5-1968

Assessing Achievement on a First-Grade Economics Course of Study

A. Guy Larkins
Utah State University

ASSESSING ACHIEVEMENT ON A FIRST-GRADE ECONOMICS COURSE OF STUDY

by

A. Guy Larkins

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF EDUCATION

in

Curriculum and Supervision

Approved:

_____
Major Professor

_____
Head of Department

_____
Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

1968

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

ABSTRACT

Assessing Achievement on a First-Grade

Economics Course of Study

by

A. Guy Larkins, Doctor of Education

Utah State University, 1968

Major Professor: Dr. James P. Shaver
Department: Elementary Education

Problem

Despite the surge of interest in economic education in the elementary school in the last two decades, there have been very few attempts to assess the ability of young children to learn economic concepts. In the primary grades, this problem is compounded by the difficulty of measuring knowledge in six and seven year old children.

Objectives

The primary objective of this dissertation was to determine whether first-grade children can learn the basic concepts in Our Working World: Families at Work. Since instruments suitable for assessing achievement on Families at Work were not available when this study was initiated, a secondary objective was to develop adequate achievement tests.

Procedures

Four Primary Economics Tests for Grade One (PET-1) were developed: The YES-NO, Matched-Pairs, All-NO, and Picture tests. These four tests were compared for reliability and validity. Reliability of the Matched-Pairs, All-NO, and Picture tests was adequate for the major purposes of this study, such as comparing group means. However, the Picture test

lacked content validity in the sense that it was not comprehensive--it sampled only a few of the major concepts in Families at Work. And the All-NO test confounded acquiescence-set with knowledge of the content of Families at Work. It was concluded that the Matched-Pairs test had adequate reliability and validity for studies such as this one.

To determine if elementary students could learn the concepts in Families at Work, control and experimental groups of children were selected from one urban, one rural, and two suburban areas of northern Utah. An experimental group of children was also tested in Elkhart, Indiana--where Our Working World: Families at Work was developed under the direction of Lawrence Senesh. Children were given the PET-1 tests and a test of mental ability. In comparing PET-1 means, analysis of covariance was used to adjust for differences in mental ability between control and experimental groups. Chi-square was used in item analyses to determine whether the first-grade children learned individual concepts in Families at Work.

Conclusions

The investigations of pupil learning led to five conclusions:

1. There were general indications that first-grade children can learn the content of Families at Work. In each of four studies--two which were preliminary to this dissertation, and two which were central to this dissertation--PET-1 means for the experimental groups were significantly larger at the .01 level than for the control groups.

2. There were no major concepts in Families at Work which first-grade children did not learn. Each concept was learned by some students at at least a simple level of abstractness and complexity.

3. Families at Work was not too easy for bright first-grade children. Even very intelligent children failed to demonstrate complete mastery of the major concepts in Families at Work. No student obtained a perfect or near-perfect PET-1 score.

4. Families at Work was not too difficult for slow students. Slow students demonstrated that they learned some of the content of Families at Work. Those students in the experimental groups who were at least six months below grade-level obtained significantly (.01 level) higher PET-1 scores than did similar students in the control groups.

5. Special training or experience does not seem to be necessary in order for teachers to adequately instruct first-grade children in the content of Families at Work. PET-1 means for students in Elkhart, Indiana did not differ at the .05 level of significance from PET-1 means for the other experimental groups.

(202 pages)

CHAPTER I

STATEMENT OF THE PROBLEM, AND REVIEW OF THE LITERATURE IN ECONOMIC EDUCATION

Problem

Since the first workshop in economic education held at New York University in the summer of 1948, and the founding of the Joint Council on Economic Education the following year, reports of numerous content and opinion surveys, evaluation committees, curriculum projects, and general recommendations for economic education have appeared in the literature. For the most part, this surge of interest in economic education has centered on the secondary school. One significant exception has been the work of Lawrence Senesh in conjunction with the public schools of Elkhart, Indiana.

Senesh is convinced that the terminology and analytic concepts of economics can be taught in ways that are comprehensible to children in the earliest grades. Following this conviction, he has produced social studies programs for Grades One to Three based on economic and other social science concepts which were formerly believed to be too difficult for six-to-eight year old children. The general title for the Senesh materials is Our Working World. The courses of study for Grades One to Three are subtitled Families at Work, Neighbors at Work, and Cities at Work.

Despite the fact that Our Working World is based on the assumption that primary-grade children can learn the basic concepts of economics and other social sciences, an extensive review of the literature uncovered

no research which tested that assumption. This appears to be consistent with the general lack of interest in research of any kind concerning primary-grades social studies. Of five-hundred and sixty-six dissertations in social studies listed in McPhie's guide (1964), only twenty-one are clearly related to the primary grades and of these only twelve are clearly specific to the primary grades. Furthermore, an extensive review of the literature for this dissertation uncovered only one attempt to measure the ability of primary-grade children to learn economic concepts. This review included more than 200 journal articles and dissertations in economic education. The one study which attempted to measure the ability of primary-grade children to learn economic concepts (Robinson, 1963) was conducted prior to the publication of the first Senesh materials--Our Working World: Families at Work (1963), and therefore it did not attempt to measure learning of the specific concepts contained in that course of study.[1]

Given the lack of interest in research of any kind concerning primary-grades social studies, it is not surprising that while there are economics tests available for the secondary school, none has been published at the primary-grade level. The test Robinson developed does not fill this gap: (1) The reliability of her instrument is too low--less than a coefficient of .50, and (2) it is not readily reproduceable.

The lack of assessment of the ability of young children to learn economic concepts in general and the concepts included in Our Working World in particular cannot be justified on the grounds that few people

---

[1]Robinson's study is reviewed in greater detail later in this chapter.

would be interested in the results of such a study. The Senesh materials have been published by a major educational publishing house--Science Research Associates. These materials have also received considerable notice in the literature--see, for instance, the September through June issues of The Instructor for 1964-65. Furthermore, Our Working World is apparently being adopted by a number of school districts, including three of the largest in Utah--Salt Lake City, Weber County, and Davis County. Therefore, development of a primary-grades economics achievement test which is based on the Senesh materials, and investigation of learning due to instruction with the Senesh program, could make a practical contribution to primary-grades education.

Although achievement tests need to be developed and assessment of learning needs to be conducted for the Senesh materials at each of the first three grades, this dissertation is limited to the first grade--Our Working World: Families at Work. The decision to restrict test development and learning assessment to one grade level was based on experience gained through an earlier study by Shaver and Larkins (1966). In that study an attempt was made to remedy both the lack of a suitable test and the lack of evidence of ability to learn economics in the first grade. A paper-and-pencil achievement test[2] based on Families at Work was developed and administered to a sample of control and experimental classes in the Salt Lake City School District in May, 1966. Although, as expected, the mean scores of control and experimental groups were different at the .01 level of significance, the results of that study

---

[2]This test and subsequent tests developed for this dissertation are titled Primary Economics Tests: Grade One, abbreviated PET-1.

clearly indicated the need for further test development. First, the reliability of the initial PET-1 test was low--.28 for the control group, and .56 for the experimental group. Second, although the mean scores for the two groups differed at the .01 level of significance, very few individual items discriminated between control and experimental groups. This could be explained either on the grounds that non-discriminating test items were poorly constructed, or on the grounds that the experimental classes failed to learn several basic economic concepts included in the Senesh materials. If first-grade children fail to learn many of the concepts as they are taught in the Senesh materials, then: (1) Expectations of those who use the materials will need to be revised, (2) the teaching methods used in Families at Work will have to be revised, (3) the course content will have to be revised, or (4) some combination of revision would be in order.

If it is assumed that the PET-1 test items were not poorly constructed, and that the children in the experimental groups were ignorant of the content of many of the test items, it still does not follow that children cannot be taught the economic concepts in question. It might be that the Salt Lake City experimental classes did not represent an optimal learning situation for the Senesh program. The Salt Lake City experimental classes were probably less than optimal in at least three ways. First, the Shaver-Larkins study was conducted at the request of the Salt Lake City School District to fulfill the requirements of Title I of Public Law 89-10. The school district had purchased Our Working World: Families at Work with federal funds for use with "economically deprived" students, but the materials were not introduced into the curriculum of

the Salt Lake City schools until midway in the school year. As a result, teachers had already begun their ordinary social studies program and some were hesitant to drop what they had begun to take up something entirely new. Of course, since Families at Work was designed to be a full year's course of study, students could not be expected to learn all of the concepts in a half-year. Second, several persons involved in initiating experimental economics courses have commented on the importance of in-service teacher training in economics (Anonymous, 1964). The Salt Lake City first-grade teachers met in an orientation meeting which was designed to introduce them to the Senesh materials in one afternoon. That meeting is not likely to have met very stringent criteria for inservice training. Third, it is possible that new courses of study are better implemented by teachers who volunteer to try them than by teachers who have no choice in the matter. In Salt Lake City, the Senesh materials were introduced by administrative fiat. A fourth way in which the Salt Lake City experimental classes were less than optimal was in the nature of the population from which the sample was drawn. As previously mentioned, Families at Work was used only in those schools in neighborhoods which qualified under Title I of Public Law 89-10 as economically deprived. This does not necessarily mean that the students were less able to learn the content of Families at Work, especially since the Salt Lake City School District reduced the class load in most of these schools and introduced special programs to overcome some of the educational disadvantages which these children might have had. Nevertheless, the sample was not representative of most schools. According to the records of the school district, children in those schools which qualified as economically deprived have not done

as well in the past on standardized measures of ability and achievement as have children in the rest of the district.

The problem, then, is:

1. There are no adequate achievement tests for assessing learning of the concepts in Our Working World: Families at Work. No such tests have appeared in the published literature. Shaver and Larkins' PET-1 instrument is based on Families at Work, but it has low reliability.

2. There have been no adequate assessments of learning the content of Families at Work. The Shaver-Larkins study was not entirely adequate for several reasons already specified.

Therefore, the objectives of this study are:

1. To develop an adequate version of the Shaver-Larkins PET-1 test, and

2. To use that test to assess the ability of first-grade children to learn terms and concepts basic to Families at Work.

As stated on the first page of this chapter, an extensive review of the literature in economic education was conducted. That review is summarized below in order to sketch the general development of interest in economic education, particularly economic education in the elementary school, and to emphasize the almost total lack of interest in determining the ability of young children to learn economic concepts.

Overview of Economic Education

Although never a serious competitor of history or geography for rank in the social science curriculum, economics has long held a minor place in the public secondary schools of the United States (Cummings, 1950).

A college level course in political economy was offered in the academies from the early days of this nation until after the Civil War. At that time separate secondary school courses in political economy were developed (Prehn, 1965; Gilbreth, 1945). At the turn of the twentieth century, the term "economics" began replacing the older "political economy." Since then, the Great Depression and World War II have stimulated separate periods of interest in economic education, with the latter period of interest extending relatively unabated to the present (Merrifield, 1959).

A major landmark was the 1948 New York University Workshop on Economic Education, which led to the founding of the Joint Council on Economic Education the following summer. As of 1966, the Joint Council on Economic Education had forty-three affiliated state and regional councils, and though an impressive number of other organizations are interested in furthering the teaching of economics (McKee and Moulton, 1951), the Joint Council occupies a dominant position.

Economic Education in the Elementary School

Compared to the secondary school, economics has only recently appeared as a separate course of study in the elementary school. Gavian and Nanassy (1955), Knoble (1939), and Sloan (1943) mention research studies and curriculum development projects relating to the teaching of economics in the elementary grades as early as the 1930's, but there was no widespread interest in teaching economics to young children until after the Joint Council was founded.

Various authors, then and now, have held divergent views concerning the nature of economic education in the elementary school. The major

approaches can be divided into three categories: (1) applied economics, (2) economics as a structure of principles, and (3) economic topics.

Applied economics

Beginning in the late 1930's the Alfred P. Sloan Foundation supported a series of attempts to determine whether greater emphasis on "applied economics" in Grades One to Twelve would improve the living conditions of families in the economic fringe areas of our society (Sloan, 1943; Olson and Nutter, 1945; Seay, 1945). School children in the "backwoods" areas of Florida, Kentucky, and Vermont were given instruction in raising and preparing food, house construction, and clothing manufacture. These projects did not stress economics in the academic sense; economic concepts such as "producer," "consumer," and "division of labor" were not taught. A limited amount of research indicated that children in some of the Sloan projects made significant gains in mental age, and in diet and health practices (Goodykoontz, 1953).

A second approach--also classifiable as applied economics--centers around the performance of some business activity. Forming "little corporations" (Logan, 1946) and operating school stores in which children sell candy and small articles to their schoolmates (Eisen, 1958; Frisina, 1962; and Gavian, 1958) are typical examples. Brunson's (1966) plan to teach children "personal economics," consisting of problems in family finance, likewise fits this category.

Economics as a structure of principles

Knoble (1939), one of the first to champion teaching economics in the elementary school, bemoaned the fact that he was not taught a few

economic principles, that he was not given a pattern of economics, as a child. More recently, Lovenstein (1961), Coon (1966), Wing (1964), among others, have likewise argued for teaching structure rather than unrelated economic facts. Darrin (1960a, 1960b, 1960c, and 1961) developed outlines for courses of study based on the notion of economic structure, as did the Northwest Council for Economic Education (1966), and the Ohio State Economics Project (Lovenstein, et al., 1967).

The most ambitious project of this type to date is being directed by Lawrence Senesh, who claims

> ... economic understanding is founded upon a unified and logical system of ideas. It is acquired by learning economic relationships rather than by isolated economic activities as they are sometimes reproduced in the classroom. A game of grocery store . . . contributes little or nothing to economic education unless conceptual meanings are made clear (Senesh, 1966b).

His series, Our Working World (1963), is completed through Grade Three, and is intended to continue through Grade Twelve. Materials published at each grade level for the first three grades include a teacher resource book, a student text, a student workbook, and phonograph records which contain a story for each lesson. Filmstrips are also available for use in teacher training. Although based on economic concepts and problems, Families at Work also includes concepts drawn from other social science disciplines as they are relevant to important social issues.

Economic topics

Some contributors to the literature have been concerned with neither applied economics nor with teaching a structure of economics. The content analyses by Gavian and Nanassy (1955) of elementary-school courses of study are typical. They scrutinized the courses available for 1930-38

and for the 1940's and noted the occurrence of terms or phrases which were related to economics. This procedure generally results in a list of terms or topics which are related more by frequency of use than by logical pattern. Such lists, according to Senesh, do not constitute a structure or model of economics.

It is not uncommon to find suggestions for lessons or units in elementary-school economics which attempt to develop a topic or a series of related concepts and terms, but which give no indication that these suggestions are based on any comprehensive rationale concerning the nature of economics (Rohrbaugh and Haines, 1960, pp. 33-39; McCombs and Hohl, 1953; Barnes, 1953; Reed, 1958; and Delva, 1955).

Justifications for Teaching Economics

Justification for teaching economics has been as diverse as the differences of opinion concerning the proper approach to economic education in the elementary school.

The Depression and the Cold War

The Sloan projects mentioned earlier were admittedly motivated by the impact of the Depression, just as the more recent filmstrips sponsored by the Sloan Foundation were admittedly stimulated by the tensions of the Cold War (Zurcher, 1965). Garwood (1962, 1964), Bond and Roehr (1952), Melby (1950), and Senesh (1958) likewise have referred to the Cold War to justify teaching economics. At times, such justification is stated in extreme language. Perry (1960, p. 19) concludes his argument with, "Turn to the business teachers for help . . . these are the

people who are unhampered and unindoctrinated with alien social and political philosophies. . .ignorance is the soil in which foreignisms thrive."

Citizenship

Others have not seen fit to appeal to the danger of communism and socialism, but argue simply that citizenship in a democracy requires the ordinary man to make decisions concerning public policy, and that decisions often require some knowledge of economics (McPherson, 1948; Coleman, 1963; Wolfson, 1950). A position frequently taken by those who argue for teaching economics as an aid to decision-making is that there are no absolutes in economics, that economic problems are not settled once and for all. They claim it does little good to indoctrinate students with the "truth" about economic issues. Rather, our aim should be to give the student the means for analyzing problems and reaching defensible conclusions (Wolfson, 1950; Coleman, 1963; Uhr, 1963; Nourse, 1966). Senesh also holds to this position (Lagemann, 1964).

Personal adjustment

Another of Senesh's arguments for economic education in the primary grades is similar, but not identical, to the one above. In an interview published in School Management (Anonymous, 1964), Senesh claimed that young children desire to order their experiences, to arrive at a sense of reasonableness concerning a rather complicated world. Supposedly, discovering the principles of economics aids the accomplishment of this end--giving the child a sense of security. Decision-making in this case is not justified solely as an aspect of citizenship, but rather as an aid to personal adjustment.

Economic illiteracy

The most frequently cited argument for economic education is that our students and citizenry are "economically illiterate" (Pierrepont, 1948; Perry, 1960; Schultz, 1953; Bond and Roehr, 1952; and Eames, 1949). Several research studies have been published which conclude that Americans young and old lack economic understanding (Sewell, 1963; Saunders, 1966; Stoner, 1962; Wilde, 1954; Brown and Daily, 1961; and Madsen, 1961). However, it should be noted that economic illiteracy is to be regretted only if one or more of the other arguments for economic education are convincing. It makes little difference how ignorant we are if the object of our ignorance is unimportant to our needs or purposes.

A few dissenters

In closing this section, it should be noted that occasionally someone has the temerity to either question the wisdom of teaching economics to children, or to question the basis of all the alarm. Robbins (1955) doubts that high school students, much less six and seven year old children, are capable of understanding economics.

> I cannot get away from the feeling that economics is essentially a subject for grown-ups . . . at any rate if it is taught as anything like a theoretical system. No simple proposition in economics is likely to be true, unless it is understood as being subject to a whole complex of assumptions not likely to be read into it, save by those who have a sufficient knowledge both of the system of propositions as a whole and of the world of reality to which they have reference. Is it sensible to expect children to possess such knowledge? And if they do not, do we not run the risk of inculcating bad intellectual habits by trying to teach an economics so simplified as to be suitable for their understanding? (Robbins, 1955, p. 579)

Tonne (1955) simply states that, in his opinion, economics is being taught fairly well in both the secondary and elementary schools. He argues that economics is no different than any other subject. All subjects could be taught better, but there is no need for drastic revision.

Of more than two-hundred opinion and research articles reviewed on this subject, Robbins and Tonne were the only authors to question the advisability of increasing our efforts in economic education. To disregard their opposition out of hand, however, would leave us open to the charge of begging the question, since no one has produced anything like conclusive evidence that this nation is suffering from the effects of economic ignorance. It is difficult to demonstrate that, even if people are economically illiterate, they are functioning poorly in society, or that economic education would help them function more adequately. While these global questions are extremely difficult, if not impossible, to answer empirically, we are capable of ascertaining the ability of various types of students to learn economic concepts taught through different approaches. That is, we are capable of doing so if appropriate research projects are conducted.

Economic Knowledge Possessed by Various Groups of Children and Adults

We have seen that Senesh is among those who believe that "economic understanding is founded upon a unified and logical system of ideas" (Senesh, 1966, p. 34), and that economic education is an important ingredient in citizenship education, as well as a means towards personal adjustment (Anonymous, 1964). While we do not intend to test all of

Senesh's claims, these claims are at least partially dependent on the assumption that children can learn certain terms and concepts which are emphasized in Our Working World. It is, therefore, appropriate to review research concerned with economic education.

Economic knowledge of adolescents and adults

A summary of research concerning economic knowledge possessed by adolescents and adults is useful, first, to illustrate by contrast the lack of research in the elementary grades, and second, as a follow-up to our brief discussion of economic illiteracy. However, this summary will be limited to a brief overview and will not detail the research designs used by the various investigators. The reason for this brevity is that the studies reviewed have little in common with the research problems anticipated for this dissertation. The only point of contact is that those studies and this dissertation both deal with assessing economic knowledge. Differences in the subjects tested--first-grade children on the one hand and adolescents or adults on the other--require substantially different assessment instruments and research designs.

Several attempts have been made to measure the economic knowledge possessed by various segments of our population. Tests have been given to school teachers, their students, preachers, white collar workers, manual workers, and businessmen. In the judgment of the majority of the investigators, persons in all of these categories have generally been found wanting in economic knowledge (Brown and Daily, 1961; Saunders, 1966; Wilde, 1954; Eames, 1949; Reinbold, 1965; and Bircher, 1964).

Other studies related to economic education have included such diverse areas of interest as the ability of students to learn economics while typing (Clayton, 1966; and Cowling, 1966), the effects of industrial arts on consumer knowledge (Jacobson, 1964), knowledge of consumer economics among home economics students and teachers (Lemmon, 1962), consumer credit knowledge of high school seniors (Thompson, 1965), and economic knowledge of school superintendents (Howell, 1965).

Most of these studies are not reported in detail, and in many cases are only tangentially related to economic education in the public schools. Three exceptions are the investigations by Deitz (1963), Madsen (1961), and Sewell (1963). Deitz tested nearly four thousand high school seniors in California, Madsen tested sixteen hundred high school students in Utah, and Sewell's instrument was administered to nine hundred secondary school students in eight states. All three investigators concluded that the students they tested were deficient in economic understanding.

Of special interest was the manner in which Sewell and Madsen instructed students to respond to their tests. Both used basically two-option response forms. Students were to mark either AGREE or DISAGREE, or occasionally, DON'T KNOW, if they were in doubt. This response form is similar to the YES-NO form which is sometimes used with young children in that it is subject to acquiescence-set. That is, students who do not know the answer tend to respond YES or AGREE. For that reason, Cronbach (1942), has advised that the YES-NO, TRUE-FALSE, or AGREE-DISAGREE response form not be used. If it is used, he recommends that the items be so written that the correct response

is always NO, FALSE, or DISAGREE. Madsen recognized this problem (1961, p. 12) and apparently followed Cronbach's advice, since DISAGREE COMPLETELY is the correct response to twenty-three of twenty-nine items on the first part of his instrument. Students responding from acquiescence would miss these items.[3]

Economic knowledge of young children

Research related to economic knowledge of adolescents and adults has been spotty, but, by comparison, research related to economic knowledge of young children has been practically non-existent. The only study which assessed the economic knowledge of young children was Robinson's (1963) investigation of the ability of kindergarten children to learn economic concepts. During the Spring of 1962, twenty-four children in the kindergarten of the Agnes Russell School at Teacher's College, Columbia University were taught economic concepts based on a structure derived from the early writings of Senesh, and on recommendations of a national task force on economic education. Some of these concepts are also found in the course of study investigated in this dissertation--Our Working World: Families at Work--for instance, "producer," "producer of goods," "producer of services," and "economic interdependence." However, only a portion of the concepts found in Families at Work was included in Robinson's test, and it could not serve as an adequate sample of the content of Senesh's course of study.

---

[3]Acquiescence was a major problem in our attempts to develop a test for young children and will be discussed in greater detail later.

The kindergarten children at the Agnes Russell School and a control group of equal size were given a pretest and posttest consisting of pictures and objects to be sorted into categories which would demonstrate the concept being tested. The children were tested individually. After performing each of these non-verbal tasks, they were asked to define the concept being tested and explain why they sorted the pictures or objects as they did. Robinson's rationale was that children learn ideas on various levels, that it is possible to learn at a certain level without being able to verbalize the concept. Verbalization presumably is indicative of greater ability to conceptualize (Ibid., p. 124).

The reliability of Robinson's test was estimated by correlating two administrations of the pretest, separated by a two week interval. A correlation coefficient of .47 was obtained. This probably underestimates the reliability of the posttest. Pretest items were based on content which the subjects had not been taught, and so responses could hardly be other than random. Unfortunately, Robinson did not estimate reliability for the posttest. If it was as low as .47, the usefulness of her instrument is obviously limited.

In fairness to Robinson, it should be noted that she was well aware of the limitations of her study (Ibid., p. 18). Both her experimental and control groups should probably be classified as "educationally privileged." Parents of children in the experimental group were university faculty members and graduate students. Similarly, the control group was chosen from two expensive, private kindergartens in New York. Furthermore, the sample was small, an N of twenty-four for each group, and the study extended over only one semester, including pretest,

treatment, and posttest. Also, her sample was not randomly selected. All of these factors severely limit the extent to which the results of her study could be generalized to other groups. Robinson, therefore, quite rightly insisted on avoiding the term "experimental" and consistently referred to her investigation as "exploratory." Exploratory studies have their place, and we would be justified in viewing Robinson's conclusions as tentative suggestions regarding the ability of children to learn certain economic concepts.

In gross terms, Robinson obtained a mean difference significant beyond the .01 level between experimental and control groups (Ibid., p. 124). Of greater interest to our purposes is the response of students to certain concepts that are also stressed in Our Working World: Families at Work. For example, eleven of twenty-four children were able to distinguish between "customers"[4] and "producers," and were able to verbalize their reasons for doing so. Robinson concluded that the remaining thirteen children were not able to conceptualize these terms.

A second area common to Families at Work and the Robinson study concerns machines and their contribution to our economy. Robinson found that most children in the experimental group were able to name machines, and some could give incomplete explanations of why machines are useful; but she concluded that the concepts involved were too difficult for most of her students.

---

[4]Senesh uses the word "consumers" which is probably harder for children to understand than the more familiar "customers."

Since Robinson's findings must be viewed as tentative, and for our purposes certainly are inconclusive, it is unfortunate that only one other investigation claims to evaluate the ability of elementary-grade children to learn economics (Darrin, 1958). For our purposes, Darrin's study was inadequate. He claimed to measure the ability of children to learn economic concepts, but his measurement took the form of asking the teachers what their children learned. While such a method might give some insight into the ability of young children to benefit from instruction in economics, it seems better suited to measuring the reactions of teachers than the achievement of children.

Besides the Shaver-Larkins PET-1 test, mentioned earlier in this chapter, there have been two other attempts to produce achievement tests based on Our Working World. However, neither of these tests have been published--they have appeared only in developmental forms. In 1960, the Elkhart Public Schools produced a developmental version of an elementary economics test, but became discouraged with the problems of test development and later devoted their energy solely to developing curricular materials. From discussions with some of the teachers in Elkhart, and with Joseph Rueff, the Coordinator of the Social Science Research projects in Elkhart, it appears that they were unable to satisfy themselves with the validity of their tests.

A more recent attempt to develop an economics achievement test based on Our Working World is being conducted by the Social Science Teaching Institute of Michigan State University. This project apparently is not completed. Their test has not been published in the literature. Although attempts to correspond with the Michigan State project have

gone unanswered, copies of their test have been examined. It appears to be questionable in at least one respect. It is a multiple-choice picture test, but rather than hire an artist to produce new drawings, the test producers used illustrations from Our Working World. For that reason, it will always be questionable whether students are answering items on this test correctly because they have learned the concept, or because they remember the pictures from the teaching materials.

Conclusions from the review of the literature

Interest in promoting the teaching of economics has not been matched by attempts to assess the ability of young children to benefit from such instruction. For instance, the Our Working World course of study has received a good deal of publicity, but there have been no published reports evaluating achievement of children who have been exposed to these materials. Although the dearth of studies in elementary-grade economic education certainly justifies further inquiry, little information is provided that is useful in constructing a suitable test. Robinson's test is not reproducible since she did not provide copies of the pictures used, nor did she describe in detail the other objects in the test. Even if such information had been provided, the low estimate of reliability, plus the fact that her instrument was not based specifically on Families at Work, make it difficult to justify using her test. The only other study which claimed to measure children's knowledge of economics--Darrin (1958)--did not use an achievement test. Likewise it would be difficult to justify using those tests which have not been published. The test produced as part of the Elkhart project

did not satisfy those involved in its construction, and there is no data available on the validity or reliability of the Michigan State instrument. Furthermore, even the developmental edition of the latter instrument was not made available to us for examination until after testing for this dissertation was completed.

