding quantitative and qualitative studies. The issue is being fought out at a philosophical as well as a methodological level. The details of that debate are not of concern here. What is important is that the reader recognize the distinction between the two types of studies. In addition, it is important to have a sense of what each type of study can contribute to education.

Quantitative studies, when properly conducted, can establish relationships among variables. However, they often tell us little about how causal relationships work. They may tell us, for example, that the use of a certain procedure, say peer tutoring, leads to higher levels of achievement among both tutors and tutees. However, the precise mechanism by which peer tutoring results in higher achievement is not ascertained in such studies. It often requires finely detailed qualitative studies such as ethnographic studies of individuals to determine the way in which peer tutoring leads to higher achievement. While some researchers and even philosophers of science see quantitative and qualitative studies as being in opposition to one another, their functions can be complementary as long as one does not expect some kind of philosophical purity in research.

This module deals exclusively with quantitative studies that are used to establish relationships among variables. Educational planners are usually required to deal with quantitative studies since they provide most of the information on which planning decisions are based. Qualitative studies do play a useful role in attempting to understand the mechanisms by which relations among variables are established. They are often undertaken after quantitative studies have established the existence of an important relationship, or they may be undertaken for purely exploratory purposes in order to gain an understanding of a particular process.



7 Experimental studies and some factors that often threaten their validity

The basic structure of experimental studies

As noted in the previous discussion, properly designed and conducted experimental studies provide a powerful means of establishing causal relationships among variables of interest. The reason why this is so are inherent in the basic structure of an experimental study. To illustrate, we may represent the basic design of a classic experimental study as follows:

Subjects

R

$$\frac{O_1 T_1 O_3}{O_2 T_2 O_4}$$

In the diagram, 'R' denotes that a random assignment procedure has been used to assign students to groups and groups to treatment conditions. 'O' denotes an observation of performance. Such observation may consist of tests that are administered to students. Note that observations 'O' are obtained both before the introduction of the treatment conditions and after completion of the treatment period. While it is considered desirable to have both before (pre-test) and after (post-test) observations, the former are not considered crucial in an experimental study and are sometimes omitted without jeopardizing the integrity of the design. The symbol 'T_1' denotes the major treatment condition that is being studied. It may be a particular instructional treatment or method of teaching, a type of organizational structure, or some other intervention that is being studied. 'T_2' denotes the alternative treatment to which it is being compared. In some studies, T_2 is often a condition of no treatment at all. In this way, researchers can assess the effect of a particular intervention in relation to no intervention at all. The 'no intervention' alternative is only occasionally studied in the field of education since there are usually legal regulations that prevent very uneven treatment of students – unless, of course, all treatments are considered equally beneficial.

To assess the effect of 'T_1', a comparison is usually made between the average levels of the post-tests for the two groups ('O_3' and 'O_4' in the above diagram). If 'O_1' and 'O_2' are present, a somewhat more complex analytic procedure, analysis of covariance, is usually employed to adjust for any existing random or chance differences between groups before the treatments are introduced.

As mentioned above, experimental studies can be used to establish causal relationships as they employ random assignment to ensure comparability of the groups being studied. The random assignment of students to groups and the random assignment of groups to treatments serves as a way of equalizing groups, except for minor random or chance differences, before the initiation of treatment.

If there are differences in performance between the groups at the completion of treatment, the most likely (and possibly the only) reason for the difference is the differential effectiveness of the treatments. The use of a group other than the one receiving the treatment of interest, technically called a 'control group', provides the necessary comparability to estimate the effectiveness of the major treatment.

In theory, experimental studies are the preferred way of estimating the effectiveness of educational treatments and organizational structures. In practice, experimental studies are subject to various limitations. These limitations stem from the way in which experimental studies are actually conducted in various settings. The limitations of experimental studies fall under the two general headings of internal invalidity and external invalidity. Internal invalidity refers to the influence of extraneous factors that can mar a well designed study. External invalidity refers to the lack of ability to generalize the findings of a particular study to other settings. Each is important and will be considered in turn.

Validity of experimental studies

The most difficult task in conducting an experimental research study in the field of education is to hold all variables in the educational situation constant except for the treatment variable. The degree to which these 'extraneous variables' may be controlled by the researcher is often referred to as the 'internal validity' of the experiment.

Campbell and Stanley (1963) wrote a classic paper that provided a comprehensive list of the factors that threaten the validity of experiments. These factors, and some examples of how they might influence and/or distort research findings have been listed below.

1. History

In educational research experiments, events other than the experimental treatment can occur during the time between the pre-test and the post-test. For example, an in-service programme to improve the knowledge and proficiency of teachers of reading may be undertaken by a particular school. At the same time, some of the teachers may be enrolled in university courses leading to an advanced degree. As part of the programmes, these teachers may be taking a course in the teaching of reading. It is certainly possible to assess the teachers' knowledge and proficiency in the teaching of reading at the conclusion of the in-service programme, but it would be virtually impossible to determine how much of their performance is due to the in-service programme and how much to their graduate course. This inability to determine the source of an effect, namely, the enhanced knowledge and proficiency in the teaching of reading renders the results of the study uninterpretable. History, as a source of internal invalidity, opens the results of a study to alternative interpretations. Readers of research reports should routinely ask themselves whether a demonstrated effect is due to the intervention under study or to something else.

2. Maturation

While an experiment is being undertaken, normal biological psychological growth and development processes are almost certain to continue to occur. These processes may produce changes in the experimental subjects that are mistakenly attributed to differences in treatment. Maturation effects are often noticed in long-term experiments in which students learn a great deal through the natural processes of exposure to stimuli that are a normal part of their socio-cultural environment. For example, students at a particular grade level in primary school who have mastered some of the rudiments of reading will undoubtedly improve in their reading

ability simply as a result of being confronted with printed material in a variety of situations – magazines, newspapers, and the like. The problem for the researcher is to determine to what extent reading improvement for such students is due to the effects of instruction and to what extent it is due to growing up in a culture where one is constantly exposed to reading material.

3. Testing

In most educational experiments a pre-test is administered before the experimental treatment which is then followed by a post-test. The very administration of the pre-test can improve performance on the post-test in a manner that is independent of any treatment effect. This occurs when pre-testing enhances later performance through providing practice in the skills required for the post-test, by improving the 'test-wiseness' (or test-taking skills) of the students, or by sensitizing the students to the purposes of the experiment. The effect of retesting can sometimes be reduced by making sure that students are given a different set of questions to answer when a test is readministered. There are two ways to eliminate a testing effect. The first would be to test only once, at the completion of the treatment period. This can be troublesome since it would deprive the researcher of information about the proficiency of students at the beginning of the programme ('O_1' and 'O_2' in the above diagram). The second way to eliminate a testing effect is to randomly divide each group of students in half and administer the test to one half of the group before the period of treatment and to the other half of the group after instruction.

4. Instrumentation

A difference in the pre-test and post-test scores for an experiment may sometimes occur because of a change in the nature or quality of the measurement instrument during the course of the

experiment. For example, the scores of essay tests may change from pre-test to post-test because different standards are used by two sets of scorers on different occasions. If the scorers of the pre-test essays are particularly stringent in their grading while the scorers of the post-tests are lenient, then gains in the essay scores may all be due to the differences in standards used by the scorers rather than the exposure of students to effective teaching. The same situation may hold for more objective measures of student performance. For example, the researcher might simply ask easier questions on a post-test than on a pre-test. Instrumentation problems also often arise when the amount of proficiency required to go from, say, a score of six to twelve is different from the amount required to go from a score of twelve to eighteen. Test scores are typically treated as if the difference between score points is uniform throughout the test, and therefore the research worker must be sensitive to the nature of the instruments that are used and the units of measurement that express performance.

5. Statistical regression

When students are selected for a treatment on the basis of extreme scores, later testing invariably shows that these students, on average, perform somewhat closer to the average for all students. This phenomenon was identified by the psychologist Lewis Terman in his studies of gifted children over half a century ago. Terman sought to identify a group of gifted children. His major criterion for classifying them as gifted was the score obtained on the Stanford-Binet Intelligence Test. As part of his initial follow-up of these children, Terman had them retested and found, to his surprise, that the average intelligence test score had 'regressed' rather dramatically (eight points) toward the average. More recently, remedial educational programmes have been developed in a number of countries to help disadvantaged students. A common practice in such programmes is to select individuals who score

extremely low on some test. On later testing, these students show a considerably higher average level of performance. While some increment in performance may have occurred, much of the apparent improvement is simply due to statistical regression: in this case the direction is upward instead of downward, as in the case of Terman's studies of gifted children. In both cases the phenomenon is the same: individuals initially selected on the basis of extreme scores will, on retesting, show less extreme scores.

6. Selection

In a study that seeks to compare the effects of treatments on different groups of students, the group receiving one treatment might be more able, older, more receptive, etc. than a group receiving another, or no treatment. In this case, a difference between groups on post-test scores may be due to prior differences between the groups and not necessarily the differences between treatments. For example, if students volunteer to participate in an experimental learning programme, they can differ considerably from students who decide to continue in a conventional programme. In this case, the act of volunteering may indicate that the volunteer students are looking for new challenges and may approach their lessons with greater zeal. If differences favouring the experimental programme are found, one faces the task of trying to decide how much such results reflect the effects of the programme and how much the special characteristics of the volunteer students.

7. Drop-out

In experiments that run for a long period of time there may be differential drop-out rates between the groups of students receiving the experimental treatments. For example, random allocation of students to several educational programmes may have ensured

comparable groups for the pre-test – but if one group incurs a loss of low-performing students during the course of the experiment, that group's average performance level will increase, regardless of the effectiveness of the programme to which it was exposed.

8. Interaction

It is possible for some of the above factors to occur in combination. For example, a source of invalidity could be selection- maturation interaction whereby, due to a lack of effective randomization, major age differences occur between the treatment groups – which, in turn, permits the possibility of differential rates of maturity or development between the groups. These latter differences may result in differences in post-test scores independently of treatment effects. Another example illustrates how the joint operation of several factors can lead one erroneously to conclude that a programme has been effective. A study was conducted on the effects of the 'Sesame Street' educational television programme in the USA. In the first year of the study of that programme's effectiveness, four different groups of children were examined to judge the instructional effects of the programme. The groups were established on the basis of the amount of time spent viewing 'Sesame Street'. This ranged from rarely or never watching the programme to viewing it more than five times a week. Scores on the 'Sesame Street' pre-tests were found to be highly related to the amount of time spent viewing the programme. That is, the higher the initial score, the more time was spent watching the programme. Scores on the post-test, and hence the gains, were in the same way highly related to the time spent viewing the programme. The combination of pre-test performance and self-selection into viewing category made it impossible to assess the effectiveness of the first year of 'Sesame Street' from the data collected.

9. Diffusion of experimental treatment

In some experiments 'the treatment' is perceived as highly desirable by members of the control group. This may lead the control subjects to seek access to the treatment – either by communicating with the treatment subjects or by some means that were not anticipated during the design of the experiment. This problem is a source of major concern in the evaluation of new curriculum programmes where the teachers and students in the treatment group are given access to an attractive curriculum based on exciting and innovative teaching materials. During the course of the experiment, curiosity might lead many of the teachers of the control group to discuss the new programme with the treatment group teachers – even if they have been instructed not to do so. Subsequently, the students in one group may learn the information intended for those in the other groups. Thus, the study may become invalid because there are, in fact, no real differences between the treatment and control curricula. While it may not be possible to have complete control over such contact in some instances, the monitoring of programme implementation in both the group receiving the experimental treatment and the group not receiving that treatment should reveal how likely a threat diffusion or imitation of treatments is to the validity of the study.

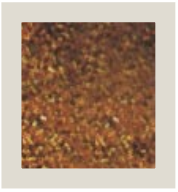
10. Resentful demoralization of students receiving less desirable treatments

The members of the group not receiving the treatment that is being studied may perceive that they are in an inferior status group and either 'lose heart' or become angry and 'act up'. This could lead to an after treatment difference between groups that may not be a consequence of treatment effectiveness but rather of resentful demoralization by the students receiving the alternative treatment. Some monitoring of the group receiving the alternate

treatment should reveal how plausible this threat is to validity. The threat can be controlled somewhat by planning that separates the group receiving the treatment of interest from the group receiving the alternate treatment in either time or space. Alternatively, arrangements can be made to enliven or 'spice up' the alternative treatment so that it appears as desirable as possible to participating students.

11. Generalizability

The ten threats to validity described above focus upon the need to ensure that the effect of the experimental treatment is not confounded by extraneous variables. All of these threats can be managed through exerting firm controls over the design and execution of an experiment. However, the implementation of these controls may lessen the 'realism' of the environment in which an experiment is conducted and consequently may affect the generalizability of the research findings to other populations, other treatments, other measurement instruments, and other social/economic/cultural/environmental settings. Any research study is conducted in a particular time and place and with particular students, treatment variables, and measurement variables. To what extent can the results of any one study be generalized to other cases? Strictly speaking, this question is unanswerable. However, at the very least, the reader of research reports must decide how similar the group under study is to the groups he/she is responsible for and how similar the conditions in his/her setting are to the one in which the study was conducted. In order to make these decisions it will be extremely helpful if the writer of a research report has carefully described the setting in which the study occurred, the students who were studied, the particular treatment that was investigated, and the types of measurements that were taken. In short, the generalizability of a study's findings for a different setting is a matter that requires careful consideration, not automatic acceptance.



8 Survey studies and some factors that often threaten their validity

The basic structure of survey studies

There is no single definition that can be used to provide a comprehensive description of the structure of survey studies. There are many types of survey studies but they all have one key feature in common: they all obtain measures from a scientific sample of subjects selected from a well-defined target population. In a cross-sectional survey these measurements are used to prepare summary statistics and then make inferences from these about the nature of the target population. In a longitudinal survey the focus is on the use of a series of time-related measurements of the same sample of individuals. Both cross-sectional and longitudinal surveys may be used for descriptive purposes, or for examining relationships between important variables, or for exploring conceptual models derived from proposed networks of variables. The following discussion of factors that threaten validity has been limited to the use of cross-sectional studies for descriptive purposes.

A survey study may be regarded as a snapshot of a situation at a particular time. Descriptive studies, because of cost, are rarely conducted for an entire population and therefore a sub-set of a population, called a sample, is chosen for closer study. The selection of a sample for a survey is a critical part of such a study since the

sample has to be chosen in such a way that it is representative of the larger population of which it is a part.

As noted earlier, survey studies can be used to establish associations between variables, and do not permit the drawing of causal relationships. Despite their limitations, survey studies play an important role in education. They can result in useful descriptions of the current state of affairs in a situation and have often been used as the basis for introducing changes, especially when the state of affairs that is described is considered unacceptable. Thus, for example, a study of the school achievement of students in a particular locality, or even a nation, may reveal levels of achievement that are deemed unacceptable to educational authorities.

Survey studies have also been used for comparative purposes. The studies conducted by the *International Association for the Evaluation of Educational Achievement* (IEA), have been used to compare the performance of students at various age and grade levels in different nations. The identification of achievement differences has often led to a closer examination of various nations' educational systems with a view to improving them. Within nations, comparisons have often been made between various types of schools, for example, single-sexed versus coeducational, in order to assess the relationship between the sex composition of schools and achievement.

The validity of survey studies

Survey studies require particular attention to be given to the scope of data collection and the design and management of data collection procedures (especially sampling, instrumentation, field work, data entry, and data preparation). If a survey study has problems with any of these areas, then the validity of the study's findings may be threatened.

In the following discussion some of the factors that often threaten the validity of survey studies have been presented. Since these factors are mainly concerned with 'generalizability' they also have the capacity to threaten the validity of experimental studies.

1. The scope of the data collection

The first step in the conduct of a survey study is specifying the entity that is to be described. This, of course, will depend on the purpose of the researcher. In some cases, one may wish to describe some features of a single locality such as the size of classes or the qualifications of teachers at a particular grade level. In other cases, one may want to assess the attitudes and achievement of students at a particular level in some school subject. Usually, researchers who carry out survey studies gather information on a number of different variables. The reason for doing this is that once one undertakes the collection of information from people, it is usually a matter of little additional time to collect information on a large number of variables rather than on just a few variables. However, there is a danger here. Sometimes researchers who are not completely clear as to their research objectives collect information on a large number of variables without knowing why they are doing so. This is often referred to as 'shotgun' research. The hope of such researchers, usually unfounded, is that if they collect information on as many variables as possible, they are apt to include in their list of variables some that may turn out to be important. Readers of research reports should generally be suspicious of studies that collect information on large numbers of variables – often hundreds.

Data collection efforts in education systems in some countries often give insufficient attention to whether it is really necessary to study the whole population of students, teachers and schools. The coverage of a whole population, because of the breadth/depth tradeoff, usually results in little information about many units. In this situation, important variables may be omitted and/or measured

with insufficient attention to reducing measurement error. For most purposes, sample surveys, when designed and executed appropriately, can provide as much information as complete censuses at considerably less cost. For example, sample surveys are often adequate for providing accurate estimates of participation and repetition rates, and are virtually mandatory for estimating national achievement levels, particularly for students in grades not regularly examined for selection purposes.

2. The sample design

The first step in the preparation of a sample design for a survey study is to develop descriptions for the desired target population (the population for which results are ideally required), the defined target population (the population which is actually studied and whose elements have a known and non-zero chance of being selected into the sample), and the excluded population (the population comprised of the elements excluded from the desired target population in order to form the defined target population). A population is defined by specifying the characteristics that all elements of the population have in common. For example, one may define a population as all students between age ten years zero months and ten years eleven months attending full-time government schools in Budapest, Hungary. Similarly, one may define a population as consisting of all students enrolled in a first year course in French in Swedish schools. In each of the above instances, the specification of the common characteristics of the members of the population defines the population.

Since it is usually not possible to study an entire population, because of cost and logistical considerations, a sub-set of the population is selected for actual study. This sub-set is called a sample. One of the challenges that researchers face is to select a sample from a population for study in such a way that it will provide precise estimates of the defined target population

characteristics. Unfortunately, in many survey studies the sample estimates provide very poor estimates of defined target population characteristics because of the following five problems.

- The defined target population and the excluded population are never clearly defined. This may arise because the researcher either does not bother to specify the size and nature of these populations or, due mainly to lack of precise information, is unable to provide precise definitions. Unfortunately, this problem often goes hand-in-hand with the researcher making generalizations about a desired target population that, upon careful scrutiny, is quite different from the defined target population.

- The participants in the study are nominated rather than sampled. This approach is often justified in terms of cost or accessibility considerations, however both of these 'constraints' can usually be addressed by adjusting the defined target population definition and then applying appropriate stratification procedures. These nonprobability samples, sometimes referred to as 'nominated samples', are generally described in scientifically meaningless terms such as 'quota', 'representative', 'purposive', 'expert choice', or 'judgmental' samples. Kish (1965) characterized data collections based on this approach as 'investigations' and pointed out that they should not be confused with appropriately designed experiments or surveys. The main problems associated with the use of nominated samples are that it is not possible to estimate the sampling errors or to have any idea of the magnitude of the bias associated with the selection procedures (Brickell, 1974). Consequently, nominated samples should be used only for the trial-testing of instrumentation or new curriculum materials because in these activities it is sometimes desirable to employ a 'distorted' sample that has, for example, a disproportionately large number of students at the extremes of a spectrum of ability, ethnicity, socio-economic status, etc.

- The sampling frame used to list the defined target population is faulty because it is out of date and/or is incomplete and/or has duplicate entries. The construction and maintenance of a comprehensive sampling frame for schools, teachers, and students may be neglected because it is considered to be too expensive or because the systematic collection of official statistics in a country is error-prone. This is sometimes the situation in countries where population growth rates are high and where large and uncontrolled movements of population from rural to urban settings are commonplace. However, there are also a number of countries that are unable to provide accurate information in this area because the management and financing of schooling is undertaken by local communities, or because there is an independently managed non-government school sector. The researcher faced with these difficulties often proceeds to use a faulty sampling frame based on poor quality official statistics in the mistaken belief that there are no other alternatives. In fact, there are well-established solutions to these problems that employ 'area sampling' (Ross, 1986) and, provided that a trained team of 'enumerators' is available to list schools within selected areas, it is possible to prepare a high quality sample design without having access to an accurate sampling frame based on a listing of individual schools.

- Confusion surrounding the terms 'total sample size' and 'effective sample size' results in the total sample size for a complex cluster sample being set at the wrong level either by the use of simple random sampling assumptions or, quite frequently, by guesswork. In school systems that are highly 'streamed', either explicitly on the basis of test scores or implicitly through the residential segregation of socio-economic groups, the use of complex cluster sampling can have dramatic effects on the total sample size that is required to reach a specific level of sampling precision. This occurs because the streaming causes larger differences in mean scores between

classes than would be the case if students were assigned at random to classes. (The magnitude of these differences can be measured by using the coefficient of intraclass correlation). Researchers with a limited knowledge of this situation often employ simple random sampling assumptions for the estimation of the required total sample size. In order to illustrate the dangers associated with a lack of experience in these matters, consider the following two examples based on schools in a country where the intraclass correlation for achievement scores at the Grade 6 level is around 0.6 for intact classes. A sample of 40 classes with 25 students selected per class would provide a total sample size of 1,000 students – however, this sample would only provide similar sampling errors as a simple random sample of 65 students when estimating the average population achievement level. Further, a sample of 50 classes with 4 students selected per class would provide a total sample size of 'only 200 students' but would nevertheless provide estimates that are more precise than the above sample of 1,000 students.

- The wrong formulae are used for the calculation of sampling errors and/or for the application of tests of significance. This usually occurs when the researcher employs a complex cluster sample (for example, by selecting intact classes within schools) and then uses the sampling error formulae appropriate for simple random sampling to calculate the sampling errors (Ross, 1986). The most extreme form of this mistake occurs when differences in means and/or percentages are described as being 'important' or 'significant' without providing any sampling error estimates at all – not even the incorrect ones. These kinds of mistakes are quite common – especially where 'treatment versus control' comparisons are being made in order to compare, for example, current practices with new curriculum content or new teaching materials (Ross, 1987).

3. Instrumentation

Once one has decided what population to study, the next step is deciding what items of information should be collected via the data collection instruments (tests, questionnaires, etc.) that have been designed for the study. One may choose to study a very limited set of variables or a fairly large set of variables. However, the research questions governing a study should determine the information that is to be collected. Some variables may be obtained at very low cost. For example, asking a member of a sample to report his or her sex will require only a few seconds of the respondent's time. On the other hand, obtaining an estimate of a student's proficiency in mathematics or science may require one or more hours of testing time. Time considerations will be a major determinant of how much information is collected from each respondent. Usually, a researcher or a group of researchers will need to make compromises between all the information that they wish to collect and the amount of time available for the collection of information. The data collection instruments should be clear in terms of the information they seek, retain data disaggregated at an appropriate level, and permit the matching of data within hierarchically designed samples or across time. Furthermore, they must be designed to permit subsequent statistical analysis of data for reliability and (if possible) validity. The basic requirements are that the questions posed do not present problems of interpretation to the respondent, and that, when forced choice options are provided, the choices are mutually exclusive and are likely to discriminate among respondents.

The presentation of test scores and sub-test scores should be accompanied by appropriate reliability and validity information. At the most minimal level for norm-referenced tests, a traditional item analysis should be undertaken in order to check that the items are 'behaving' in an acceptable manner with respect to discrimination, difficulty level, and distractor performance. An attempt should be made to establish the validity of tests where this has not been

carried out previously. The quality of the instruments prepared for a survey study is heavily dependent upon the time and effort that has been put into the pre-testing of tests, questionnaires, etc. The usual result of a failure to pre-test is that respondents can be confused and therefore answer inappropriately. When obtaining measures of student achievement, pre-testing is absolutely necessary so that items may be checked for their difficulty and discrimination levels. If items are either too hard or too easy, there will be little discrimination in the resulting test score.