CHAPTER II

DEVELOPING THE PET-1 TESTS

The first objective of this study was to develop an achievement instrument based on Families at Work. This chapter is concerned with some of the problems related to that objective.

The most apparent task associated with the development of achievement tests is the selection of appropriate content for test items. Also important is the selection of a suitable test form. As we shall see, this is particularly true when attempting to assess the academic achievement of young children.

Selecting Suitable Test Forms

Written multiple-choice tests

While there are many problems associated with testing first-grade children, such as their limited attention span and inexperience with test procedures, their limited reading ability is fundamentally related to the selection of appropriate test forms. For instance, written multiple-choice items are commonly recommended for use in achievement tests (Wood, 1961; Nunnally, 1964), but such items are seldom used in any of the primary grades and are singularly inappropriate for use in the first-grade. Confounding reading ability with knowledge of item content would unnecessarily complicate the already difficult tast of assessing achievement.

A review was made of doctoral dissertations in elementary school social studies to determine which test forms were most popular for use in assessing the achievement of young children. In one of the dissertations reviewed, the investigator claimed to have successfully administered a sixty-item written multiple-choice test to children in the first four grades (Hensen, 1964). However, on inspection, it is highly improbable that first-grade children could read all of the items on her test, given the fact that some children have a great deal of difficulty learning to read at all in the first-grade. Although the items may have been read aloud while the children followed along, Hensen unfortunately did not report the conditions under which the test was given. Equally unfortunate is the fact that she reported only a combined reliability coefficient for all four grades.

Multiple-choice picture tests

A common approach to the problem of the limited reading ability of six-year-olds is to use picture-type multiple-choice tests. However, it is possible that not all concepts can be tested with equal ease through pictures. For instance, economic interdependence is one of those concepts which is difficult to represent in a single drawing. Despite this difficulty, the picture-type multiple-choice test is the first choice of test developers who have sufficient financial and personnel resources. It is used in nearly every group test of ability or achievement produced for use with young children. However, those who lack time, money, or artistic talent--classroom teachers or directors of small-scale research projects--find the production of such tests a formidable task. For

instance, a fifty-item, five-option multiple-choice test requires 250 drawings.

Individual interviews

Apparently, others who have attempted to assess the achievement of young children have also recognized the difficulty of testing students who cannot read, and have likewise rejected the use of picture-type multiple-choice tests as too difficult for development in small-scale research projects. The most popular test form in doctoral studies concerned with achievement in the elementary school has been the individual interview (Foster, 1965; Lowry, 1963; Parker, 1963; Spodek, 1962; and Stephens, 1964). Some of the more sophisticated studies combined object and picture sorts with individual interviews (Butler, 1965; Frombert, 1965; Goldstein, 1966; Hadley, 1964; Helfrich, 1963; Johansen, 1965; Robinson, 1963; and Rush, 1964). These object and picture sorts amounted to an individually administered multiple-choice test. The students were shown a series of pictures or a series of objects and were required to select the correct one in response to a question by the tester. The investigator was thus able to adhere to a multiple-choice format without confounding reading ability and knowledge of social studies concepts.

Individual interviews have at least one serious disadvantage--they usually require a considerable amount of time to administer. In the amount of time that it takes to interview one or two children, an entire class of children could be given a paper and pencil test. If used to assess learning in the major subjects in the primary grades, individual interviews would take more time than an elementary school teacher could

give.[1] If used to assess learning as part of a research project, the number of students in the sample would have to be kept small or a number of testers would have to be used. Additional interviewers were not available for this study and it was desirable to have a larger sample than could be interviewed by one investigator. Therefore, the individual interview was rejected as an adequate test form for the purposes of this dissertation.

YES-NO tests

Shaver and Larkins (1966) used a YES-NO test in order to overcome the problem of the limited reading ability of first-grade children. The YES-NO test is similar to a TRUE-FALSE test. The items are read to the student by the tester and the student responds by circling either YES or NO on his answer sheet. When used with young children, the YES-NO test has the added advantage of not requiring the child to remember and consider four or five options, as does the multiple-choice test. It is possible that with some young children, multiple-choice tests confound knowledge of the content being tested with the ability to concentrate on multiple options. Although the YES-NO test can be produced rather

---

[1]Occasionally, throughout this dissertation, reference will be made to whether a test form is suitable for teacher-made tests. While the problem of teacher-made tests is not strictly relevant to the topic of this dissertation, the possibility of developing a test form that could be used for research projects and by primary-grade teachers occurred to this investigator during the original Shaver-Larkins (1966) study. It was assumed that if a test form could be developed which could be used in small-scale research projects at the primary-grades level, it might also be adaptable for use by classroom teachers. Therefore, a peripheral concern in evaluating test forms for use in this study was whether they were also adequate for teacher-made tests.

quickly and requires a minimum of reading ability on the student's part, it also presents some difficulties. Not the least of these are low reliability and pronounced response-set (Barnes, 1962; Cronbach, 1942, 1946, and 1950).

YES-NO tests present the subject with only two options. Since a subject may respond randomly to such a test and still be correct half of the time, YES-NO tests tend to have low reliability unless most subjects are knowledgeable concerning the content of the test and respond correctly to most of the items. Difficulties of test interpretation in YES-NO tests are further compounded by acquiescence-set; i.e., the tendency of students to respond YES when in doubt. A minority of students may even exhibit the opposite of acquiescence and respond NO when in doubt. This latter response-set is called "dissent" or "dissent-set."

The dominance of acquiescence-set in a YES-NO type first-grade economics test was investigated by Shaver and Larkins (1966). Subsequent exploratory studies (Larkins and Shaver, 1967) supported the earlier findings. Frequency of correct response to items for which the correct response is YES (YES items) was 70-75 percent. Frequency of correct response to items for which the correct response is NO (NO items) was 40-45 percent. The theoretical frequency of correct response for both types of items is 50 percent if students respond randomly. Since most students exhibit acquiescence rather than dissent, the frequency of correct response to YES items is a spuriously high estimate of knowledge.

An example illustrating how acquiescence-set can affect the interpretation of test results if the tester is not aware of the problem occurred in the Shaver and Larkins (1966) study. Control and experimental groups, each containing approximately 100 first-grade children, were asked to respond to this statement, "A specialist is a man who learns to do one job very well." Ninety children in each group correctly responded YES. This frequency of correct response is clearly higher than expected by chance. We might, therefore, be tempted to conclude that children in both the control and experimental groups knew the concept being tested. However, when the same children were asked to respond to, "A specialist can do more things for himself than a person who has not specialized," approximately thirty-five children in each group correctly responded NO. This is clearly lower than the expected chance frequency of 50, and indicates that most of the children did not know the meaning of the word "specialist." Thus the result on the previous question was apparently contradicted.

In brief, interpretation of individual YES-NO test items is difficult since there is no way of determining what portion of the responses is due to acquiescence-dissent, and what portion is due to knowledge. Of course, interpretation of scores of individual students is always difficult when tests are not reliable. Split-half[2] reliability coefficients obtained by Shaver and Larkins (1966) were .56 in the experimental group and .28 in the control group.

---

[2]Corrected with the Spearman-Brown Prophecy Formula.

All-NO tests

Cronbach (1942) suggested that higher reliability could be obtained by writing YES-NO tests containing only NO items.[3] Since most people tend to acquiesce rather than dissent, a NO response would generally be made from knowledge. However, if it is true that some people are dissentient, they would be favored by All-NO tests. A person who tends to respond NO would have a spuriously[4] high score on such a test. If for that reason the validity of the All-NO test is impaired, it makes little difference whether it is a reliable instrument or not. Despite its low reliability, the YES-NO test may be more valid than the All-NO test if the effects of acquiescence-dissent can be removed from the student's total score. If YES-NO tests are written with equal numbers of YES and NO items, any advantage gained on the YES items by an acquiescent student will be counterbalanced in his total score by his tendency to miss the NO items. Similarly, an advantage gained on the NO items by a dissentient student will be counterbalanced in his total score by his tendency to miss the YES items. There is no such balancing effect in the All-NO test.

This approach--writing balanced tests with equal numbers of YES and NO items--was suggested by Couch and Keniston (1960) in their study of

---

[3]Referred to as "All-NO tests" in the rest of this paper.

[4]"Spurious" is used in this context to mean that the student's score is higher than it would be if the test did not confound knowledge and acquiescence. It is not used to mean that the student did not really obtain a given score. This is consistent with the way in which Garrett uses the term (1958, pp. 441-443). Under the heading "Spurious correlation" he says, "We have shown elsewhere how a lack of uniformity in age level may lead to correlations which are misleadingly high." If correlations can be termed "spurious" in the sense of being misleading, then the term "spurious" should also be applicable to scores or standard deviations which are misleading.

the effects of acquiescence on personality inventories. However, the reliability of YES-NO achievement tests written with equal numbers of YES and NO items, and given to first-grade children, is still low. Larkins and Shaver (1967) reported a reliability coefficient of .35 for a 30-item test written with equal numbers of YES and NO items.

Matched-Pairs scoring

A technique, which hereafter will be referred to as "Matched-Pairs scoring," was devised to cope with the validity problems arising out of acquiescence-dissent and the problem of low reliability in the YES-NO test. Matched-Pairs scoring involves writing reversed items for each concept or bit of information tested. "Reversed items" means that for every YES item there is a NO item intended to test the same content. For example:

CHILDREN WHO JUMP ROPE ARE PRODUCERS. (NO)

CHILDREN WHO WASH DISHES ARE PRODUCERS. (YES)

In Matched-Pairs scoring, the students are required to respond correctly to both forms of an item before credit is given for either. Therefore, if students are responding from acquiescence they will respond incorrectly to the NO items. If students are responding from dissent, they will respond incorrectly to the YES items. A correct response to both items indicates either knowledge or an occasional lucky guess.

Matched-Pairs scoring should increase the reliability of the YES-NO test by decreasing the probability of correct chance responses to any item. While the ordinary YES-NO test balances the effects of acquiescence-dissent in the student's total score, Matched-Pairs scoring should also balance the effects of acquiescence-dissent in the responses to

individual items. Matched-Pairs scoring should have the advantages, without the disadvantages, of both the ordinary YES-NO test and the All-NO test.

One drawback of the Matched-Pairs techinque is that it reduces the size of the test by half. A sixty item test is reduced to thirty items because pairs of items are scored as one. In order for this technique to be useful, the positive effect of increasing the options on each item from two to four must outweigh the negative effect of halving the length of the test.

Larkins and Shaver (1967) reported an exploratory investigation of the effects of Matched-Pairs scoring. A 30-item YES-NO economics test was given to six classes of first-grade children in November, 1966. Three classes were in the experimental group, and three in the control group. The content of the test was based on Our Working World: Families at Work. Tests were first corrected in the ordinary manner and again using Matched-Pairs scoring. Split-half reliability coefficients were computed for scores based on both techniques. It was hypothesized that reliability would increase when the Matched-Pairs method was used. Means and standard deviations were also computed, and the t-test was used to compare the achievement of control and experimental groups. The following table is reproduced from Larkins and Shaver (1967, p. 8).

Table 1. Split-half reliability: Comparison of control and experimental groups using ordinary and Matched-Pairs scoring

	r_{11} Ordinary	r_{11} Matched-Pairs
Experimental	.35	.60
Control	.14	.46

Expectations in regard to reliability were supported. Reliability for control and experimental groups increased using Matched-Pairs scoring. Under both scoring methods, reliability was greater for the experimental group. This was to be expected, since the control students had not studied the material upon which the test was based and were more likely to respond either randomly or from acquiescence-dissent. A reliability coefficient of .60 for the experimental group is probably as high as one might reasonably expect for a fifteen-item test.[5] However, estimated reliability--using the Spearman-Brown Prophecy Formula--for a test twice as long is .75. Interestingly, subsequent testing of the same group with an instrument containing twice as many items produced a split-half reliability coefficient of .75.

Larkins and Shaver also wanted to determine whether Matched-Pairs scoring increased the ability of the YES-NO test to discriminate between groups. They found that differences between means, standard deviations,

[5]The original 30 items were reduced to 15 when the Matched-Pairs scoring technique was used.

and the size of t-ratios increased when Matched-Pairs scoring was used, indicating that this particular YES-NO test discriminated between groups better when scored with Matched-Pairs.

Reversals for Matched-Pairs scoring

Aside from the Larkins and Shaver study (1967), a review of the literature on acquiescence did not uncover any attempts to write reversals for achievement tests. There have been several attempts to write reversed items for personality inventories (Mogar, 1960; Chapman and Campbell, 1957; Bass, 1955; Leavit, Hax, and Roche, 1955; Rokeach, 1963, Christie, Havel, and Seidenberg, 1958; Peabody, 1961; Rorer, 1963; and Ong, 1963), but differences between writing items to test knowledge and writing items to measure personality traits limit the value of these discussions for this project.

A topic common to several of the above studies is the question of whether an intended reversal actually functions as a reversal. This topic can be used to illustrate one of the basic differences between writing reversals for personality inventories and writing reversals for achievement tests. It is fairly standard procedure to test the reversibility of items on personality inventories by correlating responses between original and reversed items. If the attempt to write reversals is successful the correlation should be negative, because a subject who responds YES on the original should respond NO on the reversal and vice versa. However, there are no right or wrong answers on a personality inventory. Every subject is assumed to "know" the answer to any item. The "correct" answer is whatever he believes it to be.

This line of reasoning cannot be applied to achievement tests. Perfect negative correlation between original and reversed items on an achievement test indicates that the subjects are completely knowledgeable. A completely knowledgeable student will respond YES to one half of a reversed pair of items and NO to the other. However, students are seldom, if ever, completely knowledgeable. Either a positive correlation or no correlation between original and reversed items indicates some ignorance. Even a low negative correlation tells the investigator very little. It may mean that the attempt to reverse items was successful but that the effects of knowledge are being confounded with acquiescence, or it may simply mean that the attempt to reverse items was only partly successful. It might even mean that the students were only partly knowledgeable of the content of the test. Confounding measurement of knowledge and reversability of items makes any single interpretation of these correlations questionable. In preliminary studies for this dissertation, correlation coefficients were computed between responses to original and reversed items. Generally a low negative coefficient was obtained, but for the reasons just stated a clear-cut interpretation of the findings was impossible.

Larkins and Shaver (1967) reported that at least one other technique, sometimes used to produce reversals for personality inventories, is inappropriate for YES-NO achievement tests for young children. Ong (1963) is one of the few researchers to unequivocally claim success in writing reversals for a personality inventory. He produced most of his reversals by including a negative qualifier in the original item; some form of the word "no" was placed in the original statement. When Larkins

and Shaver tried this they found that "no" confused first-grade children and caused them to answer NO when they meant YES. For instance, suppose that the tester is wearing a blue shirt. He instructs the child to respond YES or NO to whatever he says about his shirt. He then says, "My shirt is not red." The correct response is YES, indicating agreement with the statement. However, the child will frequently indicate agreement with the statement by saying, "No. Your shirt is not red." On the YES-NO answer booklet he then marks NO. This response spuriously indicates that the child did not know the color of the shirt. That this occurs when "no" is used in YES items is well established, and it may also occur when "no" is used in NO items. Informal trials with adults indicated that they are also confused by the insertion of "no" into otherwise straightforward questions. Therefore, the applicability of Ong's findings to achievement testing is questionable and may even need reevaluation for personality testing.

Despite the concern of researchers in developing personality tests, the problem of reversibility may not be serious with achievement tests. It is reasonable to require a person to demonstrate knowledge of a particular concept by correctly responding to a number of similar, though not identical, items. Thus, the problem of the validity of reversed items is in kind no different than the problem of content validity faced when producing any achievement test. Of course, it is important to be aware of possible ambiguity in reversed items. But, every test writer must guard against ambiguity.

Summary of the problem of selecting suitable test forms

Two objectives were stated at the conclusion of Chapter I. The first was to develop an achievement instrument based on Families at Work. This objective was then divided into two tasks: (1) the selection of suitable test forms, and (2) the selection of suitable test content. The preceding sections of the present chapter have been devoted to a discussion of some of the problems encountered in selecting test forms.

a. Written multiple-choice tests confound reading ability with knowledge of the test content. They are very seldom used to assess learning in young children.

b. The multiple-choice picture-type test does not confound reading ability with knowledge of the test content. It also produces adequate reliability because it utilizes four or five options in a single item. However, picture tests are probably limited in content validity. It is difficult, if not impossible, to adequately express complex concepts in a single small picture. Even those concepts which are easily tested with pictures require more time and talent for test construction than is possessed by most teachers and graduate students.

c. Individual interviews are the most popular test form for research carried out for doctoral dissertations in elementary education. Like the multiple-choice picture test, interviews do not confound reading ability with knowledge of the test content. However, the time required to conduct individual interviews severely limits their practicability.

d. The YES-NO test scored in the ordinary manner is unacceptable. This test form is unreliable and is of dubious value even in comparing

group means. When constructed with reversals, and scored using Matched-Pairs, the reliability of the YES-NO test is improved. Exploratory attempts to improve the reliability of the YES-NO test by Matched-Pairs scoring failed to produce coefficients acceptable for differentiating between individual students--.85 or .90. But it did produce coefficients acceptable for comparing means--.60 or higher. Furthermore, the Matched-Pairs test requires more time to construct and score than do the ordinary YES-NO tests or the All-NO tests.

e. The All-NO test is easy to construct and score, and reportedly is more reliable than the YES-NO test. However, since Matched-Pairs scoring of YES-NO tests was developed rather recently, there is no comparative reliability data on it and the All-NO test. Furthermore, there is reason to believe that the All-NO test produces scores which are invalid for comparing individual students.

Final selection of test forms

The problem of the suitability of test forms to be used in assessing achievement in young children was not decisively settled by either the review of literature or the preliminary investigations by Shaver and Larkins. In particular, a final decision was not made concerning the merits of the YES-NO Matched-Pairs test form and the All-NO test form. Therefore, it was decided to use both forms for some of the final testing.

Because of lack of funds, it was decided in the beginning not to produce a multiple-choice picture test. However, a limited amount of money became available in March, 1967, for hiring an artist. Although

the time for final testing was drawing near, the production of a multiple-choice picture test was undertaken. An artist worked two to three hours a day for the next two months and a limited version of the PET-1 test was ready in picture form the day before final testing began. Unlike the other tests, this instrument had not undergone extensive revision, nor was it as comprehensive--it did not test as many concepts. Nevertheless, the investigator believed it was important to have even a limited opportunity to compare the picture test with the other test forms. For instance, recommendations as to which test form to use should include such practical considerations as ease of administration and scoring. The investigator had no idea as to how the multiple-choice picture test compared with the YES-NO, Matched-Pairs, or All-NO tests in this regard when given to young children. Therefore, it was included as part of the test schedule.

Three tests were used, then, for at least part of the assessment reported in this dissertation: (1) The YES-NO Matched-Pairs test, (2) the All-NO test, and (3) the Picture test. Since the YES-NO Matched-Pairs test can be scored in either the ordinary manner or with matched-pairs, four sets of scores were available. In practice, this was equivalent to having four sets of tests, and at times during this paper there will be reason for reporting findings as though there were four separate tests.

Selecting Test Content

Selection of appropriate content for the PET-1 tests began in January, 1966. From January until May, the present writer spent two

hours a day analyzing the content of Families at Work and writing tentative test items. During this period, concern for the content validity of test items took two forms: (1) It was recognized that a test of reasonable length could not survey all of the concepts in Families at Work. Therefore, an attempt was made to determine which concepts occurred most frequently in the teaching materials. (2) Care was also taken that the content of test items faithfully reflected the manner in which concepts were presented in the teaching materials. This latter concern likewise took two forms: (a) Care was taken that definitions of technical terms given in test items paralleled the definitions given in Families at Work, and (b) care was taken that, aside from the technical terms being used, the vocabulary of test items did not exceed the speaking vocabulary of first-grade children. With the attempt to write test items whose content paralleled the concepts taught in Families at Work, it might be assumed that the vocabulary level of those items would also be appropriate for first-graders. However, recognizing that it is easy for adults to miscalculate the abilities of young children, test items were submitted to experienced first-grade teachers who evaluated them and offered suggestions concerning appropriate wording.

In order to determine the frequency with which various concepts appeared, the content of the teaching materials used in Families at Work--the teacher's manual, the student's text, the student's workbook, and the record albums--were analyzed and compared. A tally was made of the number of times each concept was mentioned in any of these teaching materials. Those concepts which appeared most frequently in the teaching materials were included in the test.

Besides providing a basis for determining which items were mentioned most frequently in the teaching materials, the content analysis of Families at Work acquainted the investigator with the manner in which concepts were taught. This first-hand acquaintance with the content of Families at Work provided the general basis for determining whether the statement of concepts in test items was similar to the statement of concepts in the teaching materials. Of course, reference to the teaching materials was made whenever questions arose as to whether concepts were stated properly in test items.

While the content analysis was being conducted, approximately 250 tentative test items were written. Shaver and Larkins thoroughly reviewed each of these items in terms of the criteria previously mentioned. After extensive revisions, 60 items were selected for inclusion in the first Shaver-Larkins PET-1 test.

Preliminary tryouts of this test were conducted at the Edith Bowen Laboratory School at Utah State University, and at the Plain City and the Wilson Lane elementary schools, both of which are in Weber County, Utah. Additional revisions in the test were then made, based on the tryouts and the recommendations by the cooperating teachers. In the latter part of May, the instrument was administered to control and experimental groups of first-grade children in Salt Lake City. Findings were summarized in the Shaver-Larkins report (1966).

Between May, 1966 and May, 1967, several different versions of the PET-1 test were produced. Each of these was essentially a revision of the original test. For each revision the same basic criteria were used to select items as in the original instrument. In addition, an item

analysis was performed after each of two preliminary tryouts which included control and experimental groups. The two major tryouts of the Matched-Pairs test were conducted in September and November, 1966. The latter served as the basis for the Larkins-Shaver (1967) report. Less structured trials were held in the Winter of 1966-67, including a very limited tryout of an All-NO test.

The item analyses, performed after the September and November, 1966 trials, were conducted in the following manner. The frequency of correct response on each item by the control group was compared to the frequency of correct response by the experimental group. Chi-square contingency tables were used to test whether these frequencies differed significantly. Items which discriminated between control and experimental groups were retained in revised versions of the PET-1 test. Some items which did not discriminate were rewritten to remove vagueness or ambiguity that might be confusing to the children. Some non-discriminating items were also discarded in favor of similar items which did discriminate. However, several items were retained even though they did not discriminate between control and experimental groups. These items tested concepts central to the Senesh program, and their omission would have weakened the test's content validity.

CHAPTER III

TWO INVESTIGATIONS: DESIGN AND PROCEDURES

Rationale for Having Two Investigations

More than a year had been spent trying to develop a test form which would be suitable for assessing learning in young children. There was reason to believe that either the Matched-Pairs test or the All-NO test was adequate for both the needs of the educational researcher working with limited funds and for the needs of the classroom teacher. However, although tryouts of both the Matched-Pairs and All-NO test had been held, a direct comparison between the All-NO test and the Matched-Pairs test had not been made prior to the final testing for this dissertation. Neither had the investigator had the opportunity to develop, administer and evaluate a multiple-choice picture test. It was therefore desirable to administer all three tests under similar circumstances so that they could be compared for reliability and validity. On the other hand, the substantive issue of this dissertation is whether first-grade children can learn the basic concepts in Families at Work. In order to adequately treat the substantive issue, and also compare the various test forms, it was necessary to design two studies.

Both test development and assessment of learning could not adequately be handled in a single study. In the first place, random selection of students was considered vital to investigating the

substantive issue--assessing students' knowledge. However, random selection of students was considered impractical if all three PET-1 tests were administered. Since it was considered important to obtain an estimate of the students' mental ability, children in the random samples would be tested on four different days--one day for the mental abilities test, and three days for the PET-1 tests. Thus, a number of the first-grade classes in at least two school districts would be disturbed on four separate occasions. Furthermore, on each of these occasions it would be necessary for the school to make special arrangements for a room to be available in which the testing could be conducted. Since most of the elementary schools did not have extra rooms for that purpose, it meant that someone in each school would have to be inconvenienced on each of the four test days. It was the opinion of the investigator that such an imposition would strain the hospitality of the cooperating school districts, especially when a number of schools would have to be involved in each district. If all three PET-1 tests had been administered to randomly selected students, it would also have required hiring testers for two additional days. Funds were not available to cover this additional expense.

Therefore, it was decided to design one study which would employ random selection of students, but which would use only one of the PET-1 tests. It was also decided to design a second study which would allow comparison of all of the PET-1 tests, but which would not include all of the design features considered desirable for investigating the substantive issue. However, in the second study, attention could be focused on those design features which were considered vital to comparing

the reliability and validity of test forms--for instance, counterbalancing the order in which the tests were administered.

The first investigation--employing random selection of students--is referred to throughout this paper as the WOBE study, the second is called the EPC study. WOBE and EPC stand for the cooperating schools or school districts. In the WOBE study, the Weber County School District (Utah) provided the experimental group, and the Ogden City School Disttrict (Utah) and the Box Elder County School District (Utah) provided the control groups. In the EPC study, the Elkhart Public Schools (Indiana) and the Pioneer School (Weber County, Utah) provided the experimental groups, and the Cache County School District (Utah) provided the control group.

The WOBE Study

Description of subjects

Students for the WOBE study were selected from three adjacent school districts in northern Utah. Students in the experimental group of the WOBE study were selected from seven elementary schools in District W. The control group was composed of students from three schools in District O, and four schools in District BE.

Districts W and O are the county and city school districts in the second most populous area in the state. The boundaries of District O are conterminous with the city, which contained 70,197 people in 1960 (U.S. Bureau of the Census, 1961). The boundaries of District W are the same as for the county, excluding City O, and include several suburbs contiguous to the city. County W, minus City O, contained 40,547 people in 1960.

All seven elementary schools selected in District W are located in communities which are outgrowths of City O, and are suburban rather than rural. Of the three schools selected from District O, two are in areas similar to the county suburbs. The third school is near the center of town in what appears to be an upper lower-class area, and has a fairly high pupil turnover rate. Of the fourteen schools used, this is the only one located in an area distinctly different than the others.

The four schools selected in District BE are located in City B, which had a population of 11,728 in 1960. Although City B is not a suburb of a larger city, it has characteristics of both a small town and a suburb. A few years ago, a defense industry established a plant nearby. The resulting influx of people, with the attendant growth in house construction, modified B's rural, small-town character.

In short, schools of similar size located in suburban areas were selected for the WOBE study. Five of the seven schools in the control group contained three first-grades; the others contained two. Four of the seven schools in the experimental group contained three first-grades; three schools contained four. It appeared, then, that with the exception of one school in District O, the schools selected were reasonably similar.

Description of the measures used

Students in the WOBE investigation were tested with two instruments: (1) A PET-1: YES-NO Matched-Pairs test,[1] and (2) Form A:

---

[1]YES-NO Matched-Pairs and All-NO tests are in Appendices A and C.