Where open-ended responses are to be subsequently coded into categories, pre-testing can assist in the development of the categories, or can even lead to eliminating the need for a separate coding step. For example, in one international survey, students were asked to indicate the total number of brothers and sisters in their family. The question was asked in the form of a forced choice response with the maximum value being 'five or more' and in several countries more than 80 per cent of the children indicated this category as their choice. Pre-testing would have indicated the need to extend the number of categories to allow for the very large family sizes in these countries, or to leave the question open-ended.

Many of the weaknesses and limitations of educational research stem from inadequacies in the measures that have been used. Simply put, if the measures that are used to answer research questions are deficient, it will not be possible to obtain correct answers to the research questions. Four major questions that should be asked in selecting or devising measures for research studies are:

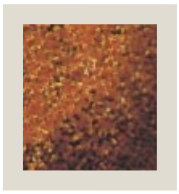
- What validity evidence is available to support the use of a measure? Validity refers to whether an instrument is measuring what it is supposed to measure. For published instruments developed by others there should be adequate documentation to justify their use with the intended target population. If such information is not available, pilot studies may need to be

undertaken to establish a basis for usage. For locally developed measures such as achievement tests, validity can be built into the measure through the use of a carefully developed test plan. However, the development of a test according to a detailed test plan may not guarantee validity of the measure. For example, a test of science achievement may contain so much verbal material that for a student to score well he/she must demonstrate a high level of reading comprehension as well as the relevant science knowledge. This would invalidate the test as a measure of science achievement. A suitable remedy in this situation would be to rewrite the items of the test in a simpler language.

- What reliability evidence is available to support the use of a measure. Reliability denotes the accuracy or precision with which something is measured. For published measures, or measures developed by others, it should be expected that reliability information will be available. If such information is not available, it will be necessary to conduct pilot studies to determine the reliability of the instrument. For locally developed instruments, a trial of the instrument will be needed to determine reliability. In general, one should avoid using instruments that test student achievement which have reliability coefficients below 0.8, and definitely not use any achievement test with a reliability lower than 0.7. Such instruments contain so much measurement error that they cannot provide adequate answers to research questions.

- Is the measure appropriate for the sample? Instruments are developed to be used with particular groups of people. A science test, developed for use in one region, may be inappropriate for use in another region where the curriculum is different. Careful review of an instrument, along with some pilot work, may be necessary to determine the suitability of an instrument for a particular group of people.

- Are test norms appropriate? Sometimes norms are available for help in interpreting scores on various measures. If the group to which a test is to be given is a sample from the target population on which the norms were developed, norms can be a useful aid in interpreting test performance. To do so, several requirements must be met. First, the sample that takes the test must be clearly a part of the population on which the norms were developed. Second, the test must be given without any alterations such as omission of certain items or changes in directions for administration. Third, the time limits for the test must be strictly followed. All of these conditions must be met if norms are to be used.



9 Other issues that should be considered when evaluating the quality of educational research

Previous sections have identified specific threats to the integrity of research studies. In this section more general considerations regarding the quality of research studies are presented. Some of these may strike the reader as being self-evident. If this is the case, then the reader has a fairly good grasp of the basic canons of research. Unfortunately, too many studies are carried out in which these canons are violated.

Every research study should contain a clear statement of the purpose of the study. Furthermore, such a statement should appear very early in the report. If specific hypotheses are being tested, these too should be clearly stated early in the research report. If no hypotheses are being tested but rather research questions are being posed, then these too should be presented early in the report. The frequency with which such dicta are violated is astonishing. Unless the reader is informed early on as to the purpose of a study and the questions to be answered, it is difficult to judge a presentation. Usually, failure to state the purpose of a study early in the report is an indication that the author is unclear about the nature of the study. If this is the case, then it is likely that the research report will contain little that is of value.

A second issue that bears some comment is the review of the literature. Reviews of the literature are intended to furnish the reader with some background for the study. These can range from just a few pages to a lengthy chapter. Practice varies considerably and there are no firm rules to follow. Also, space in publications is usually at a premium and writers are frequently asked to trim the review of the literature to a bare minimum. Despite these factors, some review of the literature is needed in a research report. The review is intended to inform the reader of the existence of previous work in the area and provide a foundation for the present study. It also furnishes some minimal assurance of the writer's familiarity with the area being studied and the likelihood that whatever errors occurred in previous studies are not apt to be repeated.

There are some types of errors that occur in a research situation that have come to be given particular 'names'. They are generally associated with experimental studies. They include the following: the *Hawthorne effect*, the *John Henry effect*, the *Pygmalion effect*, and *Demand characteristics effect*.

The *Hawthorne effect* refers to an effect detected in early studies of worker morale where the fact that subjects were aware of being involved in an experiment resulted in increased output and morale, regardless of the nature of the particular treatment to which they were exposed.

The *John Henry effect* is the opposite of the previously described threat to internal validity referred to as 'resentful demoralization'. In the John Henry effect, students and teachers in the group not receiving the experimental treatment, and knowing that they are not receiving it, join together in putting forth greater effort to perform well. Such increased effort would probably not have occurred if an experiment was not being conducted. The consequence of the John Henry effect is higher performance of the control group leading to a misleading result of no difference between the experimental and control groups.

The *Pygmalion effect* refers to experimenter expectancy effects that can influence student performance. The claim for this effect originated in a study conducted some years ago in which teachers were told that some of their students would have a spurt in mental growth during the course of the school year. In fact, the students who were supposed to show this increase were chosen at random from students in the class. Some support for this effect was found in the lowest two grades of the school, but the study's results were disputed by other researchers. Other studies have detected an expectancy effect, especially when one group is identified as low performers. If a group is labelled as low performers (whether they actually are or are not), this can result in inferior treatment and resulting low performance. While the available evidence does not show strong evidence for such an effect, the possibility for it to occur does exist in some cases.

The term *Demand characteristics effect* is concerned with all the cues available to subjects regarding the nature of a research study. These can include rumours as well as facts about the nature of the study, the setting, instructions given to subjects along with the status and personality of the researcher. These can influence the research study and, more importantly, the results. At present, research is underway to determine the conditions under which such factors may influence the outcomes of studies.

10 A checklist for evaluating the quality of educational research

The following framework for evaluating research reports (see *Box 1*) has been adapted from material developed by Tuckman (1990) and Ross et al. (1990). It is intended to furnish educational planners with a set of criteria for judging research reports. While some readers may be tempted to use the criteria included in the framework without attending to the rest of the material already presented in this module, it is strongly recommended that this not be done. The criteria represent a summarization of the concepts that have been presented earlier and are therefore likely to lack meaning unless the reader has read the entire document. Therefore, it is to be hoped that readers will devote as much attention to understanding the ideas that underlie the criteria presented below as to the criteria themselves.

To assist the reader in understanding and using the criteria, a study from the literature will be examined with regard to these criteria. The study that has been selected is 'Adult Education Project – Thailand' by Thongchua, V.; Phaholvech, N.; and Jiratatprasoot, K. It was published in *Evaluation in Education: an international review series*, 1982 Vol. 6, pp. 53-81, and is reproduced in the *Appendix*. This study sought to determine the effects of several courses of vocational instruction, namely courses in typing and sewing of 150 hours and 200 hours duration.

Box 1 Research evaluation framework

1. **Problem**
   a. is stated clearly and understandable;
   b. includes the necessary variables;
   c. has theoretical value and currency (impact on ideas);
   d. has practical value and usability (impact on practice).
2. **Literature review**
   a. is relevant and sufficiently complete;
   b. is presented comprehensively and logically;
   c. is technically accurate.
3. **Hypotheses and/or questions**
   a. are offered, and in directional form where possible;
   b. are justified and justifiable;
   c. are clearly stated.
4. **Design and method**
   a. is adequately described;
   b. fits the problem;
   c. controls for major effects on internal validity;
   d. controls for major effects on external validity.
5. **Sampling**
   a. gives a clear description of the defined target population;
   b. employs probability sampling to ensure representativeness;
   c. provides appropriate estimates of sampling error.
6. **Measures**
   a. are adequately described and operationalized;
   b. are shown to be valid;
   c. are shown to be reliable.
7. **Statistics**
   a. are the appropriate ones to use;
   b. are used properly.
8. **Results**
   a. are clearly and properly presented;
   b. are reasonably conclusive;
   c. are likely to have an impact on theory, policy, or practice.
9. **Discussion**
   a. provides necessary and valid conclusions;
   b. includes necessary and valid interpretations;
   c. covers appropriate and reasonable implications.
10. **Write-up**
    a. is clear and readable;
    b. is well-organized and structured;
    c. is concise.

Suggested guide for scoring

5 = *As good or as clear as possible; could not have been improved.*
4 = *Well done but leaves some room for improvement.*
3 = *Is marginal but acceptable; leaves room for improvement.*
2 = *Is not up to standards of acceptability; needs great improvement.*
1 = *Is unacceptable; is beyond improvement.*

The objectives of the evaluation project were as follows:

- "measuring the skills and knowledge gained by participants in typing and sewing courses of different duration (e.g., 150 hours and 200 hours);
- identifyg variables having an effect on the achievement of participants at the end of the course;
- investigating whether the graduates took up employment in typing/sewing within six months of the end of the course;
- assessing how participants utilized the skills they learned on the course six months after completing their courses." (p. 54).

1. Problem

The first criterion in the framework involves the research problem. The above statement of objectives was clear and understandable, had some theoretical value and considerable practical value. While the necessary variables were not explicitly stated, they were strongly implied. The first objective referred to skills and knowledge to be gained in typing and sewing courses. Furthermore, the third and fourth objectives specified employment in typing or sewing within six months of the end of the course and use of skills within six months after completing the course. The second objective referred to variables "...having an effect on achievement of participants" but did not specify what these variables were. These variables were presented later in the report of the study. A rating of 4 would seem suitable for the statement of the problem.

2. Literature review

The study report contained no literature review. This may have been omitted due to the extreme length of the report (28 pages) or the fact that the project was one that had an extreme practical orientation. In

any case, the omission of any literature review was troubling. If one were to be generous, one could give the study a rating of NA (Not Applicable). The alternative would be to give it a rating of 2.

3. Hypotheses and/or questions

Since the study was not an experimental one, no formal hypotheses were stated. The research questions were, however, strongly implied in the statement of objectives quoted above. They were further elaborated in the text. The expectation of the investigators was clear; training in typing and sewing was expected to have a positive effect on the skill and knowledge of the participants. Furthermore, the training course was expected to lead to employment in either typing or sewing or, at the least, use of these skills in the future (six months after the completion of the course). A rating of 4 would seem appropriate here.

4. Design and method

The authors described the procedures they followed in the conduct of the study in considerable detail (pp. 54-57). The organization of the study was presented with great clarity, including the problems encountered in selecting sites and participants. These were considerable due to the fact that there was inadequate information about what courses were offered in each of the 24 Lifelong Education Centres throughout the country. Accordingly, the investigators had to conduct no less than three surveys in order to find out what courses were being offered in what centres. The survey results led to some major modifications in the study plan. For example, it was originally envisaged that courses of 150 and 300 hours duration would be compared. However, the survey results revealed that very few centres offered courses of 300 hours duration so the study plan was adjusted to compare courses of 150 and 200 hours duration. In general, the design did fit the problem

fairly well. However, the study was not able to control adequately for either internal or external validity. The reason for this was that the researchers had virtually no control over the selection of participants for the study or the assignment of participants to the different length courses. Another design issue that the authors should have advanced was the rather limited time of 6 months that was used for the tracer study. Given the prevailing economic conditions, perhaps at least one year would have been more appropriate. At best, a rating of 3 must be given on this criterion.

5. Sampling

There were a number of difficulties encountered in sampling. First, the selection of which of the Lifelong Education Centres would be included presented problems. According to the investigators, "The selection of centres was made by purposive sampling rather than simple random sampling as originally foreseen, because of time limitations and in order to have a sufficient number of participants" (p. 55). The use of 'purposive sampling' was questioned earlier in this module. It is a somewhat elegant way of saying that the sampling was less than desirable. Second, within each centre, already established classes of 150 or 200 hours duration were selected at random with the proviso that the number of participants per centre should be at least 30. This condition was not always met. More serious, however, was the use of intact classes. The inability randomly to assign participants to classes of different durations presents real problems since it is possible that more able participants could be assigned to one type of class length. Third, the data collection was seriously compromised in some cases. The authors report, "In some cases, centres had already completed the course a week or two before the dates arranged with the centre for the administration of the tests. The teachers tried to persuade the participants to return to take the tests but, unfortunately, many of them did not do so" (p. 55). In addition, "...participants enrolled but never attended the course or dropped out of the course before it

finished because they had found a job, were already satisfied with what they had learned, became tired of the course, or were needed on the land for seasonal work" (p. 55). The effect of these events was to introduce a bias into the study whose influence is unknown. The lack of a clear definition of a target population, the lack of probability sampling, and the difficulties encountered in actually obtaining the sample raise serious questions about the adequacy of the groups that were studied. A rating of 2 on this criterion seems warranted.

6. Measures

There were several measures used to assess the achievement of the participants. Tests were developed in both typing and sewing for the study. In addition, questionnaires were developed to assess the background of participants and other variables of interest. A teacher questionnaire was also developed for use in the study. The description of the development of the instruments was quite thorough (pp. 57-64) and there is considerable evidence of content validity. The reliability of the cognitive tests of typing were rather low (Thai + English = .60 and Thai = .64). The performance tests, in contrast, showed high reliabilities (typing = .87 to .95 and sewing total = .92). Clearly, the instruments are one of the outstanding features of the study. A rating of 4 or 5 is clearly warranted.

7. Statistics

The authors used standard statistical procedures to analyze their data. They presented the means obtained on each measure for each subject for each course length and indicated whether differences between groups were statistically significant or not. Unfortunately, they did not present the standard deviations of the scores. This made it somewhat difficult to interpret the results that were obtained. For example, the authors report (p. 70) a difference

of 0.78 points between students basic knowledge on the Thai + English typing knowledge test. This difference was indicated to be statistically significant. This is all well and good, but one is also concerned as to whether the difference is meaningful or not. The inclusion of the standard deviations for the two groups would have enabled a reader to judge the meaningfulness of the difference (the means for the two groups were 7.23 and 6.45). A rating of 4 on this criterion seems warranted.

8. Results

Some comments on the results were presented under the above criterion. The general lack of differences between participants in the 150 hour course and those in the 200 hour course are notable and led to the conclusion that little was gained by having courses of more than 150 hours. This is a finding that is likely to have considerable impact on policy and practice. At the conclusion of the report, the authors suggest that, "Serious consideration should be given to abandoning the 200 hour sewing course ..." If adopted, such a recommendation would have a great impact on policy and practice and should result in a considerable saving of money. A rating of 4 seemed appropriate.

9. Discussion

The authors provided a thoughtful discussion of their results. At the beginning of the study, the authors expressed the expectation that graduates of the programme would be able to use their newly acquired skills in typing or sewing to gain employment. Subsequently, they found out that employment opportunities were quite limited and that participants used their newly acquired skills in sewing, for example, to sew clothes for themselves. The absence of economic opportunities for programme graduates was therefore not interpreted as any reflection on programme

effectiveness but rather a reflection of circumstances beyond the control of programme personnel. In contrast, the authors did not give sufficient attention to the differences between programme participants in the courses of different lengths. In the 150 hour sewing course, for example, 63 per cent of the participants were between ages eleven and twenty while only 37.5 per cent were at that age level in the 200 hour course. Furthermore, 47 per cent of the participants in the 150 hour course reported not having a sewing machine at home compared to 28 per cent in the longer course. Whether these differences might have affected performance in the course is simply not known. There were also some substantial differences between participants in the short and longer typing courses. Again, how these differences might have affected performance is unknown. It seems that in the light of the differences that existed between the groups before the start of the courses, one must be extremely careful in attributing performance differences to the treatment effect (duration of the course). It is quite possible that the differences that were found could be due to the differences that already existed between the groups. At the least, one must be quite tentative in drawing conclusions about treatment effects. It would seem that a rating of 3, at best, should be accorded the study on this criterion.

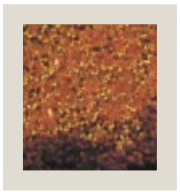
10. Write-up

Many of the comments that have been made about this study could only be given because the authors were so clear and thorough in their write-up. Despite a few omissions – standard deviations for the performance measures, for example – the authors described their study in a clear and thorough way. Some of this may be due to the fact that the authors were given a generous page allotment by the editors of the journal (28 pages). In any case, the study is almost a model for a clear, coherent presentation of a research endeavour. A rating of 4 or 5 is clearly warranted on this criterion.

A summary of the ratings is presented below:

	Criterion	Rating
1.	Problem	4
2.	Literature review	NA or 2
3.	Hypotheses and/or questions	4
4.	Design and method	3
5.	Sampling	2
6.	Manipulation and measures	4
7.	Statistics	4
8.	Results	4
9.	Discussion	3
10.	Write-up	4 or 5

A summary of the type presented above provides a quick indication of the strengths and weaknesses of the study. Clearly, there were problems in the areas of design and method and sampling. These were fully described in the report of the study and commented in the analysis presented above. A reader of the report, using the criteria in the framework, can easily detect the areas where problems occurred and direct attention to these areas in judging the adequacy of the study and how much weight to give to the results. It is hoped that readers of educational research reports will find the criteria helpful in judging the research reports they are presented with.



Summary and conclusions 11

This module has sought to furnish a guide to readers of educational research reports. The intended audience for this guide is educational planners, administrators, and policy-makers. A sincere effort has been made to make this module as non-technical as possible. It is doubtful that this effort has fully succeeded though. There are technical issues involved in research studies and any attempt to avoid them would be irresponsible. Rather, technical issues have been addressed when necessary and an attempt has been made to address them in as simple and direct a way as possible. It is felt that the ideas that underlie technical issues in research are well within the grasp of the readers of this module and that assistance with technical details can be sought when needed.

Readers of educational research reports are not asked to suspend their own judgment when they undertake to read and understand research reports. Rather, it is hoped that the same abilities of analysis and interpretation that they routinely use in their professional lives will be applied to their judgments of reports of educational research. As a colleague in philosophy once noted, "There is no substitute for judgement." It is this same quality of judgement, aided by some technical understanding, that should be used when faced with reports of educational research. There is also ample room for the reader's commonsense. Too often, educational research reports announce what are termed 'significant effects' as though this is all that is needed to make the results educationally important. The term, 'significant effects' merely denotes that an effect is not likely to be due to chance. Whether the effect is large enough to be educationally meaningful is another matter entirely and depends on the judgement of the reader. It is at this point that

the perspicacity of the reader is needed and no amount of technical competence will substitute for careful judgement.

The checklist that was presented for evaluating the quality of educational research in the previous section was intended to serve as a guide. It is hoped that it will be a useful guide, helping to raise important questions that should be addressed when judging research reports. The guide needs to be used judiciously, however. Some of the questions and criteria in the guide may not be applicable to a particular research study. If so, the reader should simply disregard them. Thoughtful use of the guide includes being able to disregard parts of it when they seem irrelevant or inappropriate.

Appendix

Adult Education Project: Thailand[1]

In 1976, a new adult education project was started within the general area of non-formal education. The overall objectives of the programme were to promote literacy skills, occupational skills, spare-time earnings, knowledge, skills and attitudes for functioning in society and, thereby, improving the living standard of the rural population. The project is still continuing.

Nationally, the Project Office of the Department of Non-Formal Education is responsible for the overall co-ordination of the project and looks after such matters as project administration, construction, procurement, expert services, fellowships and apprenticeships, and the implementation of radio correspondence programmes.

There are also four regional offices – in the north-east, north, south, and central Thailand. The regional offices are responsible for servicing the various activities in non-formal education in their region. They undertake curriculum development and the production of materials relevant to the needs of their region, training, and some research and evaluation.

One of the major activities is the work conducted by the Lifelong Education Centres. Each centre provides courses in adult continuing education, functional literacy, vocational education, and in topics of special interest to groups requesting them. Each centre also provides services for its immediate neighbourhood such as public libraries, village newspaper reading centres, certain audio-visual programmes and what are known as special activities (examples of

1. Extract [(Chapter 3) by Viboonlak Thongchua, Nonglak Phaholvech, Kanjani Jiratatprasoot] in *Evaluation in education: an international review series.* Vol. 6, No. 1, 1982 (pp. 53-107). Sawadisevee, A.; Nordin, A.B.; Jiyono; Choppin, B.H.; Postlethwaite, T.N. (Eds.). Oxford: Pergamon Press.

which are special talks on the law and elections, and participation in special activities of the province).

The vocational education work is considered to be the most important activity of these centres. The two most popular courses in 1980 were typing and sewing, and these were selected for evaluation.

The participants in these courses are thought typically to be 14 to 25 years old from poor rural homes and seeking permanent employment. At any one centre, a course is provided for a minimum of 15 persons. The overall purposes of the typing and sewing courses are that the participants will become proficient in the skills of typing and sewing. For typing the hope is that they will take jobs after the course and so increase their family income. For sewing the hope is that they will either take jobs or will sew clothes for their families thereby reducing the amount of money they spend on clothes and thus increasing the family's income. The participants all come from regions where the daily per capita income is about US$2.

Objectives of the evaluation project

The general aim was to measure achievement of participants in the typing and sewing courses, identify the variables that affected achievement and also examine how participants utilized the skills they had learned at the centres. These overall objectives were broken down into the following specific objectives:

- measuring the skills and knowledge gained by participants in typing and sewing courses of different duration (e.g. 150 hours and 200 hours);

- identifying variables having an effect on the achievement of participants at the end of the course;

- investigating whether the graduates took up employment in typing and sewing within six months of the end of the course;
- assessing how participants utilized the skills they learned on the course six months after completing their courses.

Design

A cross-sectional survey was conducted throughout the 24 Lifelong Educational Centres to acquire a sufficient number of courses for both subject areas. Practical tests and a cognitive test (typing only) were administered at the end of each course to assess the courses and to identify variables associated with achievement. Six months after the end of the courses, a tracer study was conducted to assess how and to what extent the graduates were using the skills learned.

There are 24 Lifelong Education Centres throughout the country offering the vocational courses. It was decided to select five centres offering the typing (150 hours) course and five for the typing (200), five for the sewing (150) and five for the sewing (200). Therefore, 20 of the 24 centres were involved.

In each course at each centre, one class of 30 participants was chosen at random. If 30 participants could not be found in one class, another class would be added to make up the 30 participants. Thus, there would be 150 participants for each of the four courses making a total of 600 students.

For the tracer study, a sub-sample of the graduates would be selected at random, to be followed up. This would be eight participants from each course at each centre, making a total of 40 participants for each of the four types of courses.

To obtain the sample, three surveys at different times were conducted. The first survey was launched in early October, 1980, in order to identify the centres offering 150 and 300 hours for both subjects. Originally it had been thought desirable to compare courses of 150 and 300 hours. However, the first survey indicated that very few centres conducted courses of 300 hours. The idea of comparing 150 and 300 hour courses was then discarded and a second survey was conducted to assess whether it would be possible to compare courses of 100 and 200 hours. The second survey (later in October, 1980) showed that there were insufficient classes of 100 hours duration. An attempt was then made to identify sufficient 150 and 200 hour courses. A third survey was conducted in November/ December, 1980 and finally an appropriate number of courses was identified.

However, it was difficult to identify sufficient centres providing courses which had the required number of participants and, at the same time, fell within our time limits for data collection. The time limit was therefore extended from the end of March to early June, 1981. The selection of centres was made by purposive sampling rather than simple random sampling as originally foreseen, because of time limitations and in order to have a sufficient number of participants. Purposive sampling involved all centres which completed their courses between March and early June, 1981. There were simply not enough centres to make random sampling possible.

Although the sampling design called for 30 participants per centre, the final number fell short of 30. In some cases, centres had already completed the course a week or two before the dates arranged with the centre for the administration of the tests. The teachers tried to persuade the participants to return to take the tests but, unfortunately, many of them did not do so. The reason given was that most of the participants were afraid to take the tests.