Grades K-2 of Tests of General Ability (Flanagan, 1960). The Tests of General Ability (TOGA), used to provide an estimate of the students' mental ability, was chosen because of past favorable experience. In prior usage, no special difficulties were experienced in administering this instrument. It can be given in a reasonable amount of time, is easy to score, and is reported in the test manual to produce reliability coefficients of .85 to .95.[2]

Some difficulty was experienced in deciding whether to use an All-NO test or a YES-NO Matched-Pairs test. Working with versions of YES-NO and Matched-Pairs tests for over a year had produced familiarity with their weaknesses and strengths. This was not true of the All-NO test. One All-NO test, based on the Senesh materials, had been written and administered to two classes of first-grade children. From that limited experience, it was believed that the All-NO test was more reliable than the YES-NO test scored with Matched-Pairs. There was, therefore, a tendency to favor it. However, in the end it was decided to use the YES-NO Matched-Pairs test because of its anticipated greater validity. As stated previously, however, the invalidity of the All-NO test had never been demonstrated, only suspected.

The YES-NO Matched-Pairs test contains 75 items which sample terms from the first 24 lessons of Families at Work. All but one of the 75 items were written with reversals. Although Item 75 has no reversal, it was added to the test because the student response sheet had room for 15 responses on each page. It was easier to add an extra

---

[2]Obtained split-half reliability coefficients for the TOGA administered in this study are reported in Chapter IV of this paper.

item than to explain an empty space to 200 curious first-graders. Items were not written for lessons beyond Number 24. First-grade school teachers indicated that some classes would not have studied beyond that point at the time of testing. Not all school districts ended the year during the same week and it was important to insure comparable data by testing students on material they had all covered.

The test was deliberately written in mirror-image halves with Item 1 reversed in Item 38, and Item 2 reversed in Item 39, and so on, because this facilitated scoring. Scoring items in pairs is cumbersome if reversals cannot be located quickly.[3] The order of Items 1 through 37 was determined randomly, and since the second half of the test was a mirror image of the first, the order of Items 1 through 37 determined the order of the rest.

The student response sheet contained five pages with fifteen response spaces per page. Each response space contained the number of the item and the words YES and NO. Students responded by circling the appropriate word. Even when used at the beginning of the school year, there were few indications that students were unable to distinguish YES from NO. However, students needed a few minutes of practice in following the serial order of items.

Research design and procedures

The WOBE investigation used both partial matching and random selection. Partial matching was used in that the schools selected were approximately equal in size, and were located in suburban communities.

---

[3]This scoring procedure is explained in detail in Appendix B.

The main concern was to avoid the small rural schools in some districts, and schools in disadvantaged neighborhoods in other districts. Not all of the three districts contained small rural schools, nor did all three districts contain schools in economically or culturally disadvantaged areas. Inclusion of all schools in each district could have resulted in unlike samples.

Random selection was used in that students were randomly selected from each first-grade class in each of the fourteen schools. In schools with three first-grades, five students were randomly selected from within each class. In schools with two first-grades or four first-grades, eight or four students were selected from each room. In this way, possible positive or negative effects of a particular teacher or class were spread over twenty-four classes in the experimental group and nineteen classes in the control group. This approach also facilitated testing. It was known in advance that testers would be working with groups approximately equal in size in each school. Had selection been random over an entire district it is possible that testers would have worked with groups considerably different in size from school to school.

Only posttests were used. Pretests were not given for two reasons: (1) There were no published tests available for Families at Work. (2) At the time pretests were needed--Fall, 1966--tests being constructed for this study had not yet been developed to a suitable level of reliability. However, on the chance that the partial random selection might have produced groups differing in mental ability, TOGA's were given with the intention of using the raw scores as the covariate in analysis of covariance, if needed.

Both the TOGA and the YES-NO Matched-Pairs test were administered by seven undergraduate students majoring in elementary education at Utah State University. Each student participated in two half-hour training sessions prior to administering each test. Before giving the TOGA, testers were told to adhere to the instructions in the test manual. In regard to the YES-NO test, testers were instructed to: (1) Pace themselves so as to finish in approximately 35-40 minutes, (2) give the students frequent encouragement, and (3) arrange seating to minimize students' opportunities to seek help from one another.

The TOGA was given to the control and experimental groups on Tuesday, April 11, 1967. The YES-NO Matched-Pairs test was given one month later--Tuesday, May 9, 1967. Each of the seven testers gave one test in the morning and one in the afternoon on each test day. If a tester worked with a class in the experimental group in the morning, he or she worked with a class in the control group in the afternoon.

Students were randomly selected by the testers immediately prior to administration of the TOGA. Each tester was supplied with a list of numbers selected by the investigator from a table of random numbers. Upon entering each classroom the tester numbered the students, starting with the student nearest the door. He then selected those students who numbers appeared on his list. It was anticipated that teachers would attempt to assist the testers in this task. Testers were instructed to ignore the teacher's advice and adhere to random selection. They reported they were able to do so.

Some students who were given the TOGA were not in school when the YES-NO Matched-Pairs test was given one month later. These students--25

of 221--were dropped from the study. Despite this, TOGA means for the experimental and control groups did not significantly differ at the .05 level, indicating that random selection was successful in producing groups with only chance differences in ability. The obtained F-ratio was 1.37 compared to 3.89 needed for significance.

The EPC Study

Description of subjects

Subjects in the EPC study were drawn from three school districts--two in Utah, and one in Elkhart, Indiana.[4] Elkhart is a small industrial city, and had a population of approximately 40,000 people in 1960 (U.S. Bureau of the Census, 1961). Most of the working population is employed in one of several small industries, such as the manufacture of musical instruments or mobile homes. The three classes tested were located in two schools in lower middle-class neighborhoods. Though not old, both schools appear to have been used for some time. Homes in the neighborhood are modest and for the most part appear to be at least ten years old.

The three classes in the second group were located in School P of District W, mentioned in the WOBE study. P is a new school located in a semi-rural area. This area is termed "semi-rural" because School P is surrounded by farm land, but new houses are filling in the open spaces, and less than one percent of the families are engaged in full-time farming, according to school officials. School P was built to accommodate educational innovations such as modular scheduling and team teaching. The

---

[4]Support from the Utah State University Research Council made it possible to travel to Elkhart to carry out testing for the EPC study.

teachers were placed in this school because of their stated willingness to innovate.

The third group of classes was selected from District C in northern Utah. Each class was located in a separate school, one of which is a new building on the outskirts--almost the suburbs--of a small city. The other two are older schools located in rural towns. C is the most clearly rural of any district in either study. However, it is similar to the area surrounding School P in that families engaged in full time agriculture are a distinct minority. Furthermore, it is not an isolated area. One of Utah's two state universities is located in the small city in the center of the county. Many of the fathers commute to work in defense industries located 40 to 50 miles away.

Description of the measures used

Each student in the EPC study was tested with four instruments: The TOGA, the YES-NO Matched-Pairs test, the All-NO test, and the Picture test. The TOGA and the YES-NO Matched-Pairs tests have already been discussed in connection with the WOBE study. The All-NO test contains 74 items, the correct response to each of which is NO. Eighteen of the NO items on the YES-NO Matched-Pairs test are also included on the All-NO test. The content for the remaining items was selected from each of the first 24 lessons in Families at Work. An item was included for some concept central to each lesson, one lesson at a time, in rotation. The All-NO and YES-NO Matched-Pairs tests are comprehensive in that an attempt was made to sample the content of the first 24 lessons. In contrast, the content of the Picture test is limited. However, the Picture test is

useful because it contains several concepts which are basic to the Senesh program, such as "producer," "consumer," "specialization," and "division of labor." The Picture test contains 49 items, most of which have five options. The original intention was to have 50 items, but one was deleted because the picture was inadequate.

Research design and procedures

Students in the EPC study were not selected randomly. Supervisory personnel in each district were asked to recommend three first-grade classes taught by outstanding teachers. In requesting cooperation from the school districts, teacher ability as the criterion of selection, not student ability, was emphasized. Aside from children absent on test days, all of the students in the nine classes were included in the study. This deliberately biased sample was chosen because it allowed for a comparison of optimal, average, and minimal learning environments.[5]

Each class in each group was given one test per day for three consecutive days. Table 2 describes the rotation of tests. This rotation distributed the effects of time of day and day of week over all three tests. When reading the table, remember that the fourth test--Matched-Pairs--is identical in content to the YES-NO test, only the scoring procedure is different.

---

[5]Discussed later in this chapter.

Table 2. Rotation of tests in the EPC study

	9:00 a.m.	10:00 a.m.	1:00 p.m.
Monday	Class #1 Picture test	Class #2 YES-NO test	Class #3 All-NO test
Tuesday	Class #2 All-NO test	Class #3 Picture test	Class #1 YES-NO test
Wednesday	Class #3 YES-NO test	Class #1 All-NO test	Class #2 Picture test

Besides allowing for comparisons of test forms, the EPC study was designed for a secondary purpose--the comparison of mean scores between minimal, average, and optimal learning environments for Families at Work. Since the Senesh materials are being developed with the cooperation of the Elkhart schools, it was assumed that teachers in that district would be well qualified to teach the program. Therefore, Elkhart was taken to be an optimal learning environment. In contrast, District W seemed to be typical of many others which might adopt Families at Work. Teachers in this district had used the materials for part of a year prior to 1966-67, so they were not teaching something completely unfamiliar to them, but they received no special training in economic education. District W was thus taken to be an average learning environment. Families at Work had not been used in District C. It therefore was taken to be a minimal learning environment.

The use of three groups is justified on the following grounds: (1) In both studies by Shaver and Larkins, the experimental groups scored

significantly higher than the control groups. It was anticipated that the results of the WOBE investigation would likewise favor the experimental group. In regard to the comparison of group means, then, there was little reason to simply repeat a similar design in the EPC study. (2) Item analyses in both the 1966 and 1967 Shaver-Larkins studies indicated that a minority of individual items discriminated between control and experimental groups, even when differences between groups of items were significant. One plausable explanation, other than that the test was generally ineffective, was that students in the experimental groups had not learned the content upon which the non-discriminating items were based. Assuming that this was so, the question arose whether students might learn that content if better taught. If the concepts were not learned by students in an optimal environment, there would be cause to question the likelihood of them being learned under average conditions. (3) It was necessary to include Group C--the minimal learning environment--in the EPC study in order to determine whether the responses of the other two groups were attributable to instruction. It was possible that students in an average environment might do as well on the tests as students in an optimal environment. On two tests--YES-NO and Matched-Pairs--Group OBE could have served this baseline function, except that students in the EPC and WOBE students were tested under dissimilar conditions. It was important to establish a control group as similar as possible to the experimental groups, including similarity in the rotation of tests and selection of teachers. Of course, Group OBE could not have served the baseline function for the All-NO test and the Picture test since they were not given to the children in that group.

The use of optimal and average learning environments made possible the consideration of an additional problem. A question frequently raised during discussion about the Senesh materials was, "Can average first-grade teachers adequately teach economic concepts without special training or experience?" Since the teachers in Group E had both special training and experience--they helped to develop the teaching materials--it was intended that comparison of PET-1 means among Groups E, P, and W would provide at least a tentative answer to that question.

All of the achievement tests in the EPC study were administered by the principle investigator. Funds were not available to hire additional testers, and the unrevised state of the Picture test made it difficult for anyone except the author to administer. It was also intended that the same person would administer the TOGA, but late delivery of the test booklets necessitated that it be given by the teachers in Groups E and C. Beginning with the third week in May and continuing to the first week in June, 1967, achievement tests were given in the early part of each of three consecutive weeks. The Elkhart students were tested first, then Groups C and P in that order. A lapse of 14 days occurred between the first achievement testing of the Elkhart group and the first achievement testing of Group P. Approximately twice that time lapsed between administration of the first and last TOGA tests.

All the tests for both studies were either scored by the investigator or by someone working under his direct supervision. To minimize scorer error, each test was corrected at least twice.

For both investigations, group means were compared using analysis of variance, with covariance used when needed. Individual items were

analyzed using chi-square. Reliability was estimated using split-half correlations adjusted with the Spearman-Brown Prophecy Formula. Analysis of variance and covariance were computed by the Utah State University Computer Center. All computations, including analysis of variance and covariance, were also computed by the author on a desk calculator, with each calculation performed at least twice.

CHAPTER IV

RELIABILITY AND VALIDITY OF TESTS

As pointed out in Chapter III, the EPC study was designed primarily to investigate problems of validity and reliability, and the WOBE study was designed primarily for comparing achievement gains. However, each study provided both types of information.

Reliability

As already noted, reliability coefficients were computed using odd-even split-half correlations corrected with the Spearman-Brown Prophecy Formula. Coefficients were computed for the TOGA and for each of the PET-1 instruments. Separate coefficients were computed for each of the two groups in the WOBE study and for each of the three groups in the EPC study.

Reliability coefficients for the TOGA ranged from .85 to .89 for the five subgroups in the two studies. These coefficients were nearly as high as some authors recommended for differentiating between individual students, and were considerably higher than the minimum for comparing group means (Garrett, 1958, p. 351).

Reliability coefficients were computed and compared for the following versions of PET-1: (1) The YES-NO test scored in the ordinary manner, (2) The YES-NO test scored in Matched-Pairs, (3) The All-NO test, and (4) The Picture test. Two related questions concerning reliability were of particular concern: (1) What was the range of coefficients

obtained for each instrument? and (2) How stable were the coefficients for each instrument? "Stability of reliability" is here defined as the tendency for the coefficient to remain constant regardless of the knowledge possessed by the group being tested. Concern for both the magnitude and the stability of the coefficients of reliability was necessary because the coefficients could vary in either dimension. They could consistently be relatively large or small, or they could be inconsistent--large on one testing or with one group, and small the next.

Although the word "stable" usually has positive connotation, in this case it was considered a mark of invalidity. Achievement test scores should be more reliable for knowledgeable students than for ignorant students. In the latter case, a larger proportion of the students' correct responses will be due to chance. Thus, if the reliability coefficients on a two-option test are stable--similar for control and experimental groups--the instrument is probably testing something other than, or in addition to, knowledge--assuming that the experimental treatment has had an impact.

The following table combines data from the two Shaver-Larkins studies, illustrating variance in reliability between experimental and control groups.

Table 3. Split-half reliabilities from prior studies

	YES-NO[a] 60 Items	YES-NO 30 Items	Matched-Pairs 15 Pairs
Control Group	.28	.14	.46
Experimental Group	.56	.35	.60

[a]The 60-item YES-NO test and the 30-item YES-NO test were separate instruments given to different groups at different times. The third column refers to the 30-item test scored with matched-pairs.

In both of the Shaver-Larkins studies, the reliability coefficients for the YES-NO test were higher for the experimental group than for the control group. This tendency, for the reliability to vary with the knowledge possessed by the group tested occurred with both ordinary and Matched-Pairs scoring.

Previous experience with YES-NO and Matched-Pairs tests, Cronbach's advice concerning YES-NO tests, and experience gained through one encounter with an All-NO test were used to formulate expectations concerning the comparative reliabilities of the PET-1 instruments used in this investigation.

Hypothesis 1: Reliability coefficients for the All-NO test would be higher than for the YES-NO test scored in either the ordinary manner or with Matched-Pairs.

Hypothesis 2: Reliability coefficients for the YES-NO test scored with Matched-Pairs would be higher than when the same test was scored in the ordinary manner.

Hypothesis 3: Reliability coefficients for the Picture test would be higher than for the YES-NO test scored in either the ordinary manner or with Matched-Pairs.

No predictions were made prior to analysis concerning: (1) The stability of reliability for either the All-NO test or the Picture test, and (2) The comparative magnitude of reliability coefficients between the All-NO test and the Picture test. Even though predictions were not made relevant findings will be noted on the following pages.

It was decided that for practical significance, differences in these comparisons would have to exceed statistical significance at the

.01 level. However, statistical significance was taken as a minimal standard and is not emphasized. Small differences between reliability coefficients may be statistically significant but not practically significant.

Reliability coefficients for all of the PET-1 tests were given to each group in the WOBE and EPC studies are listed in Table 4.

Table 4. Split-half reliability coefficients for the WOBE and EPC studies

Group	N	YES-NO 75 Items	YES-NO Matched-Pairs 37 Pairs	All-NO 74 Items	Picture 49 Items
WOBE					
W	96	.60	.75		
OBE[a]	100	.17	.54		
EPC					
E	77	.68	.85 (.91)[b]	.90	.84 (.89)[b]
P	59	.48	.66 (.80)	.89	.77 (.83)
C[a]	77	.29	.62 (.77)	.87	.74 (.81)

[a]Control groups.

[b]The reliability coefficients which are not in parentheses are ordinary split-half correlations corrected with the Spearman-Brown Prophecy Formula. The coefficients in parentheses are predictions of the coefficients that would be obtained if the Matched-Pairs test and the Picture test contained as many items as the YES-NO test and the All-NO test. The Spearman-Brown Prophecy Formula was applied a second time to the coefficients not in parentheses in order to make these predictions.

The findings in Table 4 indicate that Hypotheses 1 to 3 were generally supported.

1. Reliability coefficients for the All-NO test were higher than for the YES-NO test scored either way. In Group E of the EPC study,

reliability of the YES-NO Matched-Pairs test approached that of the All-NO test, but the difference between coefficients was still significant at the .01 level--computed using the SE_D between two correlations (Garrett, 1958, pp. 241-243). Theoretically, the YES-NO Matched-Pairs test was nearly as reliable as the All-NO test. That is, the coefficients for the YES-NO Matched-Pairs test are nearly as high as those for the All-NO test when computed for instruments of equivalent length--see Table 4. However, a Matched-Pairs test containing 30 pairs of items takes as long to administer as an All-NO test containing 60 items. And it would be difficult to administer more than thirty or forty pairs of items to first-grade children at any one setting.

2. In every group, the reliability coefficients for the YES-NO test scored with Matched-Pairs were higher than when scored in the ordinary manner. This increase in reliability has practical significance. Garrett (1958, p. 351) claims, "In order to differentiate between the means of two school grades of relatively narrow range, a reliability coefficient need be no higher than .50 or .60." Coefficients for the YES-NO test scored in the ordinary manner are clearly below that standard in two groups. When scored with Matched-Pairs, every group was above the mark. However, even when scored with Matched-Pairs no group attained a coefficient of .90, which Garrett claims is necessary to differentiate pupil from pupil.[1] For this purpose, the reliability of the All-No test is more nearly acceptable. Of course, high reliability is of little

---

[1]Since the research design for this dissertation does not call for differentiating pupil from pupil, the lower reliability of the Matched-Pairs test is acceptable. Of course, other things being equal, high reliability is desirable.

consequence if a test lacks validity. The questionable validity of the All-NO test will be discussed later.

3. Reliability coefficients for the Picture test were generally higher than for the YES-NO test scored in the ordinary manner. The reliability coefficient for the Picture test was not higher than for the Matched-Pairs test in Group E, but was slightly higher in Groups P and C. When coefficients were estimated for tests of equivalent length there was very little difference between the Picture test and the YES-NO Matched-Pairs test. It would be difficult, however, to extend the Matched-Pairs test beyond 37 pairs of items and administer it in one setting.

As stated previously, even though hypotheses were not formulated, two other comparisons were made concerning reliability:

1. Reliability coefficients for the YES-NO test, scored either in the ordinary manner or with Matched-Pairs, varied with the knowledge level of the group tested. E was expected to be the most knowledgeable group, followed by P, W, C, and OBE, in that order. Reliability coefficients for both scorings of the YES-NO test were consistent with this expectation, except that the order of Groups P and W was reversed. Examination of the mean scores for these two groups--presented in the next chapter--explains why this occurred; contrary to expectation, Group P was less knowledgeable than Group W. Thus, the earlier argument that reliability will fluctuate with knowledge was supported in each instance.

Because the All-NO test produced such stable reliability coefficients, its validity must be questioned. As stated earlier, a two-option test which produces reliability coefficients which do not vary from experimental

to control groups is probably testing something other than knowledge--assuming that the experimental treatment is having an effect.

2. In their present state, the All-NO test is more reliable than the Picture test. This is not true when coefficients are computed for tests of equal length. But it is doubtful whether this particular Picture test could be lengthened and still be administered to first-grade children in a reasonable amount of time.

In summary, regarding both magnitude of reliability and stability of reliability, the four tests ranked: (1) the All-NO test, (2) the Picture test, (3) the Matched-Pairs test, and (4) the YES-NO test. Considering only reliability, any of the first three tests is adequate for the major purposes of this dissertation, such as differentiating between group means. The All-NO test may also be adequate for discriminating between individual students. Reliability alone, however, is not sufficient, and there is reason to suspect the validity of the All-NO test.

Validity

Investigators attempting to assess learning in relation to new curricula may reasonably be faced with one or both of two general validity problems: (1) Is the content of the course of study valid? (2) Are the instruments valid which are used to assess learning of that content? These two questions need not be studied simultaneously. Either one is worthy of investigation. Therefore, it needs to be stressed that the investigations upon which this paper are based were not concerned with whether concepts in Families at Work adequately

represent the disciplines from which they are claimed to be drawn. Some aspects of the present study provide information on this problem, but only incidentally. However, the second problem--test validity--was a central concern.

Classifying the validity problems in the EPC and WOBE studies

Developing suitable tests for assessing learning of concepts in Families at Work required both suitable content and suitable test forms. One of the questions raised in determining the adequacy of test forms was whether the form of the test would affect the validity of the scores. On the basis of preliminary studies and a review of the literature on acquiescence-set, it was argued in Chapter II that test form can be as critical to validity as test content.

Considerations of the effect that test form might have on validity were difficult to carry on within the usual validity classifications. Test validity is commonly discussed under four headings: Content, predictive, concurrent, and construct (Borg, 1963, pp. 80-84). Only one of these headings--content validity--is clearly related to the studies reported in this dissertation, and the problem of the validity of test form does not appear to be clearly classifiable in any of the four categories.

Predictive validity refers to the degree to which test scores can be used to predict success in some activity. Although it is conceivable that PET-1 scores might be used to predict achievement in learning the concepts in Our Working World, the tests were not designed to be used for that purpose. Moreover, the problems investigated in this dissertation

did not require that the PET-1 tests have high predictive validity. For instance, whether posttest PET-1 scores could be predicted from pretest PET-1 scores was of no concern as long as the posttest scores accurately indicated knowledge of the content of Families at Work. Similarly, whether achievement on the second-grade materials could be predicted from PET-1 scores was of little concern as long as those scores accurately indicated students' knowledge of the first-grade materials.

Concurrent validity is related to predictive validity. The difference is that the criterion measure for concurrent validity is taken at the same time or nearly the same time as the predictive measure. Concurrent validity was not important to the problems investigated in this dissertation because there was no concurrent criterion of concern.

Construct validity refers to the degree to which a test is based on a particular theory, or theoretical construct, and substantiates predictions made on the basis of the theory or construct. One of the major concerns in selecting forms for the PET-1 instruments was the available knowledge concerning the effects of acquiescence-set on test responses. To make predictions based on an acquiescence-set construct when the test is not intended to measure acquiescence violates common usage of the term "construct validity." Nevertheless, if a test measures an achievement construct--in this case the knowledge of economic concepts--it should discriminate between groups which have achieved and those which have not. This is similar to the notion of confirming prediction based on a psychological construct. Acquiescence-set as a contaminating variable--one that interfers with the measurement of knowledge--might well affect construct validity in that sense. Yet, it does not fit the category very neatly.

Since none of the four common types of test validity provides a ready category for the effect of test form on validity, it was decided to present the findings concerning the validity of PET-1 tests under two headings. The first heading is the familiar category "content validity." The second heading is a stipulated category called "form validity." In the sense in which it will be used in the remainder of this paper, "form validity" refers to the degree to which the form of the test affects the validity of the findings. That is, it is assumed that changing the form of a test while holding the content constant can affect the findings--the scores, and thereby the means or standard deviations--obtained. It is further assumed that the findings obtained from some test forms, excluding differences in content, may be spurious in the sense that they are misleading, and therefore invalid, estimates of knowledge. Of course, a type of construct validity--i.e., does the test discriminate between knowledgeable and ignorant groups--will be mentioned later in this chapter.

Content validity

Content validity refers to the degree to which the content of a test represents the content of the course of study upon which it was based. Establishing the content validity of the PET-1 tests was largely a matter of comparing the content of the tests to the content of Families at Work. The manner in which the content of PET-1 items was selected, including the precautions taken to insure content validity, has been explained in Chapter II. In addition to taking care in selecting content for the test items, reactions were sought from teachers who used Families at Work and

who agreed to take part in several preliminary studies. Reactions were also sought from teachers and supervisors who took part in the WOBE and EPC studies. Included in this group were Joseph Rueff--the Coordinator of the Social Science Research Projects in Elkhart, Indiana--and the three cooperating teachers in Elkhart. These cooperating teachers and supervisors agreed that the content validity of the YES-NO, Matched-Pairs, and All-NO tests is high. However, in the EPC study, some doubts were expressed concerning the content validity of the Picture test. This was to be expected. The Picture test was completed in original form immediately prior to testing, with no opportunity for revision.[2]

In addition, a general sign of the content validity of an achievement test is whether it discriminates between knowledgeable and ignorant groups of students.[3] As noted earlier, this is a type of construct validity. As will be seen in Chapter V, all of the PET-1 tests produced means which significantly differed at the .01 level between at least one set of experimental and control groups.

Form validity

Allowing only test form to vary. Form validity is of concern only if the form of the test affects the findings. In order to check on the

---

[2]The Picture test posed some special problems which are discussed in detail in the last section of this chapter.

[3]The expectation that an achievement test will discriminate between experimental and control groups is based on the assumption that the test measures what the experimental group has been taught. That is, that the test items faithfully reflect concepts that were presented in a course of study. Of course a test may discriminate between control and experimental groups without adequately sampling all the concepts in a course of study.

effects of test form, it is necessary to hold the content of the PET-1 tests constant while varying the form of the tests. This was accomplished by using subsections of the YES-NO test; YES-NO, Matched-Pairs, and All-NO scores were taken from a single administration of the YES-NO test in the EPC study. Of course, YES-NO and Matched-Pairs scores are ordinarily obtained from the same administration of the test. The only change in procedure, then, was that the 38 NO items on the YES-NO test were treated as an All-NO test. These items are labeled "NO test" to avoid confusion with the longer All-NO test.

Since scores for all three test forms were derived from a single administration of the YES-NO test, extraneous variables other than content were held constant. These were maturation, differences in testing environment, the learning effects of multiple testing with different forms of the same test, and loss of subjects.

Another important variable was test length. Scores, means, and standard deviations cannot be compared directly unless they are derived from tests containing an equal number of items as well as similar content. The NO test included only half of the items on the YES-NO test. The Matched-Pairs test was also only half as long as the YES-NO test because Matched-Pairs scoring treats pairs of items as one. Therefore, YES-NO means and standard deviations were halved before being compared directly to means and standard deviations on the Matched-Pairs and NO tests.

Since the NO test contains only half of the items on the YES-NO and Matched-Pairs tests, the question arises whether the content of the three test forms is really held constant. Unlike the YES-NO and Matched-

Pairs tests, it is impossible by definition for any All-NO test to have content identical to a test containing YES items. However, since each of the items in the NO test is the reversal of a YES item, the content of the NO test is nearly identical to the content of the YES-NO and Matched-Pairs tests. The only differences are minor changes in wording necessitated by the production of reversals. Despite these minor changes, when reversals are carefully written the substance of the item content should remain constant.

Correlation of PET-1 scores between different test forms. The effects, if any, of test form on test scores is difficult to determine by inspection. Pearson product-moment correlations, however, provide a useful index of proportional variance between groups of scores. Therefore, to determine whether changing the test form affects PET-1 scores if all other variables are held constant, correlation coefficients were computed between scores on the YES-NO and Matched-Pairs tests, the Matched-Pairs and NO tests, and the YES-NO and NO tests. Since content was held constant among the three tests, it was expected that correlation coefficients would be large for each of the above pairings. It was also expected that correlation coefficients would be largest between groups of scores from the two test forms with the highest validity. Since theoretically the Matched-Pairs and YES-NO tests control best for acquiescence it was assumed that their form validity was higher than that of the All-NO test. Therefore, it was expected that the correlation between Matched-Pairs and YES-NO scores would be significantly higher at the .01 level than the correlation between Matched-Pairs and NO scores or between NO and YES-NO scores. Testing this expectation required

computing a test of the significance of the difference between two correlations. Therefore, Hypothesis 4 is stated in the null form.