In other cases, participants enrolled but never attended the course or dropped out of the course before it finished because they had found a job, were already satisfied with what they had learned, became tired of the course, or were needed on the land for seasonal work.

To sum up, the total achieved sample was 498, divided into 135 for sewing (150 hours), 147 for sewing (200 hours), 130 for typing (150 hours), and 86 for typing (200 hours). The achieved sample is presented in *Table 1*.

In order to obtain fairly equal numbers in both sewing groups, one centre (Samusakorn) was added to the sewing 150 hour course and one centre (Ayuthaya) was removed from the sewing 200 hour course.

Given the major constraints of money, time and manpower, we were unable to equate the groups any better. Because of the reduction in sample size, the sub-sample size for the tracer study was also affected given that about 25 per cent were to be followed up. The sub-sample design for the tracer study became that presented in *Table 2*.

In order to ensure that a 25 per cent sub-sample could be met, all course participants were invited to come to their centres. Those who did not were visited in their homes, but some could not be contacted. The final tracer study included 63 per cent of course graduates as presented in *Table 3*.

Table 1 **Total achieved samples in each center**

Sewing									
Center 150 hours	**Morning shift**	**Afternoon shift**	**Evening shift**	**Total**	**Center 200 hours**	**Morning shift**	**Afternoon shift**	**Evening shift**	**Total**
1. Chiengmai	-	23	14	37	1. Nakornswan	14	12	15	41
2. Khonken	11	13	6	30	2. Petchaboon	47	-	7	54
3. Nakornratchasima	10	-	11	21	3. Ratburi	14	-	8	22
4. Nakronswan	7	8	-	15	4. Surin	26	-	4	30
5. Samusakorn	-	-	9	9					
6. Ubonratchatanee	12	11	-	23					
Total	**40**	**55**	**40**	**135**	**Total**	**101**	**12**	**34**	**147**
Typing									
1. Chiengmai	9	5	19	33	1. Angthong	2	-	12	14
2. Khonken	9	12	9	30	2. Ayuthaya	5	1	3	9
3. Nakronratchasima	16	-	6	22	3. Petchaboon	15	5	-	20
4. Uthaithanee	12	-	1	13	4. Samusakorn	-	-	17	17
5. Ubonratchatanee	9	17	6	32	5. Surin	9	10	7	26
Total	**55**	**34**	**41**	**130**	**Total**	**31**	**16**	**39**	**86**

Table 2 **Sub-sample for tracer study**

No.	Center	Sewing		Typing	
		150 hours	200 hours	150 hours	200 hours
1.	Angthong	-	-	-	3
2.	Ayuthaya	-	-	-	2
3.	Chiengmai	9	-	9	-
4.	Khonken	8	-	8	-
5.	Nakornratchasima	5	-	5	-
6.	Nakornswan	4	10	-	-
7.	Petchaboon	-	14	-	5
8.	Ratburi	-	6	-	-
9.	Samusakorn	2	-	-	4
10.	Surin	-	7	-	7
11.	Uthaithanee	-	-	3	-
12.	Ubonratchathanee	6	-	8	-
Total		**34**	**37**	**33**	**21**

Table 3 Achieved tracer sub-sample

No.	Center	Sewing		Typing	
		150 hours	200 hours	150 hours	200 hours
1.	Angthong	-	-	-	11
2.	Ayuthaya	-	-	-	4
3.	Chiengmai	29	-	19	-
4.	Khonken	16	-	14	-
5.	Nakornratchasima	14	-	8	-
6.	Nakornswan	14	21	-	-
7.	Petchaboon	-	49	-	15
8.	Ratburi	-	19	-	-
9.	Samusakorn	8	-	-	13
10.	Surin	-	20	-	10
11.	Uthaithanee	-	-	6	-
12.	Ubonratchathanee	9	-	16	-
Sub-Total		**90 (66.7%)**	**109 (74.1%)**	**63 (48.5%)**	**53 (61.6%)**
Grand TOTAL		**315 (63.25%)**			

According to information obtained from the instructors of the courses, they represented both good and poor performers on the courses. The remaining students who did not come to the interview session gave as reasons for being absent that they did not receive the postcards, happened to be doing important business on that day, or had moved to another area. A detailed analysis was carried out to investigate whether there were important differences between the characteristics of those students who attended the tracer study interview (the tracer sample), and those who did not attend and could not be contacted (the drop-out sample). We compared these two groups on their background characteristics (sex, age, occupation, level of education, family size, previous level of skill, etc.), teacher information (sex, age, teacher experience, teacher training, additional training, etc.) and information on the students' test performance.

The differences between the tracer sample and the drop-out sample were small – being less than one third of a standard deviation score on each student characteristic. The largest differences were noted for hours of attendance in the sewing and typing courses. The maximum difference in attendance hours was 4.8 hours for the typing course.

This figure was relatively small in comparison to the total length of the courses (150 to 200 hours).

These analyses demonstrated that, although the drop-out sample consisted of some 37 per cent of the total sample, the loss of this information (in the tracer study) had not created serious problems of sample bias in our tracer study.

The construction of measures

In all, five measurement instruments were constructed. A teacher questionnaire was developed by the team members. We took it with us while conducting the field study and asked the teachers to complete it. It contained questions on educational and teaching qualifications, sex, age, teaching experience, additional training on the subject taught within the past five years, number of participants in the classes and their attendance record, and equipment facilities (quantity and quality).

There was one cognitive test for typing and practical tests in both typing and sewing with a one-page student questionnaire on student background information (sex, education, age, previous experience, motivation, siblings, father's education, mother's education, father's occupation, mother's occupation and machine at home). A three-day workshop was held in Bangkok for course content analysis and test construction. Eight teachers from Lifelong Education Centres, mobile Trade-Training Schools and the Non-Formal Education Department at the Ministry of Education, as well as two local experts, were invited to participate in this workshop. Cognitive test of typing: content analysis of the curricula for the 150 and 200 hour courses was undertaken by the instructors teaching the typing courses and team members. *Table 4* presents the topic areas and objectives, and the number of items per topic. The number of items per topic represents the weights accorded to each topic. The items were in multiple-choice form with four alternatives per item, only one was the correct answer.

Table 4 **Topics (objectives) and numbers of items in pilot and final cognitive tests for typing**

Topics	Number of items	
	Pilot	Final
1. Basic knowledge of typing	6	11
2. The parts of a typewriter and how to handle them	5	6
3. Maintenance of a typewriter	2	1
4. Body position when typing	3	2
5. How to feed and release paper and set intervals	7	8
6. How to manipulate typewriter keyboard	7	6
7. Gaining typing speed	3	4
8. Carbon copy typing	2	2
9. Typing on stencil	2	2
10. Principles of typing official letters	13	13
Total	**50**	**55**

The pilot test was administered to 134 students in a polytechnic school in Bangkok and at Adult Education centres in Khonken and Petchaboon. No difficulties were experienced with the administration of the test. An item analysis was undertaken. Thirty-five of the items were in the 20 to 80 per cent difficulty range and had point-biserial correlations of over .30. Fifteen of the items were either too easy or too difficult, or had low discrimination values. The reliability of the pilot test was low (KR -20; .62) but it was hoped that, by substituting twenty better questions for the fifteen poor ones, higher reliability would be obtained. This then made for a final test composed of 55 items (45 items for Thai typing plus 10 items for English typing) as presented in *Table 4*. The psychometric

characteristics of the final test were: for Thai + English typing with a maximum score of 55: X = 32.25 SD. = 5.621 and a KR 21 of .596; for Thai typing with a maximum score of 45: X = 23.298 Sd. = 5.408 and a KR 21 of .636.

Practical test of typing: since, in the curriculum, courses were offered for typing both in Thai and English, practical tests were constructed for both languages. In the event very few participants took courses in English typing, however, both tests were analysed.

A Thai text consisting of 678 strokes and an English text consisting of 598 strokes were selected by the typing instructors. A participant would type each text twice and would hand in the better version.

Two scores were calculated. The first was a combined speed and accuracy score which was the number of correctly typed words per minute. The formula used was:

$$\textbf{No. of words} = \frac{\textbf{No. of strokes}}{\textbf{5 or 4}}$$

$$\textbf{No. of words/minute} = \frac{\textbf{(No. of words) – (No. of wrong words x 10)}}{\textbf{Time in minutes}}$$

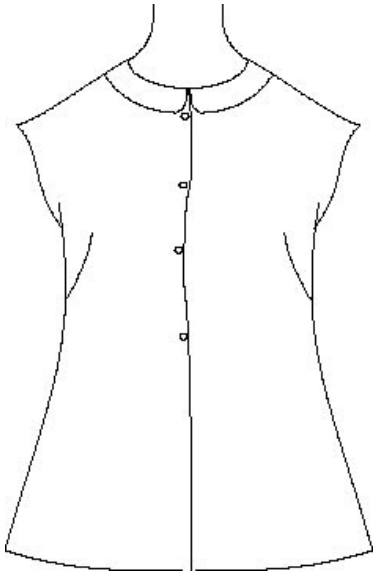
In the number-of-words formula, the denominator was 5 or 4 because this was the presumed average number of letters per word in the English and Thai languages. This was the generally used formula in non-formal education typing classes in the Lifelong Education Centres in Thailand. In the number-of-words per minute formula, the constant 'x 10' was used in the numerator because it was assumed by the instructors that time spent in correcting one mistake was equal to the time spent in typing 10 strokes. A separate score was also given for a combined format and tidiness measure. The test was piloted in the same schools as the cognitive typing test. No problems were experienced in its administration. No changes were made in the test. The psychometric characteristics of the final tests are given in *Table 5*.

Table 5 Psychometric characteristics of the final tests

	Maximum score	Mean	S.D.	KR 21
Thai + English typing				
Speed + accuracy	60	16.193	10.724	.914
Format + tidiness	10	6.521	3.220	.867
Thai typing only				
Speed + accuracy	30	7.298	7.009	.917
Format + tidiness	5	2.721	2.115	.948

Practical test of sewing: It was agreed that five types of garments where there was a good deal of variations in both the process and the product would be the subject of the sewing test. The garments were: a blouse with long sleeves, a blouse with short sleeves, a blouse with built-in sleeves, a skirt with one fold in the front and one at the back, and a six-piece skirt. After the pilot work, five garments were scored according to the criteria given in the following paragraph. The 'blouse with built-in sleeves' (see *Figure 1*) had a normal distribution of scores.

Figure 1 A blouse with built-in sleeves



The experts selected this for the final testing. The scoring criteria were set by the sewing instructors and approved by the local experts. The scoring system was set by examining the difficulty and time consumed in the sewing process. Thus, the experts agreed that out of a total of 100 points, body measurement should receive 10 points, building pattern 25 points, laying and cutting fabric 25 points, and sewing 15 points.

1. It was agreed that there were 10 basic body measurements needed for the blouse with built-in sleeves and there would be one point for each measure. These were (i) shoulder to shoulder, (ii) chest, (iii) neck to breast and breast to breast, (iv) waist, (v) round shoulder and underarm, (vi) and (vii) between arm pits (front and back), (viii) and (ix) neck to waist (front and back), and (x) shoulder to length desired.

2. Building pattern (25 points) was divided into five sections; each section was awarded five points. They were (i) front and back pieces, (ii) collar, (iii) sleeve, (iv) bent lines, and (v) calculation. In each section five points were awarded if the measurement figures were correctly converted into pattern figures and the pattern was drawn correctly. One point was subtracted for each mistake.

3. Laying, tracing and cutting (25 points) was divided into three sections. They were (i) front and back pieces (10 points), (ii) collar (10 points), and (iii) interfacing of collar and sleeves (5 points). Scoring depended on how well students laid the cloth in relation to its line (grain), on whether they left enough spare cloth for sewing, and on whether they cut out all the pieces correctly. Three points were subtracted for each mistake in laying, cutting, front-and-back pieces, and collar; and two points for incorrect interfacing.

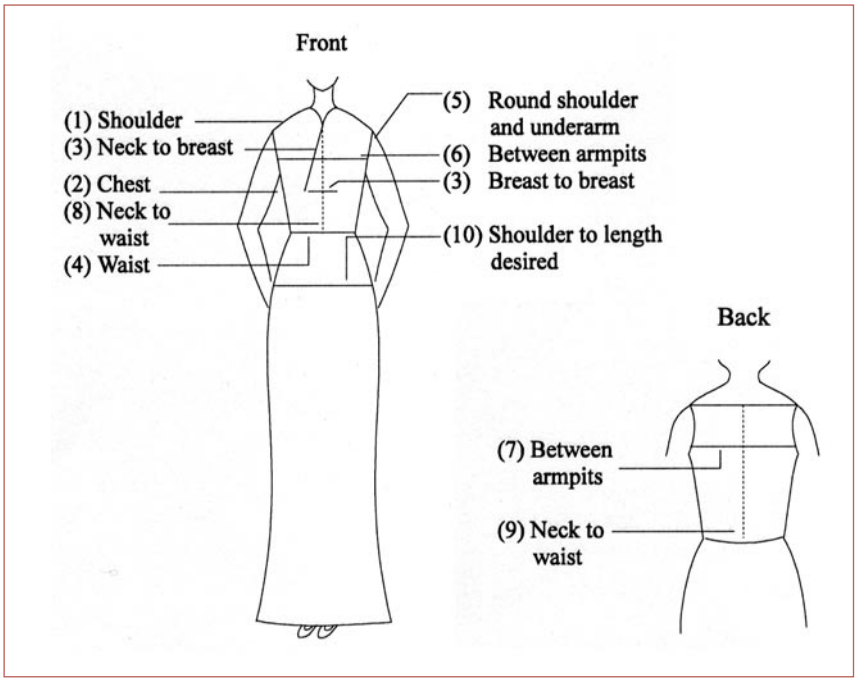
4. Sewing (15 points) was divided into five sections and each section was awarded three points. These were: (i) sewing the shoulder and side seams, (ii) sewing the collar, (iii) sewing the sleeves, (iv) sewing the button-holes and buttons, and (v) hemming. If the participants did the sewing process correctly, i.e., started from part 1 and went through to part 5 in the correct order, they would get a full score. If they jumped one step, they lost three points. Sub-scores were calculated for each of the four major process sections.

There were two major sub-scores for product: goodness of fit (10 points) and tidiness of sewing (15 points). (See *Figure 2*).

1. Goodness of fit when the dress was fitted to the model was also divided into five parts, each receiving two points. The parts were body, sleeves, collar, length of blouse, and overall goodness of fit. Two points were awarded for the fit in each part of the text.

2. Tidiness of sewing was sub-divided into five parts, each part receiving three points: collar, sewing on buttons, button-holes, stitching, and hemming. The pilot work was undertaken using 40 participants in the Khonken and Petchaboon learning centres. No serious problems were encountered. The time limit was set at five hours and this was adequate for the blouse with built-in sleeves.

Figure 2 Body measurements



At the final testing, two Ministry sewing experts were the scorers. Each scorer scored every garment according to the criteria and their average was calculated to represent a participant's score. Unfortunately, scorers were instructed only to give a global score for each of the six major sub-scores. If we were to repeat the exercise again, we should have each separate item (for example, in body measurement score). The reliability has been calculated for each

sub-score using KR 21. The means, standard deviations and KR 21s are presented in *Table 6*. All items included in the total sewing score had loadings of at least .50.

Interview schedule for tracer study: This instrument was developed by the team members. The questions contained information on the centres from which respondents had graduated, on how much money they could earn, on how much money they could save, on how they utilized the knowledge they gained from the course.

Table 6 **Means, standard deviations, and reliability coefficients for the sewing test**

		Mean	S.D.	KR 21
A.	**Process**			
1.	Body measurement (10)	9.222	1.518	.799
2.	Building pattern (25)	19.294	3.751	.715
3.	Laying and cutting fabric (25)	20.528	2.518	.862
4.	Sewing (15)	10.209	4.440	.906
B.	**Product**			
1.	Goodness of fit (10)	6.816	3.174	.895
2.	Tidiness (15)	9.102	4.180	.867
	Total score	**75.41**	**14.51**	**.921**

Data collection

The research team and staff visited and collected data at sample centres during March and early June, 1981, at the end of the typing and sewing courses. The course instructors were asked to fill out the questionnaires and return them to the research team. At that time the team administered both cognitive and practical tests for both courses to the students.

About six months after giving the tests, the tracer study was carried out by the team members and some additional staff. All data were coded and punched in Bangkok. The analyses were undertaken at the National Statistical Office in Bangkok and the D.K.I. Computer Centre in Jakarta.

Results *Table 7* presents, in percentage form, the characteristics of both students and teachers for the six courses. The data showed clearly that the students coming to the sewing courses were quite different from those who attended the typing courses. In the sewing courses we found only women (except that one man attended one of the short courses), who tended to be older, less well-educated, more often in manual jobs and from manual workers' homes but with more previous experience. They most often chose to do these sewing courses for themselves and their families, rather than to further their careers or to gain educational credits. More than half of them had sewing machines in their homes.

The teachers on these sewing courses were all women who tended to be older and more experienced than the typing teachers, but with fewer formal qualifications.

The long (200 hours) sewing course attenders were an elite. They were older women who already had some experience in sewing and were seeking to reach a higher standard. Sixty per cent of them wanted to sew for themselves and their families, and 72 per cent

Table 7 Student and teacher background variables

Student variables	Typing				Sewing	
	Short (150 hrs)		Long (200 hrs)			
	Thai+Eng	Thai	Thai+Eng	Thai	Short (150 hrs)	Long (200 hrs)
	N=72	N=58	N=47	N=39	N=135	N=147
Sex: Percent women	43.1	55.2	40.4	48.7	99.3	100
Education						
Lower primary	6.9	10.3	-	-	51.9	46.9
Upper primary	25.0	20.7	29.8	-	30.4	23.1
Lower secondary	38.9	48.3	48.9	59.0	8.1	15.6
Upper secondary	20.8	13.8	14.9	23.1	1.5	8.2
Age						
11 - 20	65.3	75.8	80.7	41.0	63.0	37.5
21 - 30	30.7	24.1	19.1	56.5	29.7	42.3
31 +	4.2	-	-	-	6.6	19.7
Previous experience: None	91.7	84.5	89.4	92.3	56.3	36.7
Occupation						
Manual	19.4	17.2	17.0	30.7	55.6	47.6
Student	59.7	60.3	46.8	38.5	8.9	5.4
Motivation						
Credit	16.6	44.8	44.7	7.7	4.4	0.7
Career	54.2	34.5	38.3	64.1	37.8	32.0
Self and family	23.6	12.1	14.9	23.1	53.3	59.9
Siblings: 7+	47.3	24.1	40.5	36.0	34.8	34.7
Mother's education						
Primary	73.6	63.7	51.1	89.7	83.7	69.4
Secondary	8.4	1.3	-	2.6	-	0.7
Father's education						
Primary	68.1	63.8	48.9	66.6	78.5	64.7
Secondary	13.9	6.8	2.1	17.9	4.4	8.1
Mother's occupation						
Manual	63.9	62.1	46.8	74.3	77.0	55.7
Housewife	6.9	13.8	8.5	12.8	3.7	15.0
Father's occupation: Manual	62.5	58.7	49.0	66.7	76.3	61.2
Machine at home: No	98.6	98.3	100.0	94.9	47.4	27.9

Table 7 (continued)

	Typing				Sewing	
	Short (150 hrs)		Long (200 hrs)			
	Thai+Eng	Thai	Thai+Eng	Thai	Short (150 hrs)	Long (200 hrs)
Teacher variables	N=72	N=58	N=47	N=39	N=135	N=147
Sex: Female	43.1	43.1	57.4	100.0	100.0	100.0
Qualification: Certificate	100.0	100.0	100.0	100.0	72.6	77.6
Additional teacher training within 5 years	100.0	77.6	-	35.9	54.1	85.7
Teacher experience						
None	-	-	23.4	84.6	28.9	39.5
11 + years	-	-	-	-	11.1	35.4
Teacher's age						
Over 35	-	-	-	-	17.8	88.4
Course variables						
Class size: Under 16	69.4	60.3	34.0	61.5	43.7	59.9
Shift						
Morning	34.7	51.7	42.6	28.3	29.6	68.7
Afternoon	40.3	8.6	10.6	15.4	40.7	8.2
Evening	25.0	39.7	46.8	56.4	29.6	22.4
Equipment: Lacking	59.1	74.1	76.6	2.6	56.3	72.1
Attendance: 75% +	90.2	82.9	79.0	64.5	92.6	63.5
Quality of facility						
Excellent	-	-	34.0	48.7	17.0	-
Fair	98.6	69.0	66.0	51.3	83.0	100
Poor	1.4	31.1	-	-	-	-

Table 8 Mean achievement scores in sewing for 150 and 200 hour courses

Course hours	Body measure	Building pattern	Cutting	Sewing	Fit	Tidiness	Total
200	9.15	19.33	29.08	11.01	7.29	9.70	77.56
150	9.29	19.44	19.90	9.56	6.37	8.57	73.12
Difference	-0.14	-0.11	***1.18	**1.45	*.92	*1.13	**4.44

had machines at home. They were taught by the oldest and most experienced teachers, generally during the morning shifts. As with all longer courses, however, their attendance record was less complete.

The four varieties of typing courses were generally about half-and-half men and women, and most had at least some secondary education. They were also much younger than the sewing course attenders (for example, 81 per cent of the long Thai plus English course were teenagers), and about half of them were students.

Their teachers, likewise, were about half-and-half men and women (though all the teachers on the long Thai typing course were women) and tended to be much younger and better qualified than the sewing teachers (all of them had at least a Certificate and some had university degrees).

Students in these four courses varied a great deal, but to some extent the long Thai typing course stood out: participants in this course had a better education, they were a little older (mostly in their twenties), they were more often in manual occupations and they were most often career-minded in their motives for doing the course.

Their parents were slightly better educated, though more often in manual occupations than other typing students. More of them attended evening shifts in rather better equipped classrooms where they were taught by an all-women staff of mostly inexperienced teachers. One got the impression here of a group that was striving for upward social mobility by doing a longer Thai typing course in their own time after work.

It was not clear why these latter students would not rather put their energies into a Thai plus English typing course; perhaps their knowledge of English was insufficient. Clearly, students did treat these courses as graded steps in accomplishment: it was not true to

say, for example, that students doing a combined Thai plus English course were already able to type in Thai (about 90 per cent in both the short and long Thai plus English courses lacked previous typing experience). Quite a large number of students in these courses were young teenagers picking up an additional credit while they were studying the adult continuing education by attending day-time classes.

The participants' achievement

The major objective of the study was to examine the relative effectiveness of the two types of courses, 150 hour and 200 hour sewing and typing. The same tests were administered to students in both courses. *Table 8* presents the results for the six sub-tests in sewing.

As can be seen, the total score for the 200 hour course was significantly higher than that for the 150 hour course. The largest differences were for sewing and cutting. However, for body measurement and building patterns, there was no difference. *Table 9* presents the results for typing. As can be seen, the total scores on the cognitive test for longer courses were higher than for shorter courses. However, it was only significant for the Thai typing course. The major difference was for the principle of typing official letters. For Thai plus English typing, the major differences were for 'how to feed and release papers and set intervals' and 'principle of typing official letters'.

For the practical test, the scores for longer courses were not significantly higher than for shorter courses on 'speed'. Remarkably, the scores on 'format plus tidiness' for longer courses were less than for shorter courses.

The objectives of the tracer study were to find out whether the graduates took employment within six months of the end of the course, and how participants used the skills they learned in the course.

The criteria used for these objectives were the amount of money the participants earned, the money they saved, the amount of time spent on typing and the reason why they did not utilize their knowledge and skills. *Table 10* presents this information.

In the sewing courses, we found over 90 per cent of the graduates for both courses used the skills they learned in the courses for themselves and their families. Only 6.7 per cent of participants (1.8 per cent for the long courses) said they did not use the skills they had learned. The reasons they gave included the fact that they had no time, no machine or that they lacked confidence. The data showed that 63.5 per cent of participants from the short sewing course and 46.3 per cent from the long sewing course saved less than 50 Bahts (1 US Dollar is equal to 23 Bahts) per month.