Hypothesis 4: There will be no significant differences at the .01 level among correlation coefficients for the Matched-Pairs and YES-NO scores, the Matched-Pairs and NO scores, and the YES-NO and NO scores.

Findings for Hypothesis 4 are presented in Table 5.

Table 5. Pearson product-moment correlations among test forms obtained by three separate scorings of a single administration of the YES-NO test

		Matched-Pairs[c] and YES-NO[b]		Matched-Pairs and NO[d]		YES-NO and NO	
Group	N	r	r^2[a]	r	r^2	r	r^2
Elkhart[e]	81	.97	.94	.88	.77	.86	.74

[a]Proportion of variance which the two sets of scores have in common.

[b]The YES-NO test scored in the ordinary manner.

[c]The YES-NO test scored with Matched-Pairs.

[d]The NO items on the YES-NO test; treated here as an All-NO test.

[e]The Elkhart group was chosen for this comparison because it produced the largest reliability coefficients for the three test forms compared. Therefore, variability due to low reliability would be less for scores taken from this group. The obtained reliability coefficients for the three tests given to the Elkhart group were: YES-NO = .68, Matched-Pairs = .85, and All-NO = .90.

Hypothesis 4 was tested by transforming the r's for Matched-Pairs and YES-NO (.97), and Matched-Pairs and NO (.88) using Fisher's z function and comparing the differences between the two z coefficients (Garrett,

1958, pp. 241-242). This method produced a critical ratio of 3.74 compared to 2.58 needed for significance at the .01 level. This technique is not strictly appropriate because the scores which were correlated are not independent; they were derived by scoring a single administration of a single test in three different ways, and are therefore based on identical or nearly identical content. However, Garrett claims that this method underestimates, rather than overestimates, the significance of the difference between two correlation coefficients (pp. 242-243).

Since the critical ratio of 3.74 was larger than the 2.58 needed for significance at the .01 level, Hypothesis 4 was not rejected. The correlation between Matched-Pairs and YES-NO scores was higher than between Matched-Pairs and NO scores, and was therefore also higher than between NO and YES-NO scores.

Even though Hypothesis 4 was not rejected, two additional points must be considered in deciding whether varying the form of the test affects scores significantly. First, the largest coefficient in Table 5 was between the test forms which had identical content--the YES-NO and Matched-Pairs tests. It is possible that differences between the YES items in these two tests and the reversed NO items in the NO test account for the lower correlation coefficients obtained in comparisons involving the NO test. Since one of the major concerns in producing the YES-NO Matched-Pairs test was to write reversals with identical or nearly identical content, it is not likely that the 20 percent difference in common variance between the Matched-Pairs and YES-NO scores (.94) and the NO and YES-NO scores (.74) can be accounted for by

differences in the content of reversals. However, it is possible, even if unlikely. Therefore, conclusions based on the correlations in Table 5 should be held with some tentativity.

The second point to be taken into consideration is the possible effect of the reliability of the three tests on the correlation coefficients in Table 5. The reliability coefficients for the three tests given to the Elkhart group were .68 for the YES-NO scores, .85 for the Matched-Pairs scores, and .84[4] for the NO scores. The scores from the Elkhart students were chosen for the comparisons in Table 5 because the reliability coefficients for all three tests were higher in this group than in any other. Thus, there would be less variability due to low reliability. It appears, however, from an examination of the coefficients in Table 5 that differences in reliability account for very little of the differences in the degree of correlation among the three tests. If reliability were a major factor the highest correlation coefficient should have occurred between the two most reliable tests--the Matched-Pairs test and the NO test. Furthermore, the coefficient between Matched-Pairs and YES-NO should have been no larger than the coefficient between NO and YES-NO.

Since neither of the two additional considerations mentioned above is likely to have significantly influenced the correlation coefficients,

---

[4]This reliability coefficient was not computed directly from the NO test. The obtained split-half reliability coefficient for the All-NO test was .90 corrected with the Spearman-Brown Prophecy Formula. Since the NO test was only half as long as the All-NO test, the reliability of the NO test was estimated by reapplying the Spearman-Brown Prophecy Formula to the All-NO reliability coefficient. Since the reliability of the All-NO test form appears to be both high and stable, this procedure probably resulted in a close approximation of the reliability of the NO test.

it appears that varying the form of a test can affect the test scores.

Comparing standard deviations. The discussion associated with Hypothesis 4 and Table 5 in the previous section was centered on the question, "If all other variables are held constant, do All-NO scores differ from Matched-Pairs scores and YES-NO scores?" This same question can be considered by comparing the standard deviations of groups of scores for each of the three test forms derived from a single administration of the YES-NO test. Findings used to make this comparison are presented in Table 6.

Table 6. Means and standard deviations for YES-NO, Matched-Pairs, and NO tests derived from a single administration of the YES-NO test

		NO[c]		Matched-Pairs		YES-NO[e]	
Group	N[a]	M[b]	SD[d]	M	SD	½M	½SD
E	77	25.16	6.31	20.46	6.58	27.75	3.76
P	59	20.05	6.58	15.14	5.41	24.15	3.38
C	46	19.15	6.46	13.65	4.78	23.30	2.97

[a]The number of students in the group.

[b]The mean.

[c]The NO half of the YES-NO test. It is treated here as an All-NO test form.

[d]Standard deviation.

[e]The YES-NO test is twice as long as the others. In order to make a direct comparison its means and standard deviations were reduced by half.

Two observations are of particular interest in regards to Table 6:

1. In all three test groups, the YES-NO test is less variable than either of the other tests--its standard deviations are smaller. The

theoretical explanation is that since students tend to be acquiescent, YES items obscure differences between ignorant and knowledgeable students. Both respond YES; one from knowledge, the other from response-set. Because there is little variability among students on the YES items, the standard deviation for the total test is reduced.

2. In Groups P and C--the least knowledgeable groups--the standard deviations for the NO test are larger than for the Matched-Pairs test. Furthermore, the standard deviations for the NO test are similar in all three groups--E, P, and C--but the standard deviations for the Matched-Pairs test and the YES-NO test decrease from Groups E to C. This is indicative of the greater validity of the Matched-Pairs and YES-NO tests, reflecting the expectation that an ignorant group would be less variable in knowledge than a group which received instruction.

It can be concluded that the All-NO and YES-NO standard deviations are spurious if taken as indicators of variability in knowledge. Scores on the YES-NO test are less variable in all groups than they would be if YES items did not obscure differences in knowledge. In contrast, scores on the All-NO test are more variable in the control groups than would be expected if the instrument were not measuring acquiescence in addition to knowledge. Moreover, standard deviations for the Matched-Pairs test are not spuriously small as indicators of variability in knowledge--the weakness of the YES-NO test; nor are they spuriously large in the control groups--the weakness of the All-NO test.

Comparing F-ratios and t-ratios. As indicated by the standard deviations in Table 6, the variability of YES-NO and All-NO scores is affected by response-set as well as knowledge. Since parametric tests

of significance utilize sample variance, i.e., the standard deviation, to estimate population variance, it is possible that spurious variability--variability confounding response-set and knowledge--will lead to spurious estimates of the significance of the difference between means. For instance, when acquiescence is confounded with knowledge, groups might appear to differ in knowledge when they do not, or groups might appear not to differ in knowledge when they do differ.

On the basis of the standard deviations presented in Table 6, expectations were formulated as to how differences in variability might affect estimates of the significance of the difference between means. In Table 6, variability of NO scores in the control group is larger than expected if the scores did not confound acquiescence and knowledge. Furthermore, the difference between All-NO means is smaller than the difference between Matched-Pairs means even though the All-NO scores are more variable. It was therefore expected that the All-NO test would produce smaller t-ratios or F-ratios than the Matched-Pairs test.

Just the opposite prediction was made for the YES-NO test. In Table 6, the standard deviations of the YES-NO scores are about 40 percent smaller in all groups than the standard deviations for the Matched-Pairs scores. It was expected that since YES-NO scores are less variable than Matched-Pairs scores, the YES-NO test would produce larger t-ratios or F-ratios. This prediction was made with less assurance because the difference between YES-NO means in Table 6 is smaller than the difference between Matched-Pairs means.

Hypothesis 5 was directed at the central problem raised in the above argument. Spurious variability might lead to spurious estimates of the

significance of the difference between means--estimates which lead to erroneous conclusions concerning the knowledge possessed by the groups being compared. Because a test of significance was not applicable to Hypothesis 5, it is stated as a research expectation rather than in the null form.

Hypothesis 5: Different estimates of the significance of the difference between means will be obtained when YES-NO, Matched-Pairs, and All-NO scores are taken from a single administration of the YES-NO test.

Data used to test Hypothesis 5 are presented in Tables 7, 8, and 9. As in Tables 5 and 6, it was necessary to compare scores from a single administration of the YES-NO test so that differences in such variables as item content or student maturation would not be confounded with differences in test form. Again, as in Tables 5 and 6, the content of the NO test is identical to half of the content of the YES-NO and Matched-Pairs tests--the NO items--and is the mirror image of the other half--the YES items. The means and standard deviations for the YES-NO test in Table 7 do not coincide with those in Table 6, because those in Table 6 were halved to make them directly comparable to the means and standard deviations for the Matched-Pairs and NO tests.

Table 7. Analysis of covariance among Groups E, P, and C on the YES-NO test

Group[e]	N	TOGA[a] M	TOGA[a] SD[c]	TOGA[a] F[d]	YES-NO M	YES-NO SD	YES-NO F	Adjusted YES-NO[b] M	Adjusted YES-NO[b] SD	Adjusted YES-NO[b] F
E	77	48.83			55.49			53.70		
P	59	38.86	7.43	30.21	48.29	6.95	29.79	50.78	6.19	19.34
C	46	45.11			46.59			46.40		

Groups[e] compared	Differences between adjusted YES-NO M's	SE_D between YES-NO M's	t-ratio[f]
E and P	2.92	1.49	1.96
P and C	4.38	1.73	2.53
E and C	7.30	1.61	4.53

$t_{.05} = 1.98$, $t_{.01} = 2.61$ $\quad F_{.05} = 3.04$, $F_{.01} = 4.71$

[a] TOGA stands for Tests of General Ability. Raw scores from this test were used to adjust for initial differences among groups.

[b] These are the scores on the PET-1 YES-NO test after adjustments were made for initial differences in mental ability.

[c] SD stands for standard deviation. Only the general standard deviation, available from the analysis of variance, is given in this table.

[d] F stands for the F-ratios obtained in analysis of variance and covariance.

[e] Groups E and P are experimental groups. However, Group P was much lower in initial ability than the other two groups and its scores more closely resemble those of Group C, the control group.

[f] Differences between pairs of groups were tested for significance using the t-test in the manner outlined by Garrett (1958, pp. 302-303).

Table 8. Analysis of covariance among Groups E, P, and C on the Matched-Pairs test

Group	N	TOGA M	TOGA SD	TOGA F	Matched-Pairs[a] M	Matched-Pairs SD	Matched-Pairs F	Adjusted Matched-Pairs M	Adjusted Matched-Pairs SD	Adjusted Matched-Pairs F
E	77	48.83			20.46			18.79		
D	59	38.86	7.43	30.21	15.14	5.84	24.00	17.46	5.06	16.03
C	46	45.11			13.65			13.47		

Groups compared	Differences between adjusted Matched-Pairs M's	SE_D between Matched-Pairs M's	t-ratio
E and P	1.33	1.21	1.10
P and C	3.99	1.42	2.81
E and C	5.32	1.32	4.00

$t_{.05} = 1.98$, $t_{.01} = 2.61$ $F_{.05} = 3.04$, $f_{.01} = 4.71$

[a]Matched-Pairs stands for the YES-NO test scored using the matched-pairs technique.
Most of the symbols used in this table are identical to those used in Table 11 and are explained there.

Table 9. Analysis of covariance among Groups E, P, and C on the NO test

Group	N	TOGA M	TOGA SD	TOGA F	NO[a] M	NO SD	NO F	Adjusted NO M	Adjusted NO SD	Adjusted NO F
E	77	48.83			25.16			23.33		
P	59	38.86	7.43	30.21	20.05	6.49	16.20	22.60	5.60	9.16
C	46	45.11			19.15			18.95		

Groups compared	Differences between adjusted NO M's	SE_D between NO M's	t-ratio
E and P	.73	1.34	.54
P and C	3.65	1.57	2.32
E and C	4.38	1.46	3.00

$t_{.05} = 1.98$, $t_{.01} = 2.61$ $F_{.05} = 3.04$, $F_{.01} = 4.71$

[a]NO stands for the NO half of the YES-NO test. It is considered here to be an All-NO test form.
Most of the symbols used in this table are identical to those used in Table 11 and are explained there.

Hypothesis 5 was supported by the data in Tables 7, 8, and 9. By inspection, the F-ratios differ among the three tests. Certainly, the F-ratio for the All-NO test--Table 9--is smaller than for the other two tests--Tables 7 and 8. However, it cannot be claimed with assurance that the F-ratio for the YES-NO test--Table 7--is significantly larger than for the Matched-Pairs test--Table 8. The F-ratios for the general differences among groups are consistent, then, not only with Hypothesis 5, but also with the predictions listed prior to Hypothesis 5. It was predicted that the All-NO test would produce smaller t-ratios or F-ratios than the Matched-Pairs test. The degree to which these expectations were born out is even more apparent when the t-ratios for the comparisons between paris of groups are examined. To facilitate this examination the t-ratios from Tables 7 - 9 are reproduced in Table 10.

Table 10. T-ratios reproduced from Tables 7 - 9

Groups	Adjusted NO	Adjusted Matched-Pairs	Adjusted YES-NO
E and P	.54	1.10	1.96
P and C	2.32	2.81	2.53
E and C	3.00	4.00	4.53

$t_{.05} = 1.98$ $t_{.01} = 2.61$

The t-ratios in Table 10 support in three ways the expectations concerning the validity of the PET-1 tests. First, Groups E and P were both experimental and were not expected to differ significantly on the PET-1 tests after adjustments were made for initial differences in mental ability. The findings from two PET-1 tests were consistent with that

expectation; Groups E and P did not differ at the .05 level of significance on the Matched-Pairs and NO tests. The t-ratio of 1.96 between Groups E and P on the YES-NO test, however, barely fell short of the 1.98 needed for significance at the .05 level. This supports the expectation that the YES-NO test produces spuriously high t-ratios. That is, had an investigator used only the YES-NO test he would have been tempted to tentatively conclude that Groups E and P significantly differ in knowledge of the content of Families at Work. In light of the expectations concerning the achievement of Groups E and P, and in light of the t-ratios from the Matched-Pairs and NO tests, the conclusion which would likely have been made on the basis of the YES-NO test alone would be misleading.

Second, since P is an experimental group and C is a control group, they were expected to differ at the .01 level of significance on the PET-1 tests. This expectation was born out only on the Matched-Pairs test. Groups P and C differed at the .05 level on the NO and YES-NO tests, but did not differ at the .01 level of significance.

Third, the pattern of t-ratios for the PET-1 tests is, in general, consistent with the expectations listed prior to Hypothesis 5. For all three pairs of groups--E and P, P and C, and E and C, in Tables 7, 8, and 9--the NO test produced the smallest t-ratios. For two pairs of groups--E and P, and E and C--the YES-NO test produced the largest t-ratios and the Matched-Pairs test produced intermediate t-ratios. For only one pair of groups--P and C--the Matched-Pairs test produced the largest t-ratio and the YES-NO test produced the intermediate t-ratio.

In brief, the major theoretical argument in this section was that spurious variability might lead to spurious estimates of significance. Hypothesis 5 was not rejected; different estimates of the significance of the difference between means were obtained when YES-NO, Matched-Pairs, and All-NO scores were taken from a single administration of the YES-NO test. It was concluded that the YES-NO test is likely to overestimate the significance of the difference between means, and that the All-NO test is likely to underestimate significance.

Validity of the PET-1 Picture test

The Picture test has three dimensions: (1) the number of options on each item, (2) the pictures, and (3) the instructions for each item. Any of these dimensions might affect student's scores.

The first dimension--the number of options on each item--is clearly related to the problem of reliability. That is, the number of correct chance responses is largely determined by the number of options on each item. The number of options may also be related to the problem of validity, because there is some evidence that multiple-choice tests are subject to response-set (Barnes, 1962). Response-set, however, is less serious in multiple-choice tests than in two-option instruments such as the YES-NO or All-NO tests (Cronbach, 1950). Therefore, the effects of response-set on the validity of the Picture test were not investigated.

The major validity question centered on the test's content. The content of the Picture test is contained in both the pictures and the instructions which accompany the pictures for each item. That is, on each item the children are told to look at a set of pictures and to

select one of the pictures according to the instructions given. If either the content of the pictures or the content of the instructions is inappropriate, the content validity of that item is affected.

The following examples may help to clarify the above point. Concerning the content of the pictures, some difficulty was experienced in representing certain concepts pictorially--for example, the concept "economic interdependence." Perhaps creative test designers could resolve these difficulties, but it may be that certain concepts cannot validly be represented by pictures alone. The probability of such inherent invalidity likely increases as attempts are made to test for something more than the simple recognition of the correct application of a term to a concrete situation. On the other hand, little difficulty was experienced in converting into pictures those concepts which were stressed most often in the Senesh materials. Picture items testing concepts such as "producer" and "consumer" discriminated between groups as well or better than items testing similar concepts on the YES-NO or All-NO forms.

Misunderstanding between the author and the artist was the most obvious source of invalidity in the pictures. For instance, in a picture intended to test the concept, "producer of goods," the artist drew a production line in a factory but did not draw a person working on the line. However, such oversights are readily recognized and easily corrected.

A more serious problem was the manner in which the content of pictures was fitted to the multiple-choice form. Weaknesses in design are evident in several of the items dealing with "producer" and "consumer."

Converting these concepts to pictures was not difficult, and items discriminated very well between control and experimental groups. Nevertheless, something in the pictures appeared to clue students to the correct answers. This is indicated by the fact that students in the control group did better on these items than would be expected on the basis of chance even though they did not do as well as experimental group students. Of course, this could also be explained by the assumption that control group students were not completely ignorant of the content of the test.

In Table 11, the expected frequencies of correct response are compared to the obtained frequencies of correct response in the control group. These data were obtained from the administration of the Picture test to 82 students in Group C--the control group in the EPC study. The column labled "Items" refers to groups of items which test a single concept. The first thirty items on the Picture test contain five options, and the last 19 contain four options. On the basis of chance, 16 students should have correctly responded to each of the first 30 items, and 21 students should have correctly responded to each of the last 19 items (see the column labled "Expected Frequency). An obtained frequency of 26 for items 1-30 differs from the expected frequency of 16 at the .01 level of significance. An obtained frequency of 32 for items 31-39 differs from the expected frequency of 21 at the .01 level of significance. The first entry should be read, "For items 1 through 10, which test the concept 'producer,' the expected frequency of correct response is 16, the frequency needed to differ from chance at the .01 level of significance is 26, and the obtained mean frequency was 51."

Table 11. Expected and obtained frequencies of correct response by Group C to various types of items on the Picture test

Items	Concept	Expected Frequency	Frequency[a] Greater than Chance	Mean Obtained Frequency	
1-10	Producer	16	26	51	< .01
11-15	Consumer who is not producing	16	26	38	< .01
16-21	Producer of goods	16	26	17	
22-27	Producer of services	16	26	21	
28-30	Specialist	16	26	18	
31-33	Consumer who is not producing	21	32	16	
34-37	Producer	21	32	37	< .01
38-39	Consumer who is not producing	21	32	11	
40-45	Specialist	21	32	15	
46-49	Division of Labor	21	32	41	< .01

[a]Chi-square was used to determine the minimum frequency which is larger than the expected frequency.

The obtained frequencies of correct response for items 1-15, 34-37, and 46-49 were, on the average, larger than expected by chance.[5] These larger-than-chance frequencies could be interpreted as meaning either that the items were cluing the students to the correct answer, or that they already knew the concepts. Although not included in the original research design, it was decided to check on these possibilities by interviewing some of the control group classes. Therefore, immediately after testing was completed in Group C, individual classes were informally interviewed. Student responses in these interviews indicated that both

[5]It also appeared that some of the obtained frequencies might be lower than expected by chance--Items 31-33, 38-39, and 40-45. Chi-square values were computed comparing these frequencies to the chance expected frequency of 21. None were significantly lower than 21 at the .01 level.

of the above interpretations are correct: Some items tended to clue students to the correct response, and some students possessed relevant knowledge prior to testing. It was impossible to determine exactly the extent to which prior knowledge influenced students' responses. It is clear that in at least a few cases it had an affect. But it was the investigator's opinion that, compared to interviews with students in the experimental groups, the students in Group C were largely ignorant of the terms being tested. Only six students were able to give even approximate definitions of the word "producer," and at least that many students offered definitions that were completely inapplicable to the term.

Since the control group students gave little evidence of knowing the terms being tested, the content of the high frequency items was examined. One characteristic was apparent in the first 30 items. Each item contrasts four things that are alike to one thing that is different. For instance, when the students are instructed to mark the picture of the producer the other four pictures show consumers. It is possible that a first-grade student who has been trained in the readiness program to differentiate things that are alike from things that are different could obtain a better-than-chance score on these items without knowing the meaning of "producer" or "consumer."

This explanation is supported by incorrect definitions given by the students for the word "consumer." Several students variously described consumers as people who do not work, as people who are resting, or as people who are lazy. These definitions correspond to the pictures in items 11-15, and do not correspond to the definitions which a student

would get from his teacher or parents.

"Division of labor" was the only other concept tested in items which had a higher-than-chance frequency of correct response. These are four-option items, with two correct pictures and two incorrect pictures in each. The pictures were drawn in contrasting pairs. If one picture shows a family dividing the labor as they wash the car, the contrasting picture shows a family not dividing the labor as they wash the car. During the interviews, one girl gave a clue as to why so many students correctly answered these items. She marked the pictures that showed people doing things the way they are done at her house. Dividing the labor is the "natural" way of doing things.

Most of the preceding remarks about the content validity of the Picture test were directed at the pictures rather than at the verbal instructions accompanying the pictures. Obviously, if a picture misleads the students, or supplies them with extraneous clues, the content validity of the item is affected. However, as indicated in the following examples, the instructions for the items may also either mislead the student or present a concept in an inappropriate manner.

The Picture test had not been revised prior to its use in the EPC study. Taking this into consideration, along with the fact that the test was not comprehensive, it was decided to make revisions in the instructions to individual items during the course of the testing, if needed. Instructions were revised in two places. After the first class in Group E was tested, the instructions for items 28-30 and 40-45 were changed from "Mark the picture which shows a man who specialized," to "Mark the picture which shows a man who is not a specialist." The second

change in instructions occurred when Group P was being tested. Items 31-33 and 38-39 were changed from "Mark the consumer," to "Mark the consumer who is not a producer."

The first change was made after one of the teachers pointed out that the word "specialized" is not used in Families at Work. The instruction, then, did not have content validity when compared to the teaching materials. The second change was made for similar reasons. Every producer is also a consumer, so that children who were instructed to "Mark the consumer" could be marking producers and not receive credit for knowing the concept.

Part of the justification for making these changes during testing was that the teachers' cooperation was needed for a period of three days. If they felt the Picture test was invalid, and thus unfair to their students, their reaction to the two-option tests might be affected. Of course, failure to use identical instructions with all groups may have affected the validity of the comparisons between groups on the Picture test. Nevertheless, the Picture test still provided some useful information, and little was lost in content since the YES-NO, Matched-Pairs, and All-NO tests were comprehensive.

Groups E, C, and P were tested in that order. Changes in instructions were made during testing of Groups E and P. It was assumed that these changes increased the opportunity of students to correctly respond to the revised items. If this assumption is correct, then Group P should have scored higher in relation to the other two groups than on the two-option tests. Table 12 presents the means for Groups E, P, and C on the Picture test, plus the data from an analysis of covariance to determine the significance of the difference between means.

Table 12. Analysis of covariance among Groups E, P, and C on the Picture test

Group	N	TOGA M	TOGA SD	TOGA F	Picture M	Picture SD	Picture F	Adjusted Picture M	Adjusted Picture SD	Adjusted Picture F
E	77	48.83			24.84			23.81		
P	59	38.86	7.82	27.27	23.85	5.90	32.43	25.32	5.60	37.77
C	77	45.10			17.69			17.59		

Groups compared	Differences between adjusted Picture M's	SE_D between Picture M's	Differences needed for significance .05	.01
E and P	1.51	1.34	2.64	3,48
P and C	7.73	1.34	2.64	3.48
E and C	6.22	1.23	2.42	3.20

$t_{.05} = 1.98$, $t_{.01} = 2.61$, $F_{.05} = 3.04$, $F_{.01} = 4.71$

Means in Table 12 indicate, when compared to means in Tables 7, 8 and 9, that Group P scored somewhat higher in relation to the other two groups than on the YES-NO, Matched-Pairs, and All-NO tests.

An attempt was made to determine whether the improved showing by Group P was due to changes in instructions. As stated previously, instructions were revised in two places. The first change occurred for items 28-30 and 40-45. When the frequencies of correct response for each of these items were compared among Groups E, P, and C, using chi-square contingency tables, there were no differences significant at the .01 level. Only one of the nine items produced a significant difference at the .05 level. In addition to comparing the frequencies of correct response on these altered items among the three groups, chi-square was used to compare the obtained frequencies of correct response to the frequencies expected on the basis of chance. Frequencies of

correct response to individual items 28-30 and 40-45 did not differ at the .01 level of significance from frequencies expected on the basis of chance for any of the three groups. It appears, then, that when compared to Group E--which received the original instructions--Groups P and C did not benefit by the changes on these items.

The second change in instructions involved items 31-33 and 38-39, and occurred when Group P was being tested. In this instance it appears that the change in instructions made a difference. Chi-square contingency tables were used to compare the frequencies of correct response for Groups E, P, and C for each of these five items. On three of the five items, Group P--which responded to the altered instructions--had significantly more correct responses at the .01 level than Group E--which responded to the original instructions. On a fourth item, Group P did significantly better at the .05 level.

Having determined that the change in instructions affected the responses of Group P to items 31-33 and 38-39, the analysis was carried one step further. An attempt was made to determine whether the affect was large enough to account for the fact that on the Picture test Group P did nominally better than Group E, while on the YES-NO, Matched-Pairs, and All-NO tests Group E did nominally better than Group P. On the five altered Picture items combined, Group P had 74 more correct responses than Group E. When these excess correct responses were divided by the number of students in Group P--74/59--the mean number of correct responses attributable to changing the instruction was 1.25. When the additional adjustment in analysis of covariance for initial differences in mental ability between groups was taken into account, the mean number of correct

responses attributable to changing the instructions increased to approximately 1.33. Since the difference between adjusted Picture means for Groups E and P was 1.51 (see Table 12) nearly all of the improvement in Group P's performance on the Picture test compared to the other tests can be accounted for by changes in instructions to the above mentioned items.

It should be noted that although the changes in instructions produced significant differences in the frequency of correct response to items 31-33 and 38-39, these changes did not produce a significant difference in Picture means. Groups E and P did not differ in Picture means at the .05 level of significance (see Table 12). This is consistent with previous findings, as they also did not differ in Matched-Pairs or All-NO means at the .05 level of significance (see Tables 8 and 9). In brief, the Picture test, like the Matched-Pairs test, produced the expected findings for the comparison of group means. E and P were both experimental groups and their means were not expected to be significantly different at the .05 level. Since C was the control group, its PET-1 means were expected to differ at the .01 level from the PET-1 means for Groups E and P, and they did (see Tables 8 and 13). The general validity of the Picture test can also be defended on the grounds that a comparatively large number of its items discriminated between groups. The only block of items which did not discriminate were those dealing with specialization. As will be noted again, this was to be expected since the items tested something the students were not taught. Of 49 items on the Picture test, 33 discriminated between control and experimental groups at the .05 level or better when differences in frequencies were tested using chi-square

contingency tables. Most of these discriminations were significant at the .01 level, with the chi-square values frequently being much larger than the 6.64 needed for significance. Twenty-one items produced chi-square values ranging from 10.00 to 46.73. Despite its limitations, the Picture test has high validity for the concepts "producer," "consumer," and "division of labor."

Most of the previously mentioned questions concerning the content validity of the Picture test were raised by the cooperating teachers in the EPC study. Their suggestions were generally well taken and would probably improve the validity of the Picture test. In a sense, however, there was one significant exception. All of the teachers in the experimental groups questioned the content validity of items 28-30 and 40-45. These items deal with the term "specialist." The objection to these items was that they test the concept in a manner dissimilar to the way it was taught.

Items testing the concept of specialization require the student to select one medical doctor who is a specialist from among four others who are not specialists, or to select one baker who is a specialist from among four who are not, etc. The students are taught, on the other hand, that all doctors, bakers, and school teachers are specialists. The result of this instruction appears to be that the students believe that there are only two categories--specialists and non-specialists. This presents some difficulties because people specialize to some degree in nearly every economic activity. Rather than a world populated by specialists and non-specialists, we have a world populated by people who have specialized in varying degrees. Although the items in question

do not validly represent the content of Families at Work, they do represent reality. Perhaps in this instance it would be better to change the content of the course of study.

CHAPTER V

STUDENTS' KNOWLEDGE OF THE CONTENT OF FAMILIES AT WORK

Investigations reported in this paper had two major thrusts: (1) test development, and (2) achievement assessment. Statistical analyses relevant to the first thrust--reliability and validity of the PET-1 instruments--were presented in the preceding chapter. Statistical analyses relevant to the second thrust--analysis of covariance for the comparison of PET-1 means between control and experimental groups, and chi-square values computed for the item analyses of PET-1 tests--are presented in the present chapter.

Because assessment of knowledge of concepts in Families at Work involved comparisons of PET-1 responses by control and experimental groups, the composition of the WOBE and EPC studies is summarized below.[1] Students in the WOBE study were selected from 14 similar schools in three adjoining school districts in Northern Utah. The experimental group in the WOBE study consisted of 96 students selected randomly from within the first-grade classrooms in seven schools in District W. The control group consisted of 100 students selected randomly from within the first-grade classrooms of four schools in District O and three schools in District BE. The students in the EPC study were chosen from three school districts; two in Northern Utah and one in Elkhart, Indiana. For reasons explained in Chapter III, students in the EPC study were not selected randomly.

---

[1] For a more detailed description of the procedure used in these studies, see Chapter III of this dissertation.

Instead, all of the students in three first-grade classes in each of Groups E, P, and C were tested. Another major difference between the EPC and WOBE studies is that all of the PET-1 tests were given to each of the EPC students, while only the YES-NO Matched-Pairs test was given to the WOBE students. All of the PET-1 tests were given to each of the students in the EPC study because that study was primarily designed to allow for comparisions of reliability and validity among the PET-1 tests. The WOBE study, utilizing random selection of students, was primarily designed to allow for comparison of PET-1 means. However, both studies provided useful information concerning the adequacy of the PET-1 tests--the central concern of Chapter IV, and both studies also provided useful information concerning the achievement of control and experimental groups of students--the central concern of the present chapter.

In assessing achievement of the control and experimental groups, analysis of the PET-1 scores was directed at the following questions:

1. Can samples of children from suburban and suburban-rural schools learn the content of Families at Work?
   a. Can they learn the content of Families at Work in general?
   b. If there are general indications that children can learn the content of Families at Work, to what extent do they learn specific concepts?
   c. Is the content of Families at Work too easy for bright children?
   d. Is it too difficult for slow children?
2. Is the achievement of students on the PET-1 tests dependent on the training or experience of their teachers?

Can First-Grade Children Learn the Content of FAMILIES AT WORK?

Were there general indications that children can learn the content of Families at Work?

Comparing PET-1 means of control and experimental groups could indicate whether one group performed better in general than the other, but could not indicate which specific terms or concepts were better learned. Nevertheless, comparing means is an important preliminary step; if there is no general difference between groups, as indicated by their PET-1 means, it is unlikely that they differ in knowledge of specific terms.

In the WOBE and EPC studies it was expected that Group W would score significantly higher than OBE, and that E and P would score higher than C. However, for statistical analysis Hypothesis 6 is stated in the null form and should not be taken as a research expectation.[2]

Hypothesis 6: There will be no significant difference in PET-1 means between control and experimental groups in the WOBE and EPC studies.

Analysis of covariance for the various PET-1 tests given to these groups are summarized in Tables 13 to 18. The last three tables are

---

[2]The reader may have noticed that not all hypotheses in this paper are testable by statistical procedures, and that not all hypotheses are stated in the null form. The use of non-statistical hypotheses is justified on the grounds that not all useful research questions require statistical tests of significance. Some, in fact, are not amenable to such tests. Of course, when tests of significance are not employed the null hypothesis is not required. Generally, throughout the present chapter, both research expectations and null hypotheses formulated to test those expectations will be stated, and an attempt will be made to distinguish between them for the reader.

identical to 7, 8, and 9 in the previous chapter. Remember that Group W in the WOBE study, and Groups E and P in the EPC study, are experimental.

Table 13. Analysis of covariance between Groups W and OBE on the YES-NO test

Group		TOGA[a]			YES-NO			Adjusted YES-NO[b]		
		M	SD	F	M	SD	F	M	SD	F
W	96	41.79	8.45	1.37	48.91	6.45	27.44	48.63	5.54	28.90
OEB	100	40.38			44.08			44.35		

df = 1/200 $F_{.05} = 3.89$ $F_{.01} = 6.76$

[a]TOGA is a mental abilities test and was used to adjust for initial differences between groups.

[b]This column gives the means, Standard Deviations, and F-ratios for the YES-NO test after adjustments were made for initial differences on the TOGA

Table 14. Analysis of covariance between Groups W and OBE on the Matched-Pairs test

Group		TOGA[a]			Matched-Pairs			Adjusted Matched-Pairs		
		M	SD	F	M	SD	F	M	SD	F
W	96	41.79	8.45	1.37	15.84	4.97	36.74	15.62	4.21	40.90
OBE	100	40.38			11.54			11.75		

df = 1/200 $F_{.05} = 3.89$ $F_{.01} = 6.76$

[a]See Table 13 for explanation of symbols.

Table 15. Analysis of covariance among Groups E, P, and C on the All-NO test

Group	N	TOGA M	TOGA SD	TOGA F	All-NO M	All-NO SD	All-NO F	Adjusted All-NO M	Adjusted All-NO SD	Adjusted All-NO F
E	77	44.83			47.10			43.85		
P	59	38.86	7.43	30.21	36.03	11.95	19.08	40.55	10.49	8.46
C	46	45.11			36.02			35.57		

	Differences between adjusted All-NO means	SE_D between All-NO means	t-ratios
E and P	3.30	2.52	1.31
P and C	4.98	2.94	1.69
E and C	8.28	2.73	3.03

df = 150, $F_{.05} = 3.06$ $F_{.01} = 4.75$, $t_{.05} = 1.98$ $t_{.01} = 2.61$

Table 16. Analysis of covariance among Groups E, P, and C on the YES-NO test

Group	N	TOGA M	TOGA SD	TOGA F	YES-NO M	YES-NO SD	YES-NO F	Adjusted YES-NO M	Adjusted YES-NO SD	Adjusted YES-NO F
E	77	48.83			55.49			53.70		
P	59	38.86	7.43	30.21	48.29	6.95	29.79	50.78	6.19	19.34
C	46	45.11			46.59			46.40		

	Differences between adjusted YES-NO means	SE_D between YES-NO means	t-ratios
E and P	2.92	1.49	1.96
P and C	4.38	1.73	2.53
E and C	7.30	1.61	4.52

df = 150, $F_{.05} = 3.06$ $F_{.01} = 4.75$, $t_{.05} = 1.98$ $t_{.01} = 2.61$

Table 17. Analysis of covariance among Groups E, P, and C on the Matched-Pairs test

Group	N	TOGA M	TOGA SD	TOGA F	Matched-Pairs M	Matched-Pairs SD	Matched-Pairs F	Adjusted Matched-Pairs M	Adjusted Matched-Pairs SD	Adjusted Matched-Pairs F
E	77	48.83			20.46			18.79		
P	59	38.86	7.43	30.21	15.14	5.84	24.00	17.46	5.06	16.03
C	46	45.11			13.65			13.47		

	Differences between adjusted Matched-Pairs means	SE_D between Matched Pairs means	t-ratios
E and P	1.33	1.21	1.10
P and C	3.99	1.42	2.81
E and C	5.32	1.32	4.03

df = 150, $F_{.05} = 3.06$ $F_{.01} = 4.75$, $t_{.05} = 1.98$ $t_{.01} = 2.61$

Table 18. Analysis of covariance among Groups E, P, and C on the Picture test

Group	N	TOGA M	TOGA SD	TOGA F	Picture M	Picture SD	Picture F	Adjusted Picture M	Adjusted Picture SD	Adjusted Picture F
E	77	48.83			24.84			23.81		
P	59	38.86	7.82	27.27	23.85	5.90	32.43	25.32	5.60	37.77
C	77	45.10			17.69			17.59		

	Differences between adjusted Picture means	SE_D between Picture means	t-ratios
E and P	1.51	1.34	1.13
P and C	7.73	1.34	5.77
E and C	6.22	1.23	5.06

df = 150, $F_{.05} = 3.06$ $F_{.01} = 4.75$, $t_{.05} = 1.98$ $t_{.01} = 2.61$

Of the ten comparisons testing Hypothesis 6--there will be no significant difference in PET-1 means between control and experimental groups in the WOBE and EPC studies--nine led to its rejection, and one did not. Furthermore, findings from both tests in the WOBE study--the better study--and from all three comparisons involving the Matched-Pairs test--the best test--led to the rejection of the null hypothesis. It is concluded that first-grade children can learn at least some of the content of Families at Work.

The following three paragraphs explain the findings summarized above.

In Tables 13 and 14, the F-ratio for adjusted means--28.09 and 40.90--exceed the ratio needed for significance at the .01 level--6.76. Thus, the null hypothesis was rejected for the YES-NO and Matched-Pairs tests in the WOBE study.

In the EPC study there were two experimental groups--E and P. In Tables 15 and 18 the differences between adjusted means for Groups E and C on the YES-NO, Matched-Pairs, All-NO, and Picture tests exceed the differences needed for significance at the .01 level. Thus, the null hypothesis was rejected for the comparisons between Groups E and C in the EPC study.

For the same comparisons between Groups P and C--Tables 15 to 18--the null hypothesis was rejected for the YES-NO, Matched-Pairs, and Picture tests. It was not rejected for the All-NO test. The difference between adjusted All-NO means is not significant at the .05 level. The difference between YES-NO means is significant at the .05 level. The differences between Matched-Pairs means and between Picture means are significant at the .01 level.

Thus the expectation was supported; PET-1 means for children instructed with Families at Work were significantly higher than for those who were not.

To what extent did they know or learn the specific concepts?

Were there concepts for which children in the experimental groups failed to exhibit knowledge? It was assumed that this question could be investigated by asking the more testable question, "Were there items[3] on which the frequency of correct response by children in the experimental groups was not higher than expected by chance?"

Were there concepts for which children in the control groups exhibited knowledge? It was assumed that this question could be investigated by asking the more testable question, "Were there items on which the frequency of correct response by children in the control groups was higher than expected by chance?"

Were there concepts for which children in the experimental groups made gains in knowledge? It was assumed that this question could be investigated by asking, "Were there items on which the frequency of correct response by children in the experimental groups was higher than for children in the control groups?"

In other words, two types of item analysis were used to answer the questions listed above: (1) Comparison of the responses of control and experimental groups to individual items, and (2) Comparison of responses

---

[3]The same question could be asked for clusters of similar items. However, asking the question in that form could obscure important differences between items which are assumed to be similar.

of either group on individual items to responses expected by chance.

It was expected that:

1. The experimental groups would have more correct responses than the control groups on a number of items, but no exact expectation was formulated.

2. The control groups would not have more correct responses than the experimental groups on any items.

3. The control groups would have more correct responses than expected by chance on some items. Some items test knowledge that a young child would obtain even if he did not study Families at Work. Other items may clue the student to the correct response.

4. The experimental groups would have more correct responses than expected by chance on more items than the control groups.

However, for statistical analysis Hypotheses 7 and 8 are stated in the null form; they should not be taken as research expectations.

Hypothesis 7: There will be no significant difference between control and experimental groups on frequency of correct response to individual items.

Hypothesis 8: For either group, there will be no significant difference between observed and expected frequencies of correct response to individual items.

Chi-square was used to test both hypotheses. However, chi-square cannot be used to correct for initial differences between groups. Therefore, it was necessary to determine whether W was similar to OBE in mental ability, and whether E, P, and C were similar in mental ability. The F-ratio for the difference between TOGA means for W and OBE--Table 13--

is not significant at the .05 level. They were thus comparable in initial mental ability. TOGA means and F-ratios in Table 15 indicate that Groups E, P, and C were not comparable in initial mental ability. Nevertheless, E and C were made comparable by removing the scores for one of the classes in Group E. The t-ratio for the difference in TOGA means between E and C was then .84, compared to 1.98 needed for significance at the .05 level.

The Matched-Pairs test was used in the following comparisions because it controls best for acquiescence, has the lowest probability of correct response by chance, and was given to groups in both studies. The data necessary for testing Hypotheses 7 and 8 are contained in Table 19. W and E are the experimental groups. Following each item is the frequency of correct response for each group. Chi-square values are underlined. "Insp" means "not significant by inspection."

Table 19. Items on the Matched-Pairs test ranked according to frequency of correct response, with chi-square values for group comparisons

Levels of significance for chi-square: $P_{.05} = 3.84$, $P_{.01} = 6.64$.
Number of students: W = 96, OBE = 100, E = 58, C = 48.
Frequencies significantly larger than those expected by chance:
W = 36, OBE = 37, E = 24, C = 21.

4. We must have food. (YES)
41. We can get along without food. (NO)
(W = 83, OBE = 83, insp) (E = 50, C = 42, insp)

6. Eskimos, Bushmen, and Indians live in different kinds of houses. (YES)
43. Eskimos, Bushmen, and Indians live in the same kind of houses. (NO)
(W = 79, OBE = 77, insp) (E = 51, C = 40, insp)

16. Tools and machines make it harder to do work. (NO)
53. Tools and machines make it easier to do work. (YES)
(W = 72, OBE = 50, 12.26) (E = 50, C = 23, 17.96)

28. Income is money people get for doing work. (YES)
65. "Income" means "Come in the house." (NO)
(W = 77, OBE = 37, 36.27) (E = 43, C = 26, 4.61)

32. Some families save part of their income. (YES)
69. Every family spends all of its income. (NO)
(W = 70, OBE = 52, 8.49) (E = 48, C = 19, 21.05)

30. If two stores sell things that are just alike, the store with the lowest prices will usually have more customers. (YES)
67. If two stores sell things that are just alike, the store with the highest prices will usually have more customers. (NO)
(W = 69, OBE = 48, 10.93) (E = 33, C = 27, insp)

10. When people shovel snow off the sidewalk they are producing a service. (YES)
47. When people shovel snow onto the sidewalk they are producing a service. (NO)
(W = 70, OBE = 43, 17.12) (E = 30, C = 29, insp)

26. Specialists usually do their work away from home. (YES)
63. Specialists usually stay home to do their work. (NO)
(W = 65, OBE = 44, 10.55) (E = 34, C = 30, insp)

Table 19. (continued)

Levels of significance for chi-square: $P_{.05} = 3.84$, $P_{.01} = 6.64$.
Number of students: W = 96, OBE = 100, E = 58, C = 48.
Frequencies significantly larger than those expected by chance:
W = 36, OBE = 37, E = 24, C = 21.

12. When Mother washes the dishes and Sister dries them they are dividing the labor. (YES)
49. When Mother and Sister watch T.V. they are dividing the labor. (NO)
(W = 64, OBE = 40, 13.33) (E = 42, C = 21, 8.95)

17. Father would usually save money if he stayed home from work to wash the car. (NO)
54. Father would usually lose money if he stayed home from work to cut the grass. (YES)
(W = 60, OBE = 45, 6.03) (E = 30, C = 23, insp)

14. When Brother sweeps the floor and Sister makes the bed they are dividing the labor. (YES)
51. When two babies are playing with dolls they are dividing the labor. (NO)
(W = 52, OBE = 26, 15.69) (E = 43, C = 17, 16.03)

5. We must have T.V. (NO)
42. We can get along without T.V. (YES)
(W = 48, OBE = 33, 5.53) (E = 35, C = 19, 4.53)

8. A farmer who raises potatoes is a producer of goods. (YES)
45. A farmer who raises weeds is a producer of goods. (NO)
(W = 48, OBE = 36, 3.66) (E = 20, C = 27, 5.04 favors C)

9. Children who jump rope are producers. (NO)
46. Children who wash dishes are producers. (YES)
(W = 48, OBE = 38, 2.86) (E = 27, C = 17, insp)

13. It is faster and cheaper to divide the labor. (YES)
50. It is faster and cheaper for one man to produce all of his own goods. (NO)
(W = 42, OBE = 27, 5.75) (E = 35, C = 18, 5.48)

22. Rich people want more things than they can have. (YES)
59. Rich people can have everything they want. (NO)
(W = 43, OBE = 27, 6.46) (E = 26, C = 16, insp)

25. A specialist knows how to do one job very well. (YES)
62. A specialist knows how to do many different kinds of jobs very well. (NO)
(W = 37, OBE = 10, 21.48) (E = 26, C - 12, 4.49)

Table 19. (continued)

Levels of significance for chi-square: $P_{.05} = 3.84$, $P_{.01} = 6.64$.
Number of students: W = 96, OBE = 100, E = 58, C = 48.
Frequencies significantly larger than those expected by chance:
W = 36, OBE = 37, E = 24, C = 21.

23. Customs and rules help us to know what other people will do. (YES)
60. Customs and rules make it hard to know what other people will do. (NO)
(W = 34, OBE = 26, 1.90) (E = 33, C = 17, 4.86)

20. If we worked harder we could have everything we want. (NO)
57. People who work very hard still want more things than they have. (YES)
(W = 34, OBE = 32, insp) (E = 32, C = 15, 6.09)

18. We have more free time because we divide the labor. (YES)
55. People who divide the labor have very little free time. (NO)
(W = 35, OBE = 30, insp) (E = 29, C = 18, insp)

33. Banks loan money to anyone who needs it. (NO)
70. Banks loan money only to people who will pay it back. (YES)
(W = 35, OBE = 24, 3.43) (E = 22, C = 11, 3.76)

27. Transportation makes it harder for specialists to trade their goods and services. (NO)
64. Transportation makes it easier for specialists to trade their goods and services. (YES)
(W = 35, OBE = 27, 1.88) (E = 20, C = 11, insp)

11. Everyone except babies and sick people is a producer. (YES)
48. Everyone is a producer. (NO)
(W = 32, OBE = 32, insp) (E = 30, C = 9, 10.90)

37. When people stop buying goods, more businesses are started. (NO)
74. When people buy many goods, more businesses are started. (YES)
(W = 31, OBE = 24, insp) (E = 24, C = 15, insp)

7. Everyone in the family is a consumer. (YES)
44. Mother and Father are the only consumers in the family. (NO)
(W = 29, OBE = 28, insp) (E = 35, C = 9, 18.72)

19. Most pioneers lived in cities. (NO)
56. Most pioneers lived on farms. (YES)
(W = 29, OBE = 22, 1.60) (E = 24, C = 14, insp)

Table 19. (continued)

Levels of significance for chi-square: $P_{.05} = 3.84$, $P_{.01} = 6.64$.
Number of students: W = 96, OBE = 100, E = 58, C = 48.
Frequencies significantly larger from those expected by chance:
W = 36, OBE = 37, E = 24, C = 21.

2. Almost every family in the world has a T.V. (NO)
39. In some places only a few families have T.V. (YES)
(W = 29, OBE = 18, 4.01) (E = 16, C = 12, insp)

34. Profit is money the businessman gets for worrying. (YES)
71. Profit is money the workers get for worrying. (NO)
(W = 28, OBE = 30, insp) (E = 9, C = 9, insp)

29. When many people try to get the same job the wages will usually be lower. (YES)
66. When many people try to get the same job the wages will usually be higher. (NO)
(W = 25, OBE = 24, insp) (E = 23, C = 7, 8.14)

36. When people buy more goods, more workers have jobs. (YES)
73. When people buy fewer goods, more workers have jobs. (NO)
(W = 25, OBE = 18, insp) (E = 19, C = 10, insp)

31. Our schools are not usually paid for by taxes. (NO)
68. Our schools are usually paid for by taxes. (YES)
(W = 23, OBE = 19, insp) (E = 26, C = 13, 3.56)

21. Pioneers are people who live in a different country. (NO)
58. Pioneers lived a long time ago. (YES)
(W = 24, OBE = 18, insp) (E = 6, C = 8, insp)

1. Your brothers or sisters are part of your close family. (YES)
38. Your mother and father are part of your distant family. (NO)
(W = 22, OBE = 4, 15.01) (E = 22, C = 4, 12.43)

15. Nations who trade with each other divide the labor. (YES)
52. Nations who trade with each other do not divide the labor. (NO)
(W = 22, OBE = 29, 1.02) (E = 22, C = 14, insp)

24. A specialist depends on others to produce the things he needs. (YES)
61. A specialist produces for himself everything he needs. (NO)
(W = 22, OBE = 11, 4.82) (E = 19, C = 15, insp)

Table 19. (continued)

Levels of significance for chi-square: $P_{.05} = 3.84$, $P_{.01} = 6.64$.
Number of students: W = 96, OBE = 100, E = 58, C = 48.
Frequencies significantly larger than those expected by chance:
W = 36, OBE = 37, E = 24, C = 21.

35. Before he can go into business a man needs a wife, a car, materials and workers. (NO)
72. Before he can go into business a man needs materials, workers, tools, and a workplace. (YES)
(W = 20, OBE = 8, 6.43) (E = 25, C = 7, 10.14)

3. Almost every family in the world has a telephone. (NO)
40. In some places only a few families have telephones. (YES)
(W = 16, OBE = 18, insp) (E = 21, C = 8, 5.05)

75. A businessman who sells a vacuum for $40 makes a $40 profit. (NO)
(There was no reversal for this item.)
(W = 30, OBE = 34, insp) (E = 25, C = 15, insp)

Hypotheses 7 and 8 are applicable to each of the 37 pairs of items in Table 19. If the .05 level of significance is accepted, then Hypothesis 7 was rejected each time either of the two chi-square values for each item exceeded 3.84. For the .01 level of significance, chi-square must be 6.64. Hypothesis 8 was rejected each time the observed frequency of correct response was either larger or smaller than the frequency expected by chance. The frequencies expected on the basis of chance are: W = 24, OBE = 25, E = 15, and C = 12. If the .01 level of significance is accepted, then Hypothesis 8 was rejected each time the frequency of correct response in any group reached the following levels: W = 12 or 36, OBE = 13 or 37, E = 6 or 24, and C = 3 or 21. At the .05 level of significance the frequencies were: W = 15 or 33, OBE = 16 or 34,

E = 8 or 22, and C = 5 or 19. For example, if the frequency of correct response by Group C on any item was 5 or less that frequency was smaller than expected by chance. If it was 19 or larger it was greater than expected by chance.

Table 19 is useful because the reader has the content of the items before him when testing for Hypotheses 7 and 8. On the other hand, it is inconvenient because it is too long for the reader to determine the total pattern of acceptance or rejection of hypotheses. For that reason, Tables 20 and 21 summarize the number of times Hypotheses 7 and 8 were rejected.

Table 20. Number of pairs of items for which the frequencies of correct response significantly differed between groups

Levels of Significance	W and OBE	E and C	Both W and OBE, and E and C	Either W and OBE, or E and C, or both
.05	17[a]	16[b]	10[c]	23[d]
.01	11	9	5	14

[a]Since there were 37 pairs of items this entry should be read, "For the comparison between Groups W and OBE, the frequencies of correct response significantly differed at the .05 level for 17 of 37 pairs of items."

[b]Read the same as for the comparison between W and OBE. One other pair of items also was significant at the .05 level, but favored the control group.

[c]Although 17 items discriminated at the .05 level between Groups W and OBE, and 16 discriminated between E and C, only 10 items discriminated at the .05 level between both W and OBE, and E and C.

[d]Although only 10 items discriminated between both W and OBE, and E and C, 23 items discriminated between either W and OBE, or E and C, or both. In other words, if all possible comparisons were taken into account, there were 23 discriminating items.

Table 21. Number of pairs of items for which the frequencies of correct response significantly differed from expectation

W[a] .01		W .05		OBE .01		OBE .05		E[a] .01		E .05		C .01		C .05	
12	36	15	33	13	37	16	34	6	24	8	22	3	21	5	19
0	17	0	22	3	1	3	12	1	25	1	29	1	10	2	12

[a]Groups W and E are experimental.

The summaries in Tables 20 and 21 indicate that Hypotheses 7 and 8 were rejected a number of times. As reported in Table 20, 23 of 37 pairs of items discriminated at the .05 level between either Groups W and OBE, or E and C. Therefore, Hypothesis 7 was rejected for 23 of 37 pairs of items in the Matched-Pairs test. As reported in Table 21, the control groups produced frequencies of correct response greater than those expected by chance on 12 of 37 pairs of items; the experimental groups did so on 31 of 37 pairs of items at the .01 level of significance. Thus Hypothesis 8 was rejected 12 of 37 times in the control group, and 31 of 37 times in the experimental group.

In general, the four expectations listed just prior to Hypotheses 7 and 8 were supported. The lone exception is listed in the footnote to Table 20. Expectation 2--the control groups would not have more correct responses than the experimental groups on any items--is violated in Item 8:45. However, this is not considered a serious exception to the expectation. The content of the item focuses on farmers; a subject the children in suburban Group E would know less about than the children in rural-suburban Group C. That the expectation was supported for the

comparison between Groups W and OBE for the same item indicates that this assumption is probably correct.

Referring again to Table 20, 10 of the 23 discriminating items discriminated in both comparisons. That is, seven of the items that discriminated between Groups W and OBE did not discriminate between Groups E and C, and six items that discriminated between Groups E and C did not discriminate between W and OBE. It is assumed that this pattern would have continued if more experimental and control groups had been compared; even more items would have discriminated in one comparison or another. Furthermore, three pair of items in Table 19 are nearly significant at the .05 level, but are not included in the total of 23 discriminating items. Also, some items which produced large chi-square values for the comparison between groups in pilot studies undertaken earlier in the year did not do so in either the WOBE or EPC studies. Whether items discriminate appears to be a function of differences in the groups that are tested and the time of year when the testing is done. It is likely that more concepts are learnable than is indicated by any specific testing.