Table 9 Mean achievement score in typing (150 and 200 hour courses)

I

Course	Basic knowledge	Parts	Maintenance	Body position	Feed return	Manipulation
Thai						
200 hours	2.18	2.54	0.69	1.77	2.74	4.90
150 hours	1.84	2.67	0.90	1.76	2.66	4.50
Difference	0.34	-0.13	**-0.21	0.01	0.08	0.40
Thai + English						
200 hours	7.23	3.55	0.77	1.70	2.85	5.06
150 hours	6.45	4.14	0.47	7.78	3.97	4.07
Difference	**0.78	-0.59	0.30	-0.08	***-1.12	***0.99

II

Course	Speed typing	Carbon	Stencil	Principal	Total	Practical	
						Speed	Format
Thai							
200 hours	1.46	1.23	1.13	6.21	24.85	8.59	2.74
150 hours	1.29	1.16	0.90	4.59	22.26	7.47	3.02
Difference	0.17	0.07	0.23	***1.62	*2.59	1.12	-0.28
Thai + English							
200 hours	2.53	1.45	1.02	7.02	33.19	16.47	6.28
150 hours	2.07	1.36	1.44	5.97	31.84	16.01	6.68
Difference	*0.46	0.09	***-0.42	*1.05	1.35	0.46	-0.4

Table 10 Information on tracer study

	Typing 150 hours		Typing 200 hours		Sewing	
	Thai + English	Thai	Thai + English	Thai	150 hours	200 hours
	N = 32	N = 31	N = 32	N = 22	N = 90	N = 109
Earn money or not?						
No	84.4	96.8	93.5	95.5	56.7	58.7
Yes	15.6	3.2	6.5	4.5	43.3	41.3
• ***If they earn, how much/month?***						
	(N=5)	***(N=1)***	***(N=2)***	***(N=1)***	***(N=39)***	***(N=45)***
1 - 50 Bahts	40	-	50	-	46.3	59.6
51 - 100 Bahts	-	100.0	-	-	12.9	13.2
101 - 150 Bahts	-	-	-	-	10.4	2.2
151 - 200 Bahts	-	-	-	-	10.2	-
200 + Bahts	60	-	50	100.0	20.5	24.3
• ***If they don't earn money, why?***						
	(N=27)	***(N=30)***	***(N=29)***	***(N=21)***	***(N=51)***	***(N=64)***
No employment	29.6	36.6	44.8	38.1	19.6	7.8
No time	7.4	33.3	34.5	23.8	27.5	46.9
No machine	22.2	6.7	17.2	33.3	15.7	7.8
Lack of confidence	3.7	6.7	3.4	-	31.4	26.6
Continue studying	18.5	10.0	-	-	3.9	1.6
Work free of charge	-	6.7	-	4.8	2.0	9.4
Sew for self/family?						
No	-	-	-	-	6.7	1.8
Yes	-	-	-	-	93.3	98.2
• ***If sewing, how much money saved per month?***						
					(N=84)	***(N=107)***
1 - 50 Bahts	-	-	-	-	63.5	46.3
51 - 100 Bahts	-	-	-	-	15.6	33.4
101 - 150 Bahts	-	-	-	-	1.2	11.7
151 - 200 Bahts	-	-	-	-	3.6	3.7
200 + Bahts	-	-	-	-	15.6	3.6

Table 10 (continued)

	Typing 150 hours		Typing 200 hours		Sewing	
	Thai + English	Thai	Thai + English	Thai	150 hours	200 hours
	N = 32	N = 31	N = 32	N = 22	N = 90	N = 109
• *If they don't sew, why?*					*(N=6)*	*(N=2)*
No machine	-	-	-	-	33.3	-
Lack confidence	-	-	-	-	33.3	-
No time	-	-	-	-	33.3	100.0
After graduation, ever typed?						
No	21.9	19.4	32.3	40.9	-	-
Yes	78.1	80.6	67.7	59.1	-	-
• *If typed, how many minutes per week?*	*(N=25)*	*(N=25)*	*(N=21)*	*(N=13)*		
1 - 60	64.0	52.0	81.3	61.5	-	-
61 - 120	12.0	8.0	4.8	-	-	-
121 - 180	8.0	-	9.6	-	-	-
180 +	16.0	40.0	4.8	38.5	-	-
• *If don't type, why?*	*(N=7)*	*(N=6)*	*(N=10)*	*(N=9)*		
No machine	85.7	66.7	80.0	85.9	-	-
No employment	-	-	-	-	-	-
No time	-	33.3	10.0	-	-	-
Isn't part of job	14.3	-	10.0	11.1	-	-

In this research study, we defined participants who earned money from sewing or typing as those who took up either permanent or part-time jobs. Only 43.3 per cent of the short course participants and 41.3 per cent of those in the long course took up jobs. Of those, 46 per cent from the short course and 60 per cent from the long course earned less than 50 Bahts per month. Again, the major reasons for not taking up a job were lack of time and lack of confidence.

From the four types of typing courses, many participants stated that they continued to type, i.e., use the skill six months after the end of the course. Seventy-eight per cent from the short course (Thai + English) and 59 per cent from the long course (Thai only) used their typing skills. Those who did not use their skills gave as reasons that they had no typewriter at home, no time or that their occupation did not require the typing skills they had learned in the course. Only a few (15.6 per cent) earned money from typing because many of them were unable to find a job (there were very few employment possibilities in their area) and they did not possess a typewriter at home. Some stated that they lacked confidence or that they typed free of charge for their friends. Many of them had no time because they were still studying.

However, just over 50 per cent typed for their own pleasure up to one hour a week and some of them up to three hours per week. Those who had taken the Thai typing courses spent more time typing than those who had taken the Thai plus English courses. In fact, 39 per cent of Thai typing course participants typed more than three hours a week.

Regression analyses

Two main sets of regression analyses were conducted. The first concerned sewing and the second typing.

Sewing included three regression analyses. The first predicted end-of-course achievement and the criterion was the total sewing score. The second concerned the tracer study and the criterion was the amount of money earned per month six months after the end of the course. The third was money saved per month six months after the course.

For typing, it was intended that a measure of typing speed which incorporated a correction for typing accuracy, and a measure of typing knowledge would provide criterion variables for regression analysis. However, a detailed investigation of the distributions and factorial structures of the two measures indicated that neither would be appropriate for use as a criterion variable. The measure of typing speed had a severely bimodel distribution which occurred because the correction for typing accuracy was applied in a fashion that many students whose 'actual' corrected score was negative were given a score of zero. The measure of typing knowledge was subjected to a principal components' analysis followed by Varimax rotation and it was discovered that there was no sound evidence for assuming that the total score on the test was assessing a single dimension which could be readily interpreted.

Consequently, the regression analyses for typing were limited to the use of only a single criterion: the number of minutes per week spent in typing six months after the end of the course. For each analysis the possible independent variables were scrutinized for skewness. High skewed variables were dropped. Correlations were calculated among independent variables and between these variables and the criterion.

At the same time, we had a chronological model in mind. We assumed that various pre-course variables would be important and that these would be included in one cluster or block of variables because the participants had been exposed to these before coming to the course. Secondly, there were a set of variables which characterized the course itself. These would be entered as a second cluster or block of variables.

Sewing: On the basis of correlation with the criterion and with other variables, the following were selected. For all regressions, father's occupation, participant's age, kits (availability of a sewing machine at home) and previous ability (meaning previous experience in sewing) were included in Block I. For the tracer study criteria, one extra variable was added, namely 'objective', which was a coding of the reason why the participant wished to take the course (1 = for a career, 2 = to gain a credit, and 3 = to sew for the family or for herself).

For the end-of-course achievement, regression Block II included teacher training (the level of pre-service teacher training), additional training, (0 = no additional training, 1 = some additional training in the last five years), facility (0 = teachers perceived the facilities for sewing at the centre to be inadequate, 1 = adequate), quality of facility (1 = very poor, 2 = poor, 3 = fair, 4 = excellent), the shift in which participants attended the centre (1 = morning, 2 = afternoon, and 3 = evening), the age of the teacher, and whether the course was a 150 hour or 200 hour course (1 = 150, 2 = 200). It had been proposed to enter shift as a dummy variable, but because of the correlations, it was decided to leave it coded as 1, 2, 3.

For the tracer study regression analyses, Block II consisted of teacher training, shift, 150-200 hour course, and size of class. For the tracer study, the sewing test score at the end of the course was entered as a third block.

Let us now examine the results. *Tables 11 and 12* report the results for achievement in course and tracer study respectively. The complex multi-stage sample designs employed in this study did not conform to the well-known model of simple random sampling. Consequently it was not possible to employ the standard error estimates provided by SPSS computer programme package (Nie et al., 1975). Instead, it was decided to use the formula provided by Guilford and Fruchter (1973: p. 145) for the standard error of correlation coefficients to obtain approximate estimates for the standard error of a standardized regression coefficient. This decision provided more conservative error limits than the SPSS programme and consequently represented a more realistic approach to testing the significance of the standardized regression coefficients (Ross, 1978). Accordingly, as the number of students in Table 11 is 282, .12 represents two standard errors while the number of students in Table 12 is 199 so 0.14 represents two standard errors. Only 28 per cent of the variance was explained by the variables in the model of predicted achievement at the end of the course (*Table 11*). This is disappointing and clearly more effort will have to be made to identify other variables which are likely to be influencing sewing achievement, and include these in future studies of this kind. The only variables to survive the regression were participants' age, the additional training of the teacher, and the quality of the facilities as perceived by the teacher. Older participants had higher achievement scores, better quality facilities were associated with higher achievement and so was the fact that teachers who had received some in-service course on sewing in the last five years had participants who scored higher on the end-of-course achievement test. The latter two variables were clearly policy variables and it would seem that it would be advantageous to attempt to supply machines and materials of adequate quality and to give teachers special in-service courses.

At the same time, it is interesting to note that being in the 150 or 200 hour course did not produce significantly different achievement scores, other things being equal. Nor did the shift in which the participant was enrolled. The initial pre-service training of the teachers was not associated with achievement.

Table 11 Simple correlations, Beta coefficients, and R-squared for sewing achievements

	Block I		Block II	
	r	ß	r	ß
Father's occupation	-.16	*		*
Kits	.20	*		*
Participant's age	.30	.25		.15
Previous experience	.16	*		*
Teacher training			-.01	*
Additional training			.43	.34
Quality of facility			.16	.21
Facility			-.11	*
Shift			.15	*
Teacher age			.10	*
150/200 hour course			.15	*
R^2	.12		.28	

Note: N = 282, 2 se (ß) = 0.12.
ß not exceeding two standard errors are asterisked (*).

Table 12 Simple correlations, Beta coefficients, and R-squared for earning and saving money

	Money earned						Money saved					
	Block I		Block II		Block III		Block I		Block II		Block III	
	r	ß	r	ß	r	ß	r	ß	r	ß	r	ß
Father's occupation	.00	*		*		*	-.34	-.33		-.28		-.24
Participant's age	.00	*		*		*	.15	*		*		*
Objective	-.21	-.23		-.17		*	-.23	-.26		-.19		*
Kits	.18	.31		.34		.31	.07	*		*		*
Previous experience	-.09	-.24		-.25		-.26	.13	*		*		*
Teachers' training			.13	*		*			.38	.15		.19
Size of class			-.03	*		*			-.30	-.24		-.26
Shift			-.15	-.23		-.29			.20	*		*
150/200 hours			.17	-.26		-.30			-.13	-.15		-.19
Total sewing score					**.28**	**.35**					**.38**	**.36**
R^2	**.13**		**.21**		**.31**		**.20**		**.33**		**.44**	

Note: N = 199, 2 se (ß) = 0.14.
ß not exceeding two standard errors are asterisked (*).