Two additional items, for which neither experimental group produced frequencies of correct response greater than those expected by chance, discriminated between control and experimental groups. Item 3:40 produced a chi-square value of 5.05 between Groups E and C--3.84 is significant at the .05 level. For this item, the frequency of correct response in Group E was 21; one less than that needed to significantly differ from chance at the .05 level. Item 2:39 produced a chi-square value of 4.01 between Groups W and OBE. For this item, the frequency

of correct response in Group W was 29; four less than that needed to significantly differ from chance at the .05 level.

The findings reported in the above paragraph are unusual, but can be explained in terms of ethnocentrism in young children. The pairs of items--2:39 and 3:40--are similar. Both test knowledge of whether people in other countries have as many telephones or television sets as people in the United States. Most young children believe that nearly every family in the world has household items that are common in this country. Because this is a positively held belief, the frequency of correct response is not significantly larger than expected on the basis of chance. However, a statistically significant number of students in the experimental groups apparently remembered this part of the content of Families at Work. Although the frequency of correct response was not large for either group, it was significantly larger for the experimental group than for the control groups. It is also interesting to note that in preliminary studies conducted earlier in the year, more than half of the students in the experimental group correctly responded to items similar to 2:39 and 3:40. Apparently, a considerable amount of forgetting occurs between November and May, as would be expected.

To summarize the two preceding pages, when the data from both types of item analysis is considered, evidence that a statistically significant number of students possessed knowledge related to the concepts being tested is absent for only 4 of 37 pairs of items. And for three of those four items the frequency of correct response between the experimental group and the control group is nearly significant at the .05 level.

Two different types of item analysis were used in the foregoing discussion. The first analysis compared the frequencies of correct response to individual items between control and experimental groups, was related to Hypothesis 7, and is summarized in Table 20. The second analysis compared the frequency of correct response to individual items to the frequency expected by chance, was related to Hypothesis 8, and is summarized in Table 21. To further probe the data relevant to the question--To what extent did the children know or learn the concepts in Families at Work?--these two item analyses were combined into a matrix. This approach allows the results of both analyses to be examined simultaneously for any item. With modification, it also allows for the items to be rated in nine categories, rather than simply dichotomized. In short, it allows for a more detailed examination of the data.

Table 22. Matrix for the two dimensional item analysis of the Matched-Pairs test

First Dimension: Magnitude of the frequency of correct response	Second Dimension: Discrimination between groups[a] High (.01 or above)	Moderate (.10)	Low (less than .10)	
High	H-H[b] 16:53[c], 28:65 32:69, 30:67 10:47, 26:63 12:49, 14:51 (8)[d]	H-M (0)	H-L 4:41, 6:43 (2)	(10)
Moderate	M-H 25:62, 11:48 7:44, 29:66 35:72, (5)	M-M 17:54 5:42, 13:50 8:45, 22:59 23:60, 20:57 33:70, 31:68 (10) 9:46	M-L 27:64, 37:74 19:56 18:55 (4)	(19)
Low	L-H 1:38 (1)	L-M 24:61, 3:40 2:39 (3)	L-L 34:71, 36:73 21:58, 15:52 (4)	(8)
	(14)	(13)	(10)	(37)

[a]When the results of the comparison of W to OBE did not agree with the comparison of E to C, then the results that gave the item the highest rating were used. This decision was based on the preceding argument, "Whether items discriminate appears to be a function of differences in the groups that are tested and the time of year when the testing is done. It is likely that more concepts are learnable than is indicated by any specific testing."

[b]H-H stands for High-High. The other capitalized letters also represent the row-column intersection.

[c]The numbers on either side of the colon stand for the paired items.

[d]The numbers in parentheses stand for the entries in each box. Those in the margins stand for the total entries in the rows and columns.

For the matrix, three levels of performance were established in each of two dimensions. The first dimension is the magnitude of the frequency of correct response to individual items, and the second dimension is the ability of individual items to discriminate between control and experimental groups of students. Levels of performance in either dimension were labeled "Low," "Moderate," and "High."

In the first dimension, the frequency of correct response needed to significantly differ from chance at the .05 level was accepted as the upper limit of the Low category. The remaining possible frequencies of correct response were divided equally to establish the limits of Moderate and High. The .05 level was chosen over two other alternatives in establishing the upper limits of the Low category. The first alternative was to divide the total possible frequency of correct response in each of the experimental groups into thirds. This approach was rejected because it would place in the Moderate category too many frequencies which did not significantly differ from chance in Group E. The expected frequency of correct response for E is 15, one-third of the total possible responses is 17, and a frequency of 22 correct responses would significantly differ from chance at the .05 level. Thus, the frequencies 18-22 would be labeled Moderate even though they do not significantly differ from that expected by chance. The second alternative was to set the upper limits of Low at the .01 level of significance. This approach was rejected because it placed too large a proportion of the possible responses into the Low category--41 percent in Group E. The .05 level was accepted as a practical compromise;

it shares the strong points of each of the other alternatives. The .05 level is a reasonably strict level of significance, yet does not depart as far as the .01 level does from the division of total frequencies into thirds.

In the second dimension, items which discriminated between control and experimental groups at the .01 level or better were placed in the High category. Items which discriminated at or above the .10 level but less than .01 were placed in the Moderate category. All others were classified as Low.

In establishing the categories for the second dimension, little difficulty was experienced in deciding upon the .01 level of significance as the cut-off point between High and Moderate. It was decided that items should not be classified as High discriminators unless the probability was small that the discrimination was a chance occurrence. However, some difficulty was experienced in deciding whether to use the .05 or the .10 level of significance to separate Moderate from Low. In this case, the .05 level was considered to be too strict. It was expected that even if the cut-off point were placed at .10, the items classified as Low, if taken together, would discriminate between control and experimental groups. As long as the items in the Low category, when taken together, discriminated between groups there was some justification for claiming that the content of those items was learned by some students. It was decided to reserve the Low category for concepts for which instruction seemed to have minimal effect on the experimental groups. Raising the cut-off point to the .05 level of significance would violate the standard of "minimal effect."

The sign test was used to check this expectation[4]--that items which did not discriminate at the .10 level when taken separately, would discriminate at the .05 level when taken together. Siegel (1956, pp. 68-75) states that the only assumptions made in using the sign test are: (1) that the variable being tested is continuous, and (2) that the groups being compared are alike. Concerning the first requirement, while frequencies of correct response are discrete, the basis for the response--student's knowledge--is continuous. Concerning the second requirement, since Groups W and OBE, and Groups E and C, produced TOGA means which did not differ at the .05 level of significance, it was concluded that they are comparable.

The sign test was computed in the following manner. Frequencies of correct response were inspected for control and experimental groups for each of the items in question. If the frequencies of correct response to a given item were apparently greater in the experimental group a plus (+) was recorded. If they were greater in the control group a minus (-) was recorded. When frequencies of correct response was greater for neither group a zero (0) was recorded. The significance of the ratio of pluses to minuses was then determined by reference to a standard table for the sign test (Siegel, 1956, p. 250).

The sign test was significant: (1) at the .01 level for the comparison between Groups W and OBE, (2) at the .05 level for the comparison between E and C, and (3) at the .001 level when both experimental groups

---

[4]This expectation could have been checked by adding chi-square values. However, chi-square values had not been computed for most of the items involved. Rather the frequencies of correct response between control and experimental groups had been declared "not significant by inspection."

were compared to both control groups. Therefore, since those items which did not discriminate at the .01 level when taken separately discriminated when taken together, the cut-off point between Moderate and Low was set at the .01 level of significance in the second dimension.

For the reader's convenience, each of the two dimensions is here defined again. The first dimension was based upon the magnitude of the frequency of correct response to individual items by the experimental groups. The second dimension was based upon the ability of individual items to discriminate between the control and experimental groups. Each dimension was divided into three categories. By combining categories from both dimensions, nine classifications were established: High-High, High-Moderate, High-Low, Moderate-High, Moderate-Moderate, Moderate-Low, Low-High, Low-Moderate, and Low-Low.

Entries in the nine cells of the matrix are to be read in the following way. Each cell is labeled with two letters such as H-H or M-L. The first letter identifies the rating given that cell in the first dimension. The second letter identifies the rating in the second dimension. Thus, H-H means that the frequency of correct response to an item by the experimental groups was high, and that the item produced a large chi-square value when frequencies of correct response were compared between groups. In other words, H-H can be interpreted as meaning that a large proportion of students possessed knowledge related to these items, and that instruction in the concepts upon which they were based appears to have been effective. Likewise, L-H can be interpreted as meaning that a small proportion of students possessed knowledge related

to these items, but that instruction in the concepts upon which they were based appears to have been effective.

It would be difficult to improve upon students' responses to those items that appear in cells H-H or H-L. In the former cell, a large proportion of students in the experimental groups correctly responded to the items, and did so in significantly greater numbers than did students in the control groups. In the H-L cell, a large proportion of students in all groups correctly responded to the item; effects of instruction, if any, could not be demonstrated since the control students also performed well.

Teachers who are using the Senesh materials and who are interested in improving instruction related to Families at Work will want to give their closest attention to the content of cells L-L, M-L, L-M, and M-M in that order. Student performance on the items in these cells was less than High in both dimensions, which may indicate that the best opportunity for improvement is connected with the content of these items. Therefore, the following non-statistical analysis focused on those four cells.

The analysis took the following form: (1) The content of each item in these four cells was examined and other items were isolated which had similar content. (2) The location in the matrix of these similar items was noted, with special attention paid to those similar items in higher rated cells. (3) It was assumed that if a concept appeared in an item in a cell with a High rating, then children were capable of learning that concept. (4) When a concept appeared in similar items with dissimilar ratings, the content of the items was examined more closely to

determine why children responded differently to them. Findings from this analysis are presented below.

Two important trends emerged from the examination of the four cells which are less than High in either dimension: (1) Items in cells L-L and M-L were frequently similar to one or more other items with higher ratings. This was true for three out of four items in L-L, and for two out of four items in M-L. Those items which were similar to one or more other items in a higher cell were frequently more complex. They either illustrated the concept in a setting further removed from the student's experience, or they compounded several concepts into one item. In order to recheck this analysis, items in cells H-H and M-H were examined to see if they were similar to items in lower cells. This was true for eight of thirteen entries. Again, items in the higher cells appeared to be less complex or abstract than those on which the students' performance was rated Moderate or Low. (2) Items in cells L-M and M-M were infrequently similar to one or more other items with higher ratings. Eight of thirteen items in these cells were either similar to no other item, or were similar to only one other item and that item was in the same cell. However, if the test is valid in proportion of representation of concepts, then the content of those items which are not similar to other items probably received less emphasis in instruction. In that case, a higher than Moderate performance by the students on those items is not to be expected.

A more definite answer can now be given to the question that prompted this part of our inquiry--To what extent did the children know or learn the specific concepts in Families at Work? Aside from those few concepts

which were tested in only one item, student performance was rated Moderate or High for at least one item for each concept. Even those items on which student performance was rated Low discriminated between control and experimental groups when the items were taken together. This study failed to demonstrate that any concepts in Families at Work were too difficult for first-grade children, at least at a minimum level of complexity or abstractness of application.

Is the content of FAMILIES AT WORK suited to either above average or below average children?

In this section, the third and fourth questions raised in the introduction to this chapter are considered--Is the content of Families at Work too easy for bright children? Or too difficult for slower ones?

These questions can be partly answered by examining the matrix in Table 22. Totals in the right hand margin indicate that High, Moderate, and Low proportions of students correctly responded to 10, 19, and 8 items. It seems reasonable to conclude that if the test validly represents the content of Families at Work, then this course of study contains concepts appropriate to the ability of able, average and slow students.

The conclusion, however, goes beyond the data. It may be that above average students obtain perfect or near perfect scores on the Matched-Pairs test; Families at Work being too easy for them. Likewise, below average students may be unable to learn any of the concepts in Families at Work; obtaining scores no better than expected by chance. Hypotheses 9 and 10 are directed at these possibilities. Although both

hypotheses are stated in the null form, only Hypothesis 10 required a statistical test of significance. Hypothesis 9 is stated in the null form because it is consistent with the research expectation.

Hypothesis 9: No student will obtain a perfect score on the Matched-Pairs test.

Hypothesis 10: There will be no significant difference between control and experimental groups in the number of below average students who obtain scores significantly greater than expected on the basis of chance.

Hypothesis 9 was tested by inspection. Table 23 gives the ten highest possible scores on the Matched-Pairs test, and the number of students who obtained each.

Table 23. Ten highest possible scores on the Matched-Pairs test

Group[a]	37	36	35	34	33	32	31	30	29	28
W	0	0	0	0	0	0	0	0	0	1
E	0	0	0	0	1	0	3	3	3	3
P	0	0	0	0	0	0	0	0	0	0

[a]W is the experimental group in the WOBE study. E and P are the experimental groups in the EPC study. If Hypothesis 9 was not rejected in the experimental groups, it is unlikely that it would be rejected in the control groups. Results from control groups, therefore, are not given.

The TOGA mean for the fourteen students in Table 23 is 54.43, which is equivalent to a grade expectancy of 3.7. They were, therefore, above average in mental ability. Although these students were above average in ability, none of them produced a perfect or near perfect Matched-Pairs score. The highest score--33--is less than 90 percent of the total possible. Furthermore, of the experimental groups, the TOGA means for W is nearest grade expectancy, and the best student in that group received a Matched-Pairs score nine less than perfect. His Matched-Pairs score was less than 76 percent of the total possible.

It is concluded that the content of Families at Work is not too easy for above average students. Bright first-grade children should find concepts in the Senesh program which challenge their ability.

Hypothesis 10 required three tests of significance in order to determine: (1) which students could reasonably be termed "below average," (2) the minimum Matched-Pairs score which is above that expected by chance, and (3) whether more below average experimental students than control students obtained a larger-than-chance Matched-Pairs score. Chi-square was used for each test of significance, with the level of significance set at .05 in each case.

1. The phrase "below average students" was defined to mean "those students who scored significantly lower than grade-level on the TOGA." Grade-level for Groups E and C required a TOGA score of 43, and since some TOGA's were given earlier than others, grade-level for Groups P, W, and OBE required a score of 42. The largest scores significantly lower than 42 and 43 were 34 and 35, which are equivalent to the 1.0 and 1.1 grade-levels. Grade-levels at the time TOGA's were administered

should have been 1.7 and 1.8, therefore, the best students in the below average category were approximately 6 months below grade-level in mental ability. Moreover, the TOGA mean for the below average students was approximately 29; lower even than that required at the 1.0 grade-level.

2. Since the Matched-Pairs test contains 37 four-option items, 9.25 was the expected chance score. The lowest score significantly larger than 9 was 15.

3. Therefore, in order for Hypothesis 10 to be rejected the number of students who scored less than 36 or 35 on the TOGA and who scored higher than 14 on the Matched-Pairs test had to be significantly larger in the three experimental groups than in the two control groups. Table 24 summarizes the chi-square test for this comparison.

Table 24. Comparison of scores above and below chance on Matched-Pairs test for students who scored below average on the TOGA

Group	TOGA mean	Matched-Pairs Test above chance	Matched-Pairs Test below chance	Total	Chi-square
Experimental	29.05[b]	14	25	39	8.74[a]
Control	28.70	1	29	30	
		15	54	69	

[a] .01 = 6.64, chi-square was computed using Yates Correction.

[b] Since the TOGA means for the two groups of below average students were 29.05 and 28.70, they did not differ in initial ability and are comparable.

Hypothesis 10 was rejected. Fourteen below average students in the experimental group and one student in the control group had Matched-Pairs scores larger than expected by chance. When these frequencies were tested for independence, the chi-square value was 8.74, compared to 6.64 needed for significance at the .01 level.

Below average students correctly responded to a statistically significant number of items on the Matched-Pairs test. It is therefore concluded that they are capable of learning some of the concepts in Families at Work.

Summary of conclusions in response to the first four questions raised in this chapter

1. Can children learn in general the concepts found in Families at Work? They can. Of ten comparisons of PET-1 means between the control and experimental groups, nine favored the experimental groups, and one did not. Furthermore, findings from both tests in the WOBE study--the better study--and from all three comparisons involving the Matched-Pairs test--the best test--favored the experimental groups.

2. To what extent can first-grade children learn the specific concepts? We were able to find no concept that some children were not able to learn in some form. When the data from both types of item analysis were considered, evidence that a statistically significant number of students possessed knowledge related to the concepts being tested was absent for only 4 of 37 pairs of items. When both item analyses were combined into a matrix, it appeared that the concepts in Families at Work are well suited to the ability of most students, at

at least a simple level of complexity and application. Aside from those few concepts which were tested in only one item, student performance was rated Moderate or High for at least one item for each concept. Even those items on which student performance was rated Low discriminated between the control and experimental groups when the items were taken together.

3. Is the content of Families at Work too easy for bright children? No. The fourteen students with the ten highest possible scores missed from ten percent to twenty-four percent of the items on the Matched-Pairs test. Thirteen of these students were in Group E, which judging from their TOGA mean was a very bright group of first-grade children.

4. Is the content of Families at Work too difficult for the slower students. It is not completely beyond their ability. When the least able students in the experimental groups were compared with the least able students in the control groups, significantly more experimental group students scored higher than expected by chance on the Matched-Pairs test--.01 level. It was concluded that they are therefore able to learn some of the content of Families at Work.

Is experience or special training needed to teach FAMILIES AT WORK?

Although the initial thrust of the investigations reported in this dissertation was to determine whether first-grade children could learn the content of the Senesh materials, those investigations provided the framework for considering other important questions. The question heading this section has been asked repeatedly by teachers and administrators

who have considered adopting the Our Working World series.[5] An answer to this question was pursued by comparing optimal and ordinary learning environments for Families at Work.

Group E of the EPC study was judged an optimal learning environment because Families at Work was developed with the cooperation of the teachers in Group E--Elkhart, Indiana. Furthermore, school authorities in Elkhart were asked to select three of their best first-grade teachers for inclusion in this study.

Group P was judged to be between an optimal and ordinary learning environment. It is comparable to Group E in that the teachers were judged by their supervisors to be among the best in the district. However, prior to 1966-67 they had only a half-year experience with Families at Work, and had not received special inservice training in economic education.

Group W in the WOBE study was judged to be an ordinary learning environment because students were selected randomly from within 24 classrooms in seven schools in District W. First-grade teachers in this district received no special in-service training in economic education, and prior to 1966-67 they had only a half-year experience with Families at Work.

It was expected that students in optimal learning environments would score higher on the PET-1 tests than would students in the ordinary learning environments. However, for statistical analysis the following hypothesis is stated in the null form.

---

[5]Private conversations with Joseph Rueff and others.

Hypothesis 11: There will be no significant difference in PET-1 means between first-grade children who are instructed in optimal learning environments and those who are instructed in ordinary learning environments.

Two tests were common to Groups E, P, and W--the YES-NO test and the Matched-Pairs test. Since the Matched-Pairs test has greater reliability and validity, it was used in analysis of covariance to test Hypothesis 11. The findings are summarized in Table 25.

Table 25. Analysis of covariance among Matched-Pairs means for E, P, and W

Group	N	TOGA M	TOGA SD	TOGA F	Matched-Pairs M	Matched-Pairs SD	Matched-Pairs F	Adjusted Matched-Pairs M	Adjusted Matched-Pairs SD	Adjusted Matched-Pairs F
E	77	48.83			20.46			18.28		
P	59	38.86	8.05	28.78	15.15	5.80	18.46	16.95	4.84	2.60[a]
W	96	41.79			15.84			16.48		

	Differences between adjusted Matched-Pairs means	SE_D between Matched-Pairs means	t-ratios
E and P	1.33	1.16	1.14
P and W	.47	1.11	.41
E and W	1.80	1.02	1.76

$t_{.05}$= 1.97, $t_{.01}$= 2.60, $F_{.05}$= 3.04, $F_{.01}$= 4.71.

[a]Usually when the adjusted F-ratio is not significant, pairs of groups are not compared. However, since Groups E and P are compared in Tables 19-21, pairs of groups are also compared here for the benefit of the reader who may want to see how close the differences came to being significant.

Although it was expected that Group P would score significantly higher than Group W, and that Group E would score higher than either P or W, these expectations were not supported. Findings in Table 25 did not lead to the rejection of the null hypothesis. Differences between adjusted Matched-Pairs means are not significant at the .05 level for any of the three comparisons.

Hypothesis 11 can also be partially tested by comparing Groups E and P in Tables 15-18. As before, the findings did not lead to the rejection of the null hypothesis. None of the differences between adjusted PET-1 means are large enough to be significant at the .05 level. The difference between adjusted YES-NO means is nearly large enough to be significant, but the higher reliability of the All-NO test, and the higher reliability and validity of the Matched-Pairs test, cause findings based upon them to be more acceptable. This is especially true for the Matched-Pairs test because it is identical in content to the YES-NO test.

There was, then, no significant difference in PET-1 means between first-grade students who were instructed in optimal and ordinary learning environments. The claim that it is necessary for teachers to have either special training or extensive experience with the Senesh materials in order to adequately teach the program was not substantiated.

It is concluded that first-grade children are capable of learning the content of Families at Work, and that ordinary first-grade teachers can adequately utilize the materials.

CHAPTER VI

SUMMARIES OF CONCLUSIONS, AND RECOMMENDATIONS

The central concern in this dissertation has been whether a sample of first-grade children could learn the content of Families at Work. A major secondary concern was whether a valid and reliable test could be developed to assess the economic learning of first-grade children. Two investigations were designed relative to those concerns: The WOBE and EPC studies. The former study was designed primarily to answer questions related to the central concern, and the latter study was directed primarily at the secondary concern. The studies were not entirely independent, however, in that each provided information useful in answering both questions. Findings and conclusions have been reported in Chapters IV and V. Summaries of those conclusions, plus recommendations, are presented below.

The Secondary Concern: Developing Test Forms for Use With Young Children

Three Primary Economics Tests: Grade One (PET-1) were developed. Two of these--the YES-NO and All-NO tests--are variations of the YES-NO or TRUE-FALSE format. The third is a multiple-choice picture test. In addition, the YES-NO test was written to be scored either in the ordinary manner or, by matching reversed pairs of items, with the resulting sets of scores treated as two separate tests. In effect, then, four PET-1

tests were compared: The YES-NO, Matched-Pairs, All-NO and Picture tests.[1]

Conclusions supported by pertinent findings

Reliability. Split-half reliability coefficients estimated for Matched-Pairs, All-NO, and Picture tests of equivalent length were similar--approximately .90--when given to knowledgeable students.[2] And only small differences were obtained among split-half reliability coefficients estimated for Matched-Pairs, All-NO, and Picture tests of equivalent length when given to control group students. The use of tests of equivalent length was impractical, however, since the Picture test and the Matched-Pairs test require more time to administer than does the All-NO test. When the tests were ranked according to the magnitude of reliability coefficients for the unequal test lengths actually used, the order was All-NO, Picture, Matched-Pairs, and YES-NO. Differences in reliability coefficients among the various tests were particularly noticable in the control groups, where the coefficients for the ordinary YES-NO test were lower than the .50 or .60 recommended for differentiating between group means.

It was concluded that, considering only reliability, any of the first three tests was adequate for the major purposes of this dissertation, such as comparing means. The reliability of the All-NO test is

---

[1]The forms of these tests are discussed in Chapters II and III. Reliability and validity are discussed in detail in Chapter IV. The items on the YES-NO Matched-Pairs test and the All-NO test are presented in Appendices A and C.

[2]See Chapter IV, Table 4.

also adequate for discriminating between individual students. Reliability alone, however, is not sufficient, and there is reason to suspect the validity of the All-NO test.

Validity. Two headings were used to discuss the validity of the PET-1 instruments: (1) Content validity, and (2) form validity.[3] Content validity was obtained by carefully comparing test items to the content of Families at Work. With the exception of one group of items on the Picture test, teachers who used Families at Work agreed that the content validity of the PET-1 tests is high.

Extensive analysis was devoted to what in this dissertation is called "form validity." Form validity refers to the effects that the form of a test has on students' responses, apart from the effects on reliability. Most of the arguments pertaining to form validity centered on the All-NO test.

The form validity of the All-NO test was first questioned on the basis of the a priori claim that first-grade children are not uniformly acquiescent. If two similarly knowledgeable children differ in acquiescence--the tendency to respond YES when not responding from knowledge--they will obtain dissimilar All-NO scores, because the less acquiescent student will guess NO more often that the other student will. If so, All-NO scores confound acquiescence-set with knowledge of the content of the test. It was further assumed that writing a YES-NO test with equal numbers of YES and NO items would balance acquiescence-set, and

---

[3]See the last section of Chapter II for the discussion of content validity, and the second major section of Chapter IV for the discussion of form validity.

that scoring reversed YES and NO items as one item--Matched Pairs scoring--would remove the effects of acquiescence. On a priori grounds, then, it was concluded that in form validity the three two-option PET-1 tests rank: (1) Matched-Pairs, (2) YES-NO, and (3) All-NO.

Four empirical comparisons supported the above a priori argument:

1. All of the PET-1 tests, except the All-NO test, noticably decreased in reliability from knowledgeable to ignorant groups of students.[4] This could be explained on the grounds that All-NO scores contain fewer responses not made from knowledge. However, that explanation is implausible because All-NO means are larger than Matched-Pairs means for items equivalent in number and nearly equivalent in content.[5] It is probably true that All-NO scores contain fewer chance responses than YES-NO, Matched-Pairs, or Picture scores. The most plausible explanation for the All-NO test containing fewer chance responses is that correct All-NO responses not made from knowledge are also not made by chance, but are rather due to response-set. All-NO scores, then, probably confound knowledge and response-set.

2. Another indication that the All-NO test confounds knowledge and response-set was obtained by correlating Matched-Pairs, YES-NO, and All-NO scores obtained by three separate scorings of a single administration of the YES-NO test. It was assumed that, since the YES-NO and Matched-Pairs tests were designed to minimize the effects of acquiescence on form validity, the correlation between YES-NO and Matched-Pairs scores

---

[4]See Chapter IV, Table 4.

[5]See Chapter IV, Table 6.

would be higher at the .01 level of significance than the correlation between either YES-NO and All-NO scores or Matched-Pairs and All-NO scores. This assumption was supported by the findings.[6]

3. Standard deviations of the YES-NO, Matched-Pairs, and All-NO scores taken from a single administration of the YES-NO test were also compared. YES-NO and Matched-Pairs standard deviations decreased in magnitude from knowledgeable to ignorant groups of students, but All-NO standard deviations did not.[7] Since it is to be expected that ignorant groups are less variable in knowledge than knowledgeable groups, it was concluded that the All-NO test measures something in addition to knowledge.

4. Since the three previous comparisons indicated that the form of the test affects the scores--and therefore the correlation coefficients, means, and standard deviations based on those scores--it was expected that t-ratios and F-ratios comparing knowledgeable and ignorant groups of students would also vary with the form of the test. This expectation was supported.[8] In general, it was concluded that F-ratios and t-ratios based on the All-NO test underestimate the significance of the difference between means, and that significance might be overestimated by F-ratios and t-ratios based on YES-NO scores.

In addition to the above comparisons, it was noted that only two of the four PET-1 tests--the Matched-Pairs test and the Picture test--

---

[6]See Chapter IV, Table 5.

[7]See Chapter IV, Table 6.

[8]See Chapter IV, Tables 7 to 10.

produced all of the expected discriminations among groups of students. This was taken to be a general indication of their superior validity.

Recommendations for using the four test forms

The Matched-Pairs test is apparently superior to the YES-NO and All-NO tests and should be chosen over them whenever circumstances permit. Nevertheless, there may be times when a teacher or researcher will find the YES-NO or All-NO test better suited to his purposes. Determining which test is best suited to a particular purpose requires an understanding of the practical limitations of each.

The Matched-Pairs test. The practicability of the Matched-Pairs test is limited to some extent by its scoring procedure. Since this procedure is not as complex as it appears, the limitation is not severe. If handled systematically, a Matched-Pairs test can be scored nearly as quickly as any other four-option instrument. The procedure is not difficult to use and can be adapted to tests of any length.[9] Although the Matched-Pairs test is not unreasonably difficult to score, the other test forms are much easier. For instance, the first half of the scoring procedure used on the Matched-Pairs test is identical to the entire procedure used to score the YES-NO test. The All-NO test is even easier to score than the YES-NO test. The All-NO score is the number of items the student marked NO. Therefore, if teachers or researchers need to measure learning gains but lack time to score a Matched-Pairs test, a YES-NO or All-NO test might be used, depending on the circumstances.

---

[9]See Appendix B for the Matched-Pairs scoring procedure.

Of course these tests should not be used unless the tester has good reason to believe that they are reliable and valid enough for his particular purpose.

The All-NO test. Because it probably produces dissimilar scores for students with similar knowledge, the All-NO test should never be used for any purpose that requires examining or comparing individual scores. This includes grading students. Neither should it be used for item analyses that require comparison of obtained and expected frequencies of correct response to individual items; the frequency of correct response expected when responding from ignorance is difficult if not impossible to determine. Furthermore, the All-NO test apparently obscures small differences between groups. But use of the All-NO test to compare group means might be justified if the teacher or researcher adopted a lower level of significance than he would with the Matched-Pairs test. Of course, the possibility that students will catch on to the All-NO test is always a threat to its validity.

The YES-NO test. Reliability of the YES-NO test is lower than recommended for any purpose unless the test contains at least 120 items.[10] Even then the anticipated reliability coefficient would justify no more than a comparison of means. If enough items were given to produce a reliability coefficient of .90 or better, the YES-NO test might be used for comparing scores for individual students. This would probably require combining scores from at least four 60-item tests, and is not advised.

---

[10] Based on esitmates made using the Spearman-Brown Prophecy Formula and the average of several reliability coefficients obtained on various YES-NO tests.

Time spent developing, administering, and scoring four YES-NO tests could be used to develop, administer, and score at least two Matched-Pairs tests. The latter course of action would likely produce scores with greater validity and at least equal reliability.

Ordinarily, if the purpose is to compare individual scores, they should be based on at least two 60-item Matched-Pairs tests. At times, however, a single 60-item test--30 pairs of items--has been highly reliable when given to a knowledgeable group.

The Picture test. The ordinary multiple-choice picture test is not recommended for classroom teachers, or for unfunded research projects. Its construction is extremely time consuming compared to the other test forms, and requires more artistic talent than the average person possesses. Picture tests also have limitations in terms of measuring some concepts not readily represented by that form.

The Primary Concern: Ability of First-Grade Children to Learn the Content of FAMILIES AT WORK

Conclusions supported by pertinent findings

Two general conclusions were supported by the findings: (1) The content of Families at Work is well suited to the ability of most first-grade children. And (2) teachers do not require special training or experience in order for their students to learn the content of Families at Work. The first conclusion is based on findings pertinent to several subconclusions which are presented in the following four paragraphs.

In general, first-grade children can learn the content of Families

at Work. In the Shaver-Larkins (1966) study, children in economically deprived areas of Salt Lake City who studied Families at Work scored significantly higher at the .01 level on a YES-NO PET-1 test than did similar children in a control group. In the EPC and WOBE studies, experimental groups of first-grade children from suburban and suburban-rural schools scored significantly higher at the .01 level on four different PET-1 tests than did the children in control groups.

As represented in the Matched-Pairs PET-1 test, every major concept in Families at Work was learned by at least some of the children in the experimental group. When items were analyzed separately, evidence that a statistically significant[11] number of students possessed knowledge related to the concepts being tested was absent for only 4 of 37 pairs of items. This finding was obtained by combining two different item analyses in each of three experimental groups. When only one item analysis was used, and the items were analyzed separately, ten items did not discriminate between control and experimental groups at the .10 level. But when analyzed as a group, these same items discriminated at the .001 level. When both item analyses were combined into a matrix, it was found that those items on which students' performance was rated Low frequently were similar to other items on which their performance was rated Moderate or High.

The content of Families at Work is not too easy for bright first-grade children. No student obtained a perfect or near-perfect score on the Matched-Pairs test. The highest score in any group was 33, which

---

[11] At the .05 level.

is 89 percent of the total possible. Other than in Group E, the highest score in an experimental group was 28, which is 76 percent of the total possible. That the students who obtained scores between 28 and 33 were above average in ability was demonstrated by their Tests of General Ability (TOGA) raw score mean. It was 54.43, which is equivalent to a grade expectancy of 3.7.

Students who were below average in ability learned at least part of the content of Families at Work. Those students in the experimental groups who were at least six months below grade-level obtained significantly[12] higher Matched-Pairs scores than did similar students in the control groups.

The first general conclusion, then, was that the content of Families at Work is well suited to the ability of most first-grade children. The second general conclusion was that teachers do not require special training or experience in order for their students to learn the content of Families at Work. Matched-Pairs means for students taught by teachers with special training and experience did not differ at the .05 level of significance from Matched-Pairs means for students taught by teachers who did not have special qualifications.

Recommendations for further research

Content validity of FAMILIES AT WORK. Although the content validity of the PET-1 tests was a major concern in Chapter IV of this dissertation, investigation of the content validity of Families at Work was never

---

[12]At the .01 level.

intended.[13] In one instance, however, it could not be avoided. As reported in the last section of Chapter IV, there is reason to doubt the content validity of Families at Work in regard to the term "specialist." If such invalidity can be discovered incidentally to the consideration of other problems, systematic investigation might uncover other sources of invalidity.

Appropriateness of teaching strategies. A second major question which was not investigated in the WOBE or EPC studies is whether the teaching strategies suggested in the teacher's manual for Families at Work are appropriate to the objectives of the Senesh materials. This question is particularly pertinent to those objectives relating to analytic thinking and problem solving. According to Senesh, analytic thinking is ". . . a tool for understanding and solving problems, a skill which is a prime objective of social science education (1963, p. 6)."

To consider whether Our Working World offers appropriate strategies for teaching problem solving, one must separate the conditions for inquiry from the procedures for making decisions. The educational environment may be conducive to inquiry, but the student may not know how to approach the task. Or, he may have an adequate procedural model, but the educational environment may be stifling. And, of course, both conditions may be either adequate or inadequate.

The Senesh program may not be compatible with some theories of the conditions for inquiry. For instance, Suchman (1965a, 1965b, 1965c, and

---

[13]By "content validity of Families at Work" is meant the extent to which concepts contained in it represent the disciplines from which they are intended to be drawn. For instance, is "consumer" defined in Families at Work similarly to the way in which economists generally use the term?

1966) stresses the importance of allowing students to arrive at their own conclusions, especially emphasizing the importance of the teacher forming questions that allow for divergent responses. He claims that forcing the child to give a predetermined response inhibits inquiry. On inspection, it appears that the Our Working World materials frequently violate this principle. An extreme example is found near the end of the recording which accompanies Lesson 2 in Neighbors at Work. In the dramatization, the townspeople have been debating whether to try and attract more tourists to their small town. To this point both pro and con arguments have been offered.

> NYE: All right, now. All right. I guess we could argue all day. But there are three things we can do. We can open the mill--we can fix up the stores and the courthouse square--and we can advertise to let tourists know what we're doing. It doesn't make sense to do just one of these things without the other, so if no one else has any other ideas, I think we should vote. And we'll all do whatever the vote decides.
>
> NARRATOR: Oh, boys and girls, isn't this exciting! Let's vote along with the others.
>
> NYE: All in favor, say yes.
>
> NARRATOR: Oh, let's say yes, children. We certainly want to help Littleton.
>
> CROWD: Yes.
>
> NYE: All those who are against these suggestions, say no.
>
> * * *
>
> NYE: Well, then, it's decided. Let's get to work.
>
> NARRATOR: Isn't this exciting! I know Littleton is going to be a better town than ever before. And just think . . . we helped!

It would take an unusual child to withstand the sort of pressure to conform which is found in the above example. Furthermore, does this sort of experience serve as an adequate model of rational decision making, or

of analytical thinking? Do such social issues have a single correct answer? If not, then it is a distortion of rationality to coerce children into uniform responses.

After comparing teaching strategies in Our Working World to various theories of the conditions for inquiry, it would be useful to investigate the adequacy of the procedures for problem solving which are outlined in the Senesh materials. The Senesh materials for the first three grades approach problem solving in several places. Lesson 10 in Families at Work--Grade One--is titled "How Choices Are Made," and focuses on the concept of limited resources versus unlimited wants. Lesson 13 in Neighbors at Work--Grade Two--is titled, "How Neighborhoods Solve Problems," and outlines six steps in problem solving. The same approach, with the same title, is found on page 143 of the Developmental Edition of Cities at Work--Grade Three. The six steps are:

1. Evidence of the problem
2. Definition of the problem
3. Aspects of the problem
4. Size of the problem
5. Causes of the problem
6. Solution of the problem

(Senesh, 1965, p. 209)

These same steps converted into language suitable for children are listed as:

1. Observe the problem
2. Ask the big question
3. How does the problem affect our lives
4. Measure the problem
5. Find the causes
6. Solve the problem
   a. What can you do?
   b. What can neighbors do together?
   c. What can the city do?

(Senesh, 1966a, p. 143)

One of the difficulties in critiquing this approach to problem solving is lack of information concerning changes Senesh may intend to introduce for older children. Recognizing that limitation, the following suggestions are offered. First, since a proclaimed primary objective of Our Working World is to help children develop skills of analytic thinking, investigators may want to determine what proportion of each course of study is directly related to that task. On inspection, it is doubtful that, even if Senesh's analytic model is adequate, enough time is spent training children to use it. Second, investigators may want to determine whether telling children to do such things as observe the problem and measure the problem, is enough, or whether children also need more specific training in how to observe or measure. If children need specific training in carrying out the various steps in problem solving, is that training provided in Our Working World? Third, and perhaps most important, is problem solving the straight forward empirical process that Our Working World appears to make it out to be? Is problem solving limited to description and prediction? Or is problem solving also concerned with what should be; with whether a given condition is right or wrong, morally defensible or reprehensible? If "solving"--certainly a questionable term in itself--societal problems requires settling value disputes, are there procedures suitable to that task included in Our Working World? Lesson 10--"How Choices Are Made"--of Families at Work is at least peripherally related to that task, but it is difficult to see the relationship between that lesson and the six steps in analytic thinking outlined in the materials for Grades Two and Three.[14]

---

[14]These steps are listed on the preceding page of this dissertation.

It may be that the power of Senesh's model of analytical thinking and decision making is underestimated in this discussion. It may also be that more sophisticated models for analyzing societal problems cannot be adapted to a primary-grades program. Research in this area is justified, however, and would likely be welcomed by no one more than by the people involved with the Senesh projects.[15]

Affective learning. Another area worthy of investigation concerns affective as opposed to cognitive learning associated with exposure to Our Working World. In Families at Work Senesh says, "Over and beyond the introduction of certain basic understandings from the various sciences, the author tries to develop attitudes and values necessary to a free society (Senesh, 1963, p. 4)." Other investigators may want to identify attitudes which Senesh is trying to teach and measure the extent to which children acquire them as a result of such instruction. They may also want to focus their research on questions such as these. Does exposure to Families at Work:

1. Produce positive attitudes towards certain occupations--businessmen or bankers for instance?

2. Produce different affective learning in children from different socio-economic backgrounds?

3. Alter attitudes towards specific problems or topics such as taxes, or community action projects, or rural-urban change?

4. Change childrens' feelings towards ingroups and outgroups such as minority groups in our culture or people in foreign cultures?

---

[15]Discussions with Joseph Rueff indicated that there is more concern with that question than with those which formed the basis for this dissertation.

5. Alter student attitudes towards specific school subjects or towards school in general?

It might also be interesting to determine whether teachers' attitudes change in some of the above ways as a result of using Families at Work.

Cognitive learning. The first chapter of this dissertation distinguished between teaching economics as a unified, structured discipline of related concepts, and teaching economics as a list of commonly used economic terms, or a series of practical experiences. It was stressed that Senesh believes in teaching economics as a unified discipline. He modifies that position somewhat, however, in the following statement. "It is not expected that by the end of the first year the children will be able to formulate clearly the fundamental theoretical relationships of the various areas of the social sciences (1963, pp. 5-6)." The WOBE and EPC studies mirrored that expectation. They were used to test knowledge of individual terms and their related concepts, but did not attempt to determine whether children could relate these concepts into a larger system. Other researchers, then, may want to determine at what point children can be expected to do this. For instance, further investigation may be warranted to determine whether instruction in Families at Work enhances childrens' ability to systematize--to grasp larger and more abstract relationships--earlier than they ordinarily would.

Summary

The objectives of this dissertation were: (1) To develop a valid and reliable achievement test based on the content of Families at Work.

And (2) to use that instrument to determine whether a sample of first-grade children could learn the content of Families at Work. In regard to the first objective, it was concluded that at least one of the four PET-1 tests--the Matched-Pairs test--validly measures the major concepts in Families at Work.[16] It was also concluded that the reliability of the Matched-Pairs test is adequate for the types of discriminations made in this dissertation--such as the comparison of means between control and experimental groups of students. In regard to the second objective, it was concluded that the cognitive content of Families at Work is well suited to the abilities of first-grade children. No attempt was made to determine the ability of first-grade children to understand the general structure of economics, as this was not an objective of the first-grade course of study. Neither was there an attempt to assess learning in the affective domain, nor to investigate the content validity of Our Working World.

---

[16]It was also concluded that the Picture test validly measures several of the major concepts in Families at Work. The Picture test, however, is not comprehensive.

POSTSCRIPT

During discussions of the Senesh materials, other educators have usually indicated that they are not as concerned with the findings, conclusions, or recommendations of this paper as they are with my personal reaction to Our Working World. Generally, their attempts to pin me down have led them to ask some such question as, "If you were a first-grade teacher, would you use these materials?" or "Would you like to see your own first-grader study Families at Work?" After explaining some reservations--most of which have been expressed in the recommendations for further research--my answer is "Yes." In my opinion, Families at Work is superior to the traditional first-grade social studies courses of study. In the past, social studies curriculum developers have tended to grossly underestimate the intellectual capacities of young children. The major strength of the Senesh materials is that the young student is given something that adds to rather than rehashes his present fund of knowledge. Senesh and his associates have produced a pioneer work in primary-grades social studies. It is hoped that Our Working World will challenge others to turn their attention to the adequate education of primary-grade children.

LITERATURE CITED

Anonymous. 1964. How one district teaches economics in grade school. School Management 8(6): 52-56, 112-114.

Bass, Bernard M. 1955. Authoritarianism or acquiescence? Journal of Abnormal and Social Psychology 51(3): 616-623.

Barnes, Marcillene. 1953. Community economics for the third grade: A new venture in textbook writing. Social Education 17(7): 339-340.

Barnes, Eugene. 1962. Address to the Fourth Annual Military Testing Association Conference. October 25. (Mimeographed)

Bircher, Jack Lawrence. 1964. An analysis of selected educational, social, economic, and political opinions held by business leaders. EdD dissertation. Indiana University. 94 p. (Original not seen; abstracted in Dissertation Abstracts 25: 5705. 1965.)

Bond, Floyd A., and George L. Roehr. 1952. The rediscovery of economics. California Journal of Secondary Educaton 27(5): 295-300.

Borg, Walter R. 1963. Educational research: An introduction. David McKay Company, Inc., New York. 418 p.

Brown, Richard, and Victoria Daily. 1961. Is economics essential? The Balance Sheet 43(4): 162-163.

Brunson, Evelyn. 1966. Economics: From the cradle to the grave. The Balance Sheet 47(5): 204-205.

Butler, Delbert Franklin, Jr. 1965. A test for measuring selected life sciences concepts of elementary school children. EdD dissertation. George Peabody College for Teachers. 130 p. (Original not seen; abstracted in Dissertation Abstracts 26: 6505. 1966.)

Chapman, Loren J., and Donald T. Campbell. 1957. Response set in the F scale. Journal of Abnormal and Social Psychology 54(1): 129-134.

Christie, Richard, Joan Havel, and Bernard Seidenberg. 1958. Is the F scale irreversible? Journal of Abnormal and Social Psychology 56(2): 143-159.

Clayton, Dean. 1966. Teaching economics in typewriting classes: Study I. Business Education Forum 21(2): 14.

Coleman, John R. 1963. The American economy: A plea for understanding. The Bulletin of the National Association of Secondary School Principals 47(282): 182-189.

Coon, Ann. 1966. Introducing the economic world to primary-grade pupils. Social Education 30(4): 253-256.

Couch, Arthur, and Kenneth Keniston. 1960. Yeasayers and naysayers: Agreeing response set as a personality variable. Journal of Abnormal and Social Psychology 60(2): 151-174.

Cowling, Richard R., Jr. 1966. Teaching economics in typewriting classes: Study II. Business Education Forum 21(2): 15.

Cronbach, Lee J. 1942. Studies of acquiescence as a factor in the true-false test. Journal of Educational Psychology 33(6): 401-415.

Cronbach, Lee J. 1946. Response sets and test validity. Educational and Psychology Measurement 6(4): 475-494.

Cronbach, Lee J. 1950. Further evidence on response sets and test design. Educational and Psychological Measurement 10(1): 3-31.

Cummings, Howard. 1950. Economic education in the secondary schools. The Journal of Educational Sociology 23(7): 397-401.

Darrin, Garney L. 1958. Economics in the elementary school curriculum: A study of the District of Columbia laboratory schools. EdD dissertation. University of Maryland. 401 p. University Microfilms, Ann Arbor, Michigan.

Darrin, Garney L. 1960a. You asked about economics. Grade Teacher 77(7): 36-37, 117.

Darrin, Garney L. 1960b. Economics in the upper primary grades. Grade Teacher 78(3): 24-25, 122-124.

Darrin, Garney L. 1960c. Economics in the primary grades. Grade Teacher 78(1): 54-55, 110.

Darrin, Garney L. 1961. The food store. Grade Teacher 79(1): 56-57, 127-129.

Deitz, James Emery. 1963. Economic understanding of senior students in selected California high schools. EdD dissertation. University of California Los Angeles. 327 p. (Original not seen; abstracted in Dissertation Abstracts 24: 3562. 1964.)

Delva, Josephine G. 1955. Integrating economics in the elementary social studies. The Social Studies 46(8): 294-298.

Eames, Hamilton. 1949. Who gets how much for doing what in America. The Journal of Business Education 24(9): 15-16.

Eisen, Agnes. 1958. Economic education through a school store. The Elementary School Journal 58(5): 287-289.

Flanagan, John C. 1959. Tests of general ability: Form A grades K-2. Science Research Associates, Inc., Chicago, Illinois.

Foster, Dorothy Elizabeth. 1965. A study of first grade children's comprehension of typical first grade social studies content prior to systematic instruction. EdD dissertation. Colorado State College. 123 p. (Original not seen; abstracted in Dissertation Abstracts 26: 1922. 1965)

Frisina, Frances. 1962. Second-graders learn economics in a real school store. The Instructor 72(1): 144-145.

Fromberg, Doris Pronin. 1965. The reactions of kindergarten children to intellectual challenge. EdD dissertation. Columbia University. 202 p. (Original not seen; abstracted in Dissertation Abstracts 26: 904. 1965.)

Garrett, Henry E. 1958. Statistics in psychology and education. David McKay Company, Inc., New York. 477 p.

Garwood, John D. 1962. Changing economics. The Social Studies 52(7): 163-166.

Garwood, John D. 1964. The need for economic education. School and Society 29(2247): 229-291.

Gavian, Ruth Wood, and Louis C. Nanassy. 1955. Economic competence as a goal of elementary-school education. The Elementary School Journal 55(5): 270-272.

Gavian, Ruth Wood. 1958. Developing economic understanding. The Instructor 67(7): 6, 63.

Gilbreth, Harold B. 1945. Trends in teaching economics. The Journal of Business Education 20(8): 15-16.

Goldstein, Phineas. 1966. Concepts of landforms and waterforms of children beginning first grade. EdD dissertation. University of Southern California. 279 p. (Original not seen; abstracted in Dissertation Abstracts 27:1199. 1966.)

Goodykoontz, Bess. 1953. Selected studies relating to community schools, p. 64-82. In Nelson B. Henry (Ed.) The fifty-second yearbook of the National Society for the Study of Education: Part II, the community school. University of Chicago Press, Chicago, Illinois.

Hadley, Everett Edwin. 1964. Development of a pre-kindergarten test for predicting school achievement in the primary grades. EdD dissertation. University of Connecticut. 132 p. (Original not seen; abstracted in Dissertation Abstracts 25: 4541. 1964.)

Helfrich, John Edward. 1963. A descriptive study of certain science learnings known by entering kindergarten children. EdD dissertation. Wayne State University. 158 p. (Original not seen; abstracted in Dissertation Abstracts 25: 232. 1964.)

Hensen, Rosa May. 1964. A measurement of social studies achievement in the primary grades. EdD dissertation. North Texas State University. 136 p. University Microfilms, Ann Arbor, Michigan.

Howell, Jack Olen. 1965. Relationships of school superintendents' proficiency in economics to job performance. PhD dissertation. The Ohio State University. 198 p. (Original not seen; abstracted in Dissertation Abstracts 26: 3719. 1966.)

Jacobsen, James Huffman. 1964. A study of the contribution of industrial arts instruction to consumer knowledge. EdD dissertation. Colorado State College. 108 p. (Original not seen; abstracted in Dissertation Abstracts 25: 2813. 1964.)

Johansen, Nancy Keller. 1965. A comparison among differentiated socio-economic groups of first-grade children in regard to a background of information. EdD dissertation. Univeristy of Missouri. 87 p. (Original not seen; abstracted in Dissertation Abstracts 26: 5293. 1966.)

Knoble, John W. 1939. The teaching of economics in the sixth grade. Education 59(6): 363-366.

Lagemann, John Kord. 1964. Its never too early to learn the economic facts of life. The PTA Magazine 59(3): 4-7.

Larkins, A. Guy, and James P. Shaver. 1967. Matched-pairs scoring technique used on a first-grade YES-NO type economics achievement test. Utah Academy of Sciences, Arts, and Letters: Proceedings 44(1): 229-242.

Leavitt, Harold J., Herbert Hax, and James H. Roche. 1955. "Authoritarianism" and agreement with things authoritative. The Journal of Psychology 40(2): 215-221.

Lemmon, Mary Louise. 1962. A comparison of consumer economic knowledge of two Illinois populations of high school home economic teachers and their students. EdD dissertation. University of Illinois. 67 p. (Original not seen; abstracted in Dissertation Abstracts 23: 546. 1962.)

Logan, S. R. 1946. Economics in action at Skokie. The Clearning House 21(4): 201-203.

Lovenstein, Meno. 1961. Economics for grownups and kids. The Instructor. 71(4): 6, 71.

Lovenstein, Meno, Edward J. Furst, Robert Jewett, and Elizabeth Steiner Maccia. 1966. Development of economics curricular materials for the secondary schools: Cooperative Research Project Number HS-082. The Ohio State University Research Foundation, Columbus, Ohio. 935 p.

Lowry, Betty Lucille. 1963. A survey of the knowledge of social studies concepts possessed by second grade children previous to the time these concepts are taught in the social studies lessons. PhD dissertation. State University of Iowa. 252 p. (Original not seen; abstracted in Dissertation Abstracts 24: 2324. 1963.)

Madsen, Gibb R. 1961. Economic concepts and understandings of senior high school students. PhD dissertation. University of Utah. 196 p. University Microfilms, Ann Arbor, Michigan.

McCombs, N. D., and George W. Hohl. 1953. Business can help teach economics. The School Executive 72(6): 80-82.

McKee, C. W., and H. G. Moulton. 1951. A survey of economic education. The Brookings Institution, Washington, D. C. 63 p.

McPherson, J. Roland. 1948. Why teach introductory economics? The Social Studies 39(1): 28-32.

McPhie, Walter E. 1964. Dissertations in social studies education: A comprehensive guide. National Council for the Social Studies, Washington, D. C. 100 p.

Melby, Ernest O. 1950. Economic education is a must. The Journal of Educational Sociology 23(7): 378-388.

Merrifield, Charles W. 1959. Economic competence: New frontier in civic education. Social Education 23(2): 71-74.

Mogar, Robert E. 1960. Three versions of the F scale and performance on the semantic differential. Journal of Abnormal and Social Psychology 60(2): 262-265.

Northwest Council for Economic Education, and Office of the State Superintendent of Public Instruction. 1966. Economic education for Washington schools: Kindergarten through grade six. Olympia, Washington. Approximately 200 p. (pagination not continuous).

Nourse, Edwin G. 1966. Current aspects of our persistent economic problems. Social Education 30(4): 232-235.

Nunnally, Jum C. 1964. Educational measurement and evaluation. McGraw-Hill Book Company, New York. 440 p.

Olson, Clara M., and Hazen E. Nutter. 1945. Housing: The Sloan experiment in Florida. The Clearing House 19(7): 420-425.

Ong, Jin Shuck Louis. 1963. Measurement with opposite forms of an inventory. PhD dissertation. University of California, Berkeley. 80 p. University Microfilms, Ann Arbor, Michigan.

Parker, James Robert. 1963. An analysis of childrens' concepts of selected occupations. EdD dissertation. Northwestern University. 206 p. (Original not seen; abstracted in Dissertation Abstracts 24: 3631-3632. 1963)

Peabody, Dean. 1961. Attitude content and agreement set in scales of authoritarianism, dogmatism, anti-semitism, and economic conservatism. Journal of Abnormal and Social Psychology 63(1): 1-11.

Perry, Enos C. 1960. Economic illiteracy, not Russia, is our greatest threat. The Balance Sheet 42(1): 19.

Pierrepont, R. S. 1948. Economic illiteracy leads to national disunity. The Journal of Business Education 23(5): 11-12.

Prehn, Edward C. 1965. Economics in the K-12 social studies project. High Points 47(7): 36-38, 40.

Reed, Clyde L. 1958. Detroit pioneers in economic education. Social Education 22(2): 63-65.

Reinbold, John Clifford. 1965. An analysis of changes in attitudes of participants in selected clergy economic education programs. EdD dissertation. Indiana University. 130 p. (Original not seen; abstracted in Dissertation Abstracts 26: 5200. 1966.)

Robbins, L. C. 1955. The teaching of economics in schools and universities. The Economic Journal 65(260): 579-593.

Robinson, Helen F. 1963. Learning economic concepts in the kindergarten. EdD dissertation. Columbia University. 321 p. University Microfilms, Ann Arbor, Michigan.

Rohrbaugh, Barbara, and Ruth E. Haines. 1960. Economic education in the primary grades, p. 33-39. In Commission on education for economic comptence. Educating for economic competence. Association for Supervision and Curriculum Development. The National Education Association, Washington, D. C.

Rokeach, Milton. 1963. The double agreement phenomenon: Three hypotheses. Psychological Review 70(4): 304-309.

Rorer, Leonard George. 1963. The function of item content in MMPI responses. EdD dissertation. University of Minnesota. 503 p. University Microfilms, Ann Arbor, Michigan.