Table 12 presents the results of two regression analyses, the first making 'money earned' as the criterion, and the second taking 'money saved' as the criterion. Five variables were significantly associated with money earned. Those were the possession of a sewing machine at home (as we would expect), previous experience in sewing (but this was negatively associated with earning, meaning that those who had no previous experience earned more after the course than those with previous experience), shift (those in the morning shift earned more), those in the 150 hour course earned more than those in the 200 hour course (in fact, being in the 150 hour course was worth 289 baht per month more than being in the 200 hour course), and the end-of-course achievement as measured by the test.

Size of class, level of teacher training, the reason for joining the class, age and father's occupation were not associated with money earned. Again, however, the regression accounted for only 31 per cent of the variance.

For money saved, five variables were important. Participants whose fathers engaged in agriculture and labour saved more. Participants whose teachers had higher levels of pre-service education, those in smaller classes in the 150 hour course, and again, gratifyingly, those with higher end-of-course achievement scores saved more. In this case, the regression accounted for 44 per cent of the variance.

What can we glean about the adult education non-formal programme in sewing?

Firstly, the sewing course itself was important because those with higher scores earned more and saved more, but we must remember that higher scores were primarily a function of participants' age, additional teacher training and the quality of the facilities.

Secondly, there was no advantage for anyone to have been in the 200 hour course as opposed to the 150 hour course. There would seem to be good reason to abandon the 200 hour course and provide additional training for teachers and improve the quality of facilities. Why size of class should be inversely related to 'money saved' (and shift to 'money earned') is not immediately clear. We suspect that the class-size might be so large that the teachers were not able to supervise them all. Therefore, after graduation from the course, they lacked confidence to sew even for themselves and their families. Regarding the shift and 'money earned', the morning shift students might have been those who were in the waiting period for jobs and wanted to take sewing seriously as a career (as opposed to evening shift enrollees who entered to take it as a hobby).

Typing: As mentioned earlier, the criterion measure was limited to time spent per week on typing six months after the end of the course. We were hesitant about the use of this criterion variable because we believed that it would contain substantial measurement error. For example, time spent in typing could be influenced either by a desire to practice typing, or perhaps the need to type personal and/or family papers and documents. It would therefore be likely that participants might spend a great deal of time typing in one week yet very little in the following week. The students' answers to the question might therefore have represented an 'average time' spent over the time since their course had finished or it might have represented an estimate of the amount of time they had spent in the few weeks before their interview.

Variables entered into the regression were arranged in two blocks, bearing in mind the same process by which we chose and grouped the variables for sewing analysis. In fact, the two block regression model was equivalent to the first two blocks used to examine money earned and money saved in sewing. The third block was omitted because the cognitive measure was unsuitable. Variables in Block I included father's occupation, participants' age, typing equipment at home (whether or not they or their family owned a typewriter) and

personal study objective (for a career, for credit or for self, family and others). Block II included teacher qualification, size of class, the shift attended (morning, afternoon, or evening), and finally the type of course participants enrolled in (150 or 200 hours).

The achieved sample size for these interviews was 116 participants (53.7 per cent of all typing). As a result, 0.19 was calculated as being about two standard errors for the beta coefficients in this analysis. The results of the analysis are presented in *Table 13*.

Table 13 Time spent typing per week

	Block I		Block II	
	r	ß	r	ß
Father's occupation	.07	*		*
Objective	.09	*		*
Participant's age	.19	.22		.22
Kits	.16	*		*
Teacher training			-.04	*
Size of class			-.13	*
Shift			-.14	*
150/200 hours			-.12	*
R^2	.09		.14	

Note: N + 116, 2 se (ß) = .19.
ß not exceeding two standard errors are asterisked (*).

The result of the study showed that simple correlations between predictors and the criterion were all below two standard errors, except for the variable 'objective' which barely reached significance. This variable was again the only variable which continued to be

significant in both blocks at the end of the analysis. This meant that typing students whose objectives were to take the course for miscellaneous purposes, e.g., for self and family, typed more than those who did not have this objective. But in the case of sewing, the result was reversed, that was participants in the sewing courses who used sewing skills six months after the course were those whose motivation to take the course had been that of making a career in sewing.

The total amount of variance predicted at both stages was rather low, reaching only 14 per cent when all predictors had been entered. In summary, at both the simple correlational level and at the multiple correlational level, we had very limited success in explaining the amount of time spent in typing by students after they had finished their courses.

It is extremely difficult to draw any policy recommendation from this section of the analysis because of the reasons outlined above. At best, we can make two suggestions: first, examine the personal study objective variable closely. We can suggest that typing courses should be opened at all three levels, that is basic typing, intermediate typing and advanced typing so that students in the more advanced course are those who want to continue to take typing for very obvious reasons, e.g. a career. This approach should, then, lead to a strategy to reduce drop-outs. Secondly, draw upon the experience we have gained to look more closely at the potential for extreme multidimensionality in tests of practical vocational skills such as typing. To the best of our knowledge, this field of research has received limited attention in our country. We believe it is a field which presents different problems than covered in the research available from western countries because of the substantially different structure of the language.

Conclusions and recommendations

The study had four main aims. The first was whether participants in the 200 hour course learned more than those in the 150 hour course. In sewing, the participants in the 200 hour course gained significantly higher scores on the end-of-course sewing test. In typing, there was no difference in scores between the 150 and 200 hour courses. Whereas we have confidence in our sewing test, we must express some concern about our typing test, so that the typing results must be interpreted with caution.

The second aim of the study was to identify those variables influencing the achievement of participants at the end of their non-formal education course. As can be seen from the detailed discussion earlier, in the study we failed to produce a good typing test and therefore present no multivariate analyses using this criterion. However, for sewing three variables were identified, two of which were important. The first was additional teacher training. Participants with those teachers who had received at least one in-service teacher training on sewing in the last five years obtained higher sewing scores than participants with teachers who had not attended such in-service courses. The second was the quality of facilities. The higher the teachers rated the quality of the facility, the higher the scores of the participants. Facilities included the sewing machines, zigzag machines and irons. The third was characteristic of the participants. Older participants tended to score slightly higher than younger ones. Of these three variables, it was 'additional teacher training' which was by far the most important.

The third and fourth aims were to discover whether participants in the study took up employment using their skills within six months of the end of the course. The definition of employment was whether participants earned money either from full or part-time work. Using this definition, only 3.5 per cent of those in the Thai only (short and long) took up employment. In the Thai plus English typing courses,

16 per cent from the 150 hour course and 7 per cent from the 200 course were employed. In sewing, 42 per cent became employed but with no difference between the short and long courses. The result for typing is disappointing, but participants were asked how much they typed, irrespective of whether this was for employment or not. A further 66 per cent from the Thai typing course indicated that they typed, but without reimbursement. From the Thai plus English course, a further 62 per cent from the short course and 53 per cent from the long course typed, but again, without reimbursement. In general, typists were typing below one hour per week after six months. Many of the participants in the typing courses were students in other adult continuing education courses held in the centres. These courses were for obtaining educational certification equivalent to full-time schooling. Fifty per cent of all participants in the Thai typing courses were students as were just over 50 per cent in the Thai plus English courses. It could not be expected that those who were still students six months after the end of the course would be employed. Only 42 per cent of the participants in the sewing classes were employed but another 50 per cent sewed for their families and their own use. Thus, the sewing course can be regarded as highly successful.

A further analysis was conducted to determine how much money the sewing participants earned and saved six months after their course, and which were the major determinants of earning and saving. Forty-two per cent were earning money varying in amount from 25 to 3,000 Baht per month. Five factors were important in determining money earned – the possession of a sewing machine at home, their achievement at the end of the course, which course they attended (those in the 150 hour course earned more), which shift they attended (the morning shift earned more) and their previous experience (those with experience in sewing before the course earned less). Money saved was also mainly determined by five factors, only two of which were the same as for money earned. These two were the course (150 hour course participants saved

more), and the score at the end of the course. The other three factors were being from farming families, being members of smaller classes at the centre, and having teachers with higher levels of pre-service training.

To summarize the findings for sewing it would appear that the important variables which affect the work of the centres are teacher training, both pre- and in-service, the quality of facilities, the shift (morning shift performed better) and the length of the course (150 hour participants performed better and earned and saved more). It would be unwise to comment on typing because it was impossible to evaluate it accurately partly because so many of the participants are still studying.

On the basis of our conclusions, we make the following suggestions to the Department of Non-formal Education at the Ministry of Education:

- Serious consideration should be given to abandoning the 200 hour sewing course and concentrating on improving the 150 hour course by ensuring better quality of facilities and providing in-service training for all teachers.

- Repeat the typing study with two changes incorporated. First construct psychometrically sound tests of typing and, secondly, make the tracer study cover a period of 18 months after the end of the course.

EXERCISE

Select two published articles from an educational research or educational evaluation journal that have the following general features.

- An article that describes the evaluation of a new textbook, teaching method, or curriculum reform by using an *experimental approach*, which includes 'treatment' and 'control' groups of students.
- An article that describes the use of a sample survey approach in which data are collected for the purposes of 'monitoring' and/or 'evaluating' student educational achievement and the general conditions of schooling.

Use the '*Checklist for Evaluating the Quality of Educational Research*' described in this module to examine the quality of the two articles.

References

Brickell, J.L. (1974). Nominated samples from public schools and statistical bias. In *American Educational Research Journal*. Vol. 11. No. 4, pp. 333-41.

Coleman, J.; Hoffer, T.; Kilgore, S. (1987). *Public and private schools*. Washington: National Center for Educational Statistics.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Ross, K.N. (1986). *Sample design options for a multi-purpose survey of villages in Indonesia*. Assignment report. Jakarta: Office of Educational and Cultural Research and Development, Ministry of Education and Culture.

Ross, K.N. (1987). Sample design. In *International Journal of Educational Research*. Vol. 11, No. 1, pp. 57-75.

Ross, K.N.; Mählck, L. (eds.). (1990). *Planning the quality of education*. UNESCO: International Institute for Educational Planning. Oxford: Pergamon Press.

Thongchua, V.; Phaholvech, N.; Jiratatprasoot, K. (1982). Adult Education Project – Thailand. *Evaluation in Education*. Vol. 6, pp. 53-81.

Tuckman, B.W. (1990). A proposal for improving the quality of published research. *Educational Researcher*. Vol. 19, No. 9, pp. 22-25.

Additional readings

Borg, W.; Gall, M. (1989). *Educational research: an introduction* (Fifth edition). New York: Longman.

Cook, T.D.; Campbell, D.T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Hedges, L.V.; Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Keeves, J.P. (1988). *Educational research, methodology, and measurement: an international handbook*. Oxford: Pergamon Press.

Kerlinger, F. (1986). *Foundations of behavioral research* (Third edition). New York: Holt, Rinehart and Winston.

Millman, J.; Goroin, D.B. (1974). *Appraising educational research: a case study approach*. Englewood Cliffs, NJ: Prentice-Hall.

Walberg, H.J.; Haertel, G.D. (1990). *The international encyclopedia of educational evaluation*. Oxford: Pergamon Press.

Wolf, R.M. (1990). *Evaluation in education* (Third edition). New York: Praeger.

Since 1992 UNESCO's International Institute for Educational Planning (IIEP) has been working with Ministries of Education in Southern and Eastern Africa in order to undertake integrated research and training activities that will expand opportunities for educational planners to gain the technical skills required for monitoring and evaluating the quality of basic education, and to generate information that can be used by decision-makers to plan and improve the quality of education. These activities have been conducted under the auspices of the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ).

Fifteen Ministries of Education are members of the SACMEQ Consortium: Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania (Mainland), Tanzania (Zanzibar), Uganda, Zambia, and Zimbabwe.

SACMEQ is officially registered as an Intergovernmental Organization and is governed by the SACMEQ Assembly of Ministers of Education.

In 2004 SACMEQ was awarded the prestigious Jan Amos Comenius Medal in recognition of its "outstanding achievements in the field of educational research, capacity building, and innovation".

These modules were prepared by IIEP staff and consultants to be used in training workshops presented for the National Research Coordinators who are responsible for SACMEQ's educational policy research programme. All modules may be found on two Internet Websites: **http://www.sacmeq.org** and **http://www.unesco.org/iiep**.