Rush, Orville Findley, Jr. 1964. The development of scientific concepts of living and non-living among pre-school children. EdD dissertation. The University of North Carolina at Chapel Hill. 246 p. (Original not seen; abstracted in Dissertation Abstracts 26: 1486. 1965.)

Saunders, Phillip. 1966. Preparing future teachers for economic competence: Content appraisal. Social Education 30(4): 247-250.

Schultz, Frank G. 1953. Horse sense and buggy economics. School and Society 77(2008): 372-375.

Seay, Maurice F. 1945. Nutrition: The Sloan experiment in Kentucky. The Clearing House 19(7): 426-428.

Senesh, Lawrence. 1958. We cannot afford economic illiteracy. The School Executive 78(2): 51-53.

Senesh, Lawrence. 1963. Our working world: Families at work, resource unit. Science Research Associates, Inc., Chicago, Illinois. 198 p.

Senesh, Lawrence. 1965. Our working world: Neighbors at work, resource unit. Science Research Associates, Inc., Chicago, Illinois. 296 p.

Senesh, Lawrence. 1966a. Our working world: Cities at work, developmental edition (student text). Science Research Associates, Inc., Chicago, Illinois. 257 p.

Senesh, Lawrence. 1966b. What economics is. The Instructor 75(10): 34, 81.

Sewell, Edward G. 1963. Effect of classwork in economics on attitudes and understanding of a select group of secondary school pupils. The Journal of Educational Research 57(3): 131-136.

Shaver, James P., and A. Guy Larkins. 1966. Evaluation report to Salt Lake City School District: SRA economics materials in grades one and two. Bureau of Educational Research, Utah State University. (Mimeographed)

Siegal, Sidney. 1956. Nonparametric statistics for the behavioral sciences. McGraw-Hill Book Company, New York. 312 p.

Sloan, Alfred P. 1943. Teaching experiments in better living. The Nations Schools 31(6): 12-14.

Spodek, Bernard. 1962. Developing social science concepts in the kindergarten. EdD dissertation. Columbia University. 248 p. (Original not seen; abstracted in Dissertation Abstracts 23: 1563. 1962.)

Stephens, Lois Evans. 1964. What concepts of telling time can be developed by kindergarten children. EdD dissertation. University of California Los Angeles. 122 p. (Original not seen; abstracted in Dissertation Abstracts 25: 1793. 1964.)

Stoner, James K. 1962. The relative value of selected economic concepts. The Balance Sheet 44(3): 100-104.

Suchman, J. Richard. 1965a. The motivation to inquire. The Instructor 75(2): 26, 122, 125.

Suchman, J. Richard. 1965b. The conditions for inquirey. The Instructor 75(3): 30, 137-138.

Suchman, J. Richard. 1965c. The role of the teacher. The Instructor 75(4): 26, 64.

Suchman, J. Richard. 1966. Inquirey in the curriculum. The Instructor 75(5): 24, 64.

Thompson, Normal Samuel. 1965. The measurement of consumer credit knowledge. EdD dissertation. Colorado State College. 170 p. (Original not seen; abstracted in Dissertation Abstracts 26: 5141. 1966.)

Toone, Herbert A. 1955. What brand of economics shall we teach? The Balance Sheet 36(6): 247-248, 252.

Uhr, Carl G. 1963. Some generalizations from economics. Journal of Secondary Education 38(7): 54-57.

U. S. Bureau of the Census. 1961. U. S. census of population: 1960. Volume I, Characteristics of the population. Part A, Number of inhabitants. U. S. Government Printing Office, Washington, D. C.

Wilde, E. Irving. 1954. Defeciencies in economic understanding. The Journal of Business Education 29(4): 161-162, 168.

Wing, Richard L. 1964. Computer-controlled economics games for the elementary school. Audiovisual Instruction 9(9): 681-682.

Wolfson, Martin. 1950. Let's make economics meaningful. The Journal of Education 133(5): 144-145.

Wood, Dorothy Adkins. 1961. Test construction: Development and interpretation of achievement tests. Charles E. Merrill Books, Inc., Columbus, Ohio. 134 p.

Zurcher, Arnold J. 1965. Progress toward economic literacy. Bulletin of the National Association of Secondary School Principals 49(299): 106-110.

APPENDIXES

Appendix A

YES-NO and Matched-Pairs Test

Instructions to testers in WOBE study

1. Select students from classes. Do not test students who are not on your list. If a child is absent we will drop him from the sample.

2. Make sure each child has a pencil and a crayon.

3. Write the name of the school on the back of each test booklet.

4. Have each child print his name on the back of his test booklet.

5. Practice three items on the practice sheet. You may practice more if children do not seem to catch on.

6. Have children lay their pencils down when they are not being used.

7. Periodically throughout the test encourage the children to guess. Many students will feel uncomfortable guessing and will need frequent reassurance.

8. Take a short break after items 30 and 60. Have the children stand and stretch.

9. Read each item twice. After each reading say, "Circle either YES or NO."

10. Have children point to the number of the item you are on so they will not lose their place. This may not be necessary after the first page. The children should be asked to point to the first number on each page as a check against turning too many pages or the possibility that the pages were placed in the booklet in the wrong order.

11. Try to control "peeking." Spread the children out as much as possible. Remind them not to look on other's papers.

12. Pace yourself so that actual test time is 45 minutes or less. Try to keep the children working and given them frequent encouragement.

PET-1 YES-NO and Matched-Pairs items for Lessons 1-24 of FAMILIES AT WORK

Lesson #	Item #	
1	1	Your brothers or sisters are part of your close family. (yes)
2	2	Almost every family in the world has a T.V. (no)
	3	Almost every family in the world has a telephone. (no)
	4	We must have food. (yes)
	5	We must have T.V. (no)
3	6	Eskimos, Bushmen, and Indians live in different kinds of houses. (yes)
2	7	Everyone in the family is a consumer. (yes)
4	8	A farmer who raises potatoes is a producer of goods. (yes)
	9	Children who jump rope are producers. (no)
	10	When people shovel snow off the sidewalk they are producing a service. (yes)
	11	Everyone except babies and sick people are producers. (yes)
5	12	When Mother washes the dishes and Sister dries them they are dividing the labor. (yes)
	13	It is faster and cheaper to divide the labor. (yes)
	14	When Brother sweeps the floor and Sister makes the bed they are dividing the labor. (yes)
	15	Nations who trade with each other divide the labor. (yes)
6	16	Tools and machines make it harder to do work. (no)
7	17	Father would usually save money if he stayed home from work and washed the car. (no)
8	18	We have more free time because we divide the labor. (yes)

Lesson #	Item #	
11	19	Most pioneers lived in cities. (no)
10	20	If we worked harder we could have everything we want. (no)
11	21	Pioneers are people who life in a different country. (no)
9	22	Rich people want more things than they can have. (yes)
12	23	Customs and rules help us to know what other people will do. (yes)
13	24	A specialist depends on others to produce the things he needs. (yes)
	25	A specialist knows how to do one job very well. (yes)
	26	Specialists usually do their work away from home. (yes)
14	27	Transportation makes it harder for specialists to trade their goods and services. (no)
15	28	"Income" is money people get for doing work. (yes)
16	29	When many people try to get the same job the wages will usually be lower. (yes)
17	30	If two stores sell things that are just alike, the store with the lowest prices will usually have more customers. (yes)
18	31	Our schools are not usually paid for by taxes. (no)
19	32	Some families save part of their income. (yes)
20	33	Banks loan money to anyone who need it. (no)
21-23	34	Profit is money the businessman gets for worrying. (yes)
	35	Before he can go into business a man needs a wife, a car, materials and workers. (no)
24	36	When people buy more goods, more workers have jobs. (yes)

Lesson #	Item #	
24	37	When people stop buying goods, more businesses are started. (no)
1	38	Your mother and father are part of your distant family. (no)
2	39	In some places only a few families have T.V. (yes)
	40	In some places only a few families have telephones. (yes)
	41	We can get along without food. (no)
	42	We can get along without T.V. (yes)
3	43	Eskimos, Bushmen, and Indians live in the same kind of houses. (no)
2	44	Mother and Father are the only consumers in the family. (no)
4	45	A farmer who raises weeds is a producer of goods. (no)
	46	Children who wash dishes are producers. (yes)
	47	When people shovel snow onto the sidewalk they are producing a service. (no)
	48	Everyone is a producer. (no)
5	49	When Mother and Sister watch T.V. they are dividing the labor. (no)
	50	It is faster and cheaper for one man to produce all of his own goods. (no)
	51	When two babies are playing with dolls they are dividing the labor. (no)
	52	Nations who trade with each other do not divide the labor. (no)
6	53	Tools and machines make it easier to do work. (yes)
7	54	Father would usually lose money if he stayed home from work and cut the grass. (yes)

Lesson #	Item #	
8	55	People who divide the labor have very little free time. (no)
11	56	Most pioneers lived on farms. (yes)
10	57	People who work very hard still want more things than they have. (yes)
11	58	Pioneers lived a long time ago. (yes)
9	59	Rich people can have everything they want. (no)
12	60	Customers and rules make it hard to know what other people will do. (no)
13	61	A specialist produces for himself everything he needs. (no)
	62	A specialist knows how to do many different kinds of jobs very well. (no)
	63	Specialists usually stay home to do their work. (no)
14	64	Transportation makes it easier for specialists to trade their goods and services. (yes)
15	65	"Income" means "come in the house." (no)
16	66	When many people try to get the same job the wages will usually be higher. (no)
17	67	If two stores sell things that are just alike, the store with the highest prices will usually have more customers. (no)
18	68	Our schools are usually paid for by taxes. (yes)
19	69	Every family spends all of their income. (no)
20	70	Banks loan money only to people who will pay it back. (yes)
21-23	71	Profit is money the workers get for worrying. (no)
	72	Before he can go into business a man needs materials, workers, tools, and a workplace. (yes)
24	73	When people buy fewer goods, more workers have jobs. (no)

Lesson #	Item #	
24	74	When people buy many goods, more businesses are started. (yes)
	75	A businessman who sells a vacuum for $40 makes $40 profit. (no)

Appendix B

Scoring Procedure for Matched-Pairs Test

Suppose that a 60-item YES-NO test is produced and items are written in pairs with reversals. The test should be divided so that one half is the mirror image of the other. This is done by forming two groups of items, with each item in a separate group than its reversal. The 30 items in the first group are randomly assigned numbers 1-30 on the test. The reversals of each of these items receive the corresponding numbers 31-60. The reversal for Item 1 is Item 31; the reversal for Item 2 is Item 32.

One score sheet for each student is mimeographed for the teacher to use in correcting the response sheets. These score sheets contain two double columns. Each column has, side by side, the number of an item and the number of its reversal. To correct the response sheets, a line is drawn through the number of each item that is incorrectly answered. A plus is placed beside each pair of items that does not have a line through either number. The pluses are counted and the total is the student's score. In the following example the student missed 22 single items, and 19 pairs of items; his score was 11 of 30 pairs of items.

Figure 1. Sample score sheet for Matched-Pairs test

1-31		16-46	+
2-32		17-47	
3-33	+	18-48	+
4-34		19-49	
5-35		20-50	+
6-36		21-51	
7-37		22-52	
8-38		23-53	
9-39		24-54	
10-40	+	25-55	
11-41	+	26-56	
12-42	+	27-57	
13-43	+	28-58	+
14-44	+	29-59	+
15-45		30-60	

Score: 11/30

Sample page from students' response booklet for YES-NO type tests

1	YES	NO
2	YES	NO
3	YES	NO
4	YES	NO
5	YES	NO
6	YES	NO
7	YES	NO
8	YES	NO
9	YES	NO
10	YES	NO
11	YES	NO
12	YES	NO
13	YES	NO
14	YES	NO
15	YES	NO

Appendix C

PET-1 All-NO Test

(Instructions to testers, and the students' response booklet are the same as for the YES-NO Matched-Pairs test.)

Lesson #	Item #	
1	1	Your mother and father are part of your distant family.
2	2	We can get along without food, clothes, or houses.
3	3	Eskimos, Bushmen, and Indians live in the same kind of houses.
4	4	Everyone is a producer.
5	5	When each person in the family washes his own clothes they have divided the labor.
6	6	Tractors, cars, and trucks are simple tools.
7	7	Most people would save money if they grew their own food, made their own clothes, and built their own houses.
8	8	People who use many tools and machines have very little free time.
9	9	If we worked harder we could have everything we want.
10	10	People usually want just a few things.
11	11	Pioneers are people who live in a different country.
12	12	Every custom is a rule.
13	13	A specialist knows how to do many different kinds of jobs.
14	14	Transportation makes it harder for specialists to trade their goods and services.

Lesson #	Item #	
15	15	"Income" means the same as "price."
16	16	"Wages" means the same as "interest."
17	17	Most families need a car more than they need a house.
18	18	"Taxes" means "money we pay to stores."
19	19	Families always spend all of their money.
20	20	"Loan" means "putting money in the bank."
21-23	21	Profit is money a worker gets for worrying.
24	22	When people do not buy goods, more workers have jobs.
2-4	23	A man who is building a fence is consuming the fence.
4	24	A factory that builds cars is producing services.
1	25	Most families live in the same house all of their lives.
2	26	We must have television.
3	27	People everywhere use money.
4	28	When a barber cuts peoples' hair he is producing goods.
5	29	When two countries both raise bananas they have divided the labor.
6	30	Tools and machines make it harder to do work.
7	31	Father would usually save money if he stayed home from work and washed the car.
8	32	People have very little free time when they divide the labor.
9	33	A man who is very rich can have everything he wants.
10	34	If a girl got new clothes and a new doll she should not want anything else.
11	35	We have to work harder than people use to.
12	36	Customs are the same in all countries.
13	37	When people divide the labor there are fewer specialists.

Lesson #	Item #	
14	38	Washing the dishes is one kind of transportation.
15	39	"Income" means "money we pay for goods and services."
16	40	When many people try to get the same job the wages will usually be higher.
17	41	Most families need T.V. more than they need clothes.
18	42	Indiana/Utah is bigger than the United States.
19	43	Some people have all the things they want, so it is easy for them to save money.
20	44	A man who gives money away is borrowing.
21	45	A man needs to be old before he can go into business.
24	46	When people do not buy goods, more businesses are started.
4	47	A man who is eating pie is producing the pie.
1	48	Most families are the same size.
2	49	Food must be consumed before it can be produced.
3	50	Eskimos are farmers.
4	51	When a carpenter builds houses he is producing services.
5	52	When two farmers both raise pigs they have divided the labor.
6	53	People used to have better tools and machines than we have.
7	54	Father would usually save money if he stayed home from work and cut the grass.
8	55	Eskimos have more free time than we do.
9	56	Only people who work hard are consumers.
10	57	If a boy got ten dollars to spend any way he wanted he would not want anything else.
11	58	Most pioneers lived in cities.

Lesson #	Item #	
12	59	Customs and rules make it harder to know what others will do.
13	60	Specialists usually stay home to do their work.
14	61	A person who is not a specialist can make things faster than a specialist can.
16	62	When there are many jobs and not very many people looking for jobs the wages will usually be lower.
17	63	If two stores sell things that are just alike, the store with the highest prices will have more customers.
18	64	Elkhart/Ogden/Brigham City is bigger than Indiana/ Utah.
20	65	Banks loan money to anyone who needs it.
21	66	A man needs to be married before he can go into business.
4	67	A person who is teaching school is producing goods.
5	68	When each person in the family cooks his own food they have divided the labor.
16	69	A fireman is usually paid more than a doctor.
18	70	Our food is usually paid for by taxes.
9	71	Some people can have everything they want.
6	72	We have fewer tools and machines than people used to have.
16	73	A milkman is usually paid more than a doctor.
4	74	A car salesman is a producer of goods.

Appendix D

The PET-1 Picture Test

Test Instructions

(Have the name of the school and the name of the teacher on the chalkboard.)

WITH YOUR PENCIL I WANT YOU TO PRINT YOUR NAME WHERE IT SAYS "NAME" ON THE FRONT PAGE. (Pause)

PRINT THE NAME OF YOUR TEACHER ON THE NEXT LINE. (Pause)

ON THE BOTTOM LINE PRINT THE NAME OF YOUR SCHOOL. (Pause)

NOW PUT YOUR PENCIL DOWN. WE'RE NOT GOING TO USE OUR PENCILS FOR AWHILE. DON'T PICK YOUR PENCIL UP UNTIL I TELL YOU TO. (Check to make sure that each child has filled in the blanks correctly.)

OPEN YOUR BOOKLET AND FOLD IT BACK LIKE THIS SO THAT THE FIRST PAGE IS SHOWING. (Demonstrate.) (If necessary, remind children to leave their pencils on the desk.)

IN THE FIRST ROW THERE ARE PICTURES OF A BOY WITH A DRINK, A BOY IN A SWING, A BOY IN BED, A BOY MOWING THE GRASS, AND A BOY EATING AN APPLE. LEAVE YOUR PENCILS ON THE DESK. WITH YOUR FINGER POINT TO THE PICTURE OF THE BOY YOU THINK MIGHT BE TIRED. (Pause.) POINT TO THE PICTURE OF THE BOY YOU THINK MIGHT BE THIRSTY. (Pause.) GOOD. POINT TO THE PICTURE OF THE BOY YOU THINK IS HAVING FUN. (Pause.)

NOW TAKE YOUR PENCIL AND MARK AN X LIKE THIS (show on the blackboard) ON THE PICTURE IN THE FIRST ROW THAT SHOWS A PRODUCER. IF YOU DON'T KNOW, GUESS. NOBODY CARES IF YOU GUESS. (Check to see that children know what to do.) DON'T PUT AN X ON MORE THAN ONE PICTURE IN A ROW.

NOW POINT TO THE SECOND ROW OF PICTURES. (Pause and check.) PUT AN X ON THE PICTURE IN THIS ROW THAT SHOWS A PRODUCER. (Pause and repeat question.)

POINT TO THE THIRD ROW OF PICTURES. PUT AN X ON THE PICTURE THAT SHOWS A PRODUCER. (Pause. Encourage the children to guess if necessary.)

NOW FOLD YOUR PAPER BACK TO THE NEXT PAGE. (Pause and check.) ONE PICTURE IN EACH ROW SHOWS A PRODUCER. PUT AN X ON THE PICTURE IN EACH ROW THAT SHOWS A PRODUCER. (Pause.) WHEN YOU HAVE FINISHED THIS PAGE LAY YOUR PENCIL ON YOUR DESK. (Pause.)

FOLD YOUR PAPER BACK TO PAGE THREE. (Pause.) PUT AN X ON THE PICTURE IN EACH ROW THAT SHOWS A PRODUCER. (Pause.) WHEN YOU HAVE FINISHED THIS PAGE LAY YOUR PENCIL DOWN. (Pause. Check children's work by walking quickly around the room.)

FOLD YOUR PAPER BACK TO PAGE FOUR. (Pause.) THIS PAGE IS DIFFERENT. I WANT YOU TO LAY YOUR PENCILS DOWN AFTER YOU FINISH THE FIRST ROW. PUT AN X ON THE PICTURE IN THE FIRST ROW THAT SHOWS A PRODUCER. DO NOT DO THE SECOND ROW. (Pause.) LAY YOUR PENCILS ON THE DESK.

PUT YOUR FINGER ON THE SECOND ROW. NOW I WANT YOU TO PUT AN X ON THE CONSUMERS, BUT NOT ON THE PRODUCERS. PUT AN X ON THE CONSUMERS, BUT NOT ON THE PRODUCERS. (Pause. Check.)

PUT YOUR FINGER ON THE BOTTOM ROW. (Pause.) PUT AN X ON THE CONSUMERS, BUT DO NOT PUT AN X ON THE PRODUCERS.

FOLD YOUR PAPER BACK TO PAGE FIVE. (Pause.) PUT AN X ON THE PICTURE IN EACH ROW THAT SHOWS A CONSUMER. DO NOT PUT AN X ON THE PICTURES THAT SHOW PRODUCERS. LAY YOUR PENCIL DOWN WHEN YOU HAVE FINISHED THIS PAGE. (Pause.)

FOLD YOUR PAPER BACK TO PAGE SIX. (Pause.) PUT AN X ON THE PICTURE IN EACH ROW THAT SHOWS A PRODUCER OF GOODS. (Pause. Repeat. Check work.) LAY YOUR PENCIL DOWN WHEN YOU HAVE FINISHED THIS PAGE.

FOLD YOUR PAPER BACK TO PAGE SEVEN. (Pause.) PUT AN X ON THE PICTURE THAT SHOWS A PRODUCER OF GOODS. DO NOT DO THE LAST ROW. LAY YOUR PENCIL DOWN WHEN YOU HAVE FINISHED THE FIRST TWO ROWS. (Pause. Repeat. Check.)

ON THE LAST ROW PUT AN X ON THE PICTURE THAT SHOWS A PRODUCER OF SERVICES. (Pause. Repeat.) LAY YOUR PENCILS DOWN.

FOLD YOUR PAPER BACK TO PAGE EIGHT. (Pause.) PUT AN X ON THE PICTURE IN EACH ROW THAT SHOWS A PRODUCER OF SERVICES. (Pause.) PUT YOUR PENCIL DOWN WHEN YOU HAVE FINISHED THIS PAGE.

FOLD YOUR PAPER BACK TO PAGE NINE. (Pause.) DO ONLY THE TOP ROW. DO NOT MARK THE SECOND OR THIRD ROWS. PUT AN X ON THE PICTURE IN THE TOP ROW THAT SHOWS A PRODUCER OF SERVICES. (Pause.) PUT YOUR PENCILS DOWN WHEN YOU HAVE FINISHED THE TOP ROW.

IN THE MIDDLE ROW ARE PICTURES OF FIVE FARMS. THE MAN WHO OWNS ONE OF THESE FARMS IS NOT A SPECIALIST. PUT AN X ON THE FARM WHICH IS OWNED BY A MAN WHO IS NOT A SPECIALIST. (Pause. Encourage children to guess if they have to.)

IN THE BOTTOM ROW ARE PICTURES OF FIVE DOCTOR'S OFFICES. PUT AN X ON THE OFFICE OF A DOCTOR WHO IS NOT A SPECIALIST. PUT AN X ON THE ONE THAT IS NOT A SPECIALIST. (Pause.)

FOLD YOUR PAPERS BACK. IN THE TOP ROW PUT AN X ON THE BAKERY WHICH IS OWNED BY A MAN WHO IS NOT A SPECIALIST. (Pause.)

IN THE MIDDLE ROW, PUT AN X ON THE STORE WHICH IS OWNED BY A MAN WHO IS NOT A SPECIALIST. PUT AN X ON THE STORE WHICH IS OWNED BY A MAN WHO IS NOT A SPECIALIST. (Pause.)

LAY YOUR PENCILS DOWN. FOLD YOUR PAPER BACK. (Pause.) THERE IS A CONSUMER IN EACH PICTURE ON THIS PAGE. PUT AN X ON ALL OF THE CONSUMERS WHO ARE NOT PRODUCERS. PUT AN X ON ALL OF THE CONSUMERS WHO ARE NOT PRODUCERS. THERE IS A PRODUCER AND CONSUMER IN EACH PICTURE. DO NOT PUT AN X ON THE PRODUCERS. (Pause.)

FOLD YOUR PAPERS BACK TO PAGE TWELVE. THERE IS A PRODUCER IN EACH PICTURE ON THIS PAGE. PUT AN X ON ALL OF THE PRODUCERS. DO NOT PUT AN X ON THE PEOPLE WHO ARE NOT PRODUCERS. (Pause.)

FOLD YOUR PAPERS BACK TO PAGE THIRTEEN. PUT AN X ON ALL OF THE CONSUMERS. DO NOT PUT AN X ON THE PRODUCERS. (Pause.)

FOLD YOUR PAPER BACK TO PAGE FOURTEEN. LOOK AT THE FIRST TWO PICTURES IN THE FIRST ROW. PUT AN X ON THE BAKERY WHICH IS OWNED BY A SPECIALIST.

LOOK AT THE NEXT TWO PICTURES IN THE FIRST ROW. PUT AN X ON THE SHOE STORE WHICH IS OWNED BY A SPECIALIST.

LOOK AT THE MIDDLE ROW. PUT AN X ON THE FARM THAT IS OWNED BY A SPECIALIST. AND PUT AN X ON THE GARDEN THAT IS OWNED BY A SPECIALIST.

LOOK AT THE BOTTOM ROW. PUT AN X ON THE CARLOT THAT IS OWNED BY A SPECIALIST. PUT AN X ON THE STORE THAT IS OWNED BY A SPECIALIST.

FOLD YOUR PAPERS BACK TO PAGE FIFTEEN. PUT AN X ON THE PET SHOP THAT IS OWNED BY A SPECIALIST. AND PUT AN X ON THE DOCTOR'S OFFICE WHERE THE DOCTOR IS A SPECIALIST.

IN THE MIDDLE ROW, PUT AN X ON THE SCHOOL TEACHER WHO IS A SPECIALIST. AND PUT AN X ON THE NEWS STAND OPERATOR WHO IS A SPECIALIST.

IN THE LAST ROW, PUT AN X ON THE BIKE SHOP WHICH IS OWNED BY A SPECIALIST. AND PUT AN X ON THE TRAILER LOT WHICH IS OWNED BY A SPECIALIST.

FOLD YOUR PAPERS BACK TO PAGE SIXTEEN. PUT AN X ON THE FAMILY WHICH IS DIVIDING THE LABOR AS THEY CLEAN UP THE HOUSE. AND PUT AN X ON THE FAMILY WHICH IS DIVIDING THE LABOR AS THEY WASH THEIR CAR.

IN THE MIDDLE ROW, PUT AN X ON THE FARMERS WHO ARE DIVIDING THE LABOR. AND PUT AN X ON THE HOUSE BUILDERS WHO ARE DIVIDING THE LABOR.

IN THE LAST ROW, PUT AN X ON THE FAMILY WHICH IS DIVIDING THE LABOR AS THEY DO THE IRONING. AND PUT AN X ON THE FAMILY WHICH IS DIVIDING THE LABOR AS THEY CLEAN UP THE YARD.

ON THE LAST PAGE, PUT AN X ON THE FAMILY WHICH IS DIVIDING THE LABOR AS THEY FIX BREAKFAST. AND PUT AN X ON THE FAMILY WHICH IS DIVIDING THE LABOR AS THEY BUILD A FIRE.

ON THE LAST ROW, PUT AN X ON THE FAMILY WHICH IS DIVIDING THE LABOR AS THEY CLEAN THE HOUSE. AND PUT AN X ON THE FAMILY WHICH IS DIVIDING THE LABOR AS THEY STRAIGHTEN UP THE HOUSE.

(Due to limited supply, the picture test is not included in this copy of the dissertation.)

VITA

A. Guy Larkins

Candidate for the Degree of

Doctor of Education

Dissertation: Assessing Achievement on a First-Grade Economics Course of Study

Major Field: Curriculum and Supervision

Biographical Information:

1955-1957	Weber Junior College
1960-1961	University of Utah
1961-1962	Utah State University B.S.
1962-1963	Teacher, 4th, 5th, and 6th Grades, Grimmer Elementary School, Fremont, California
1963-1964	Teacher, 3rd Grade, Blacow Elementary School, Fremont, California
1964-1965	Teacher, 3rd Grade, Marshall Elementary School, Fremont, California
1962-1965	San Jose State College and California State College at Hayward
1965-1966	Graduate Assistant in Elementary Education, Utah State University
1966-1967	Research Associate in the Bureau of Educational Research, Utah State University
1967-	Assistant Professor of Educational Research Utah State University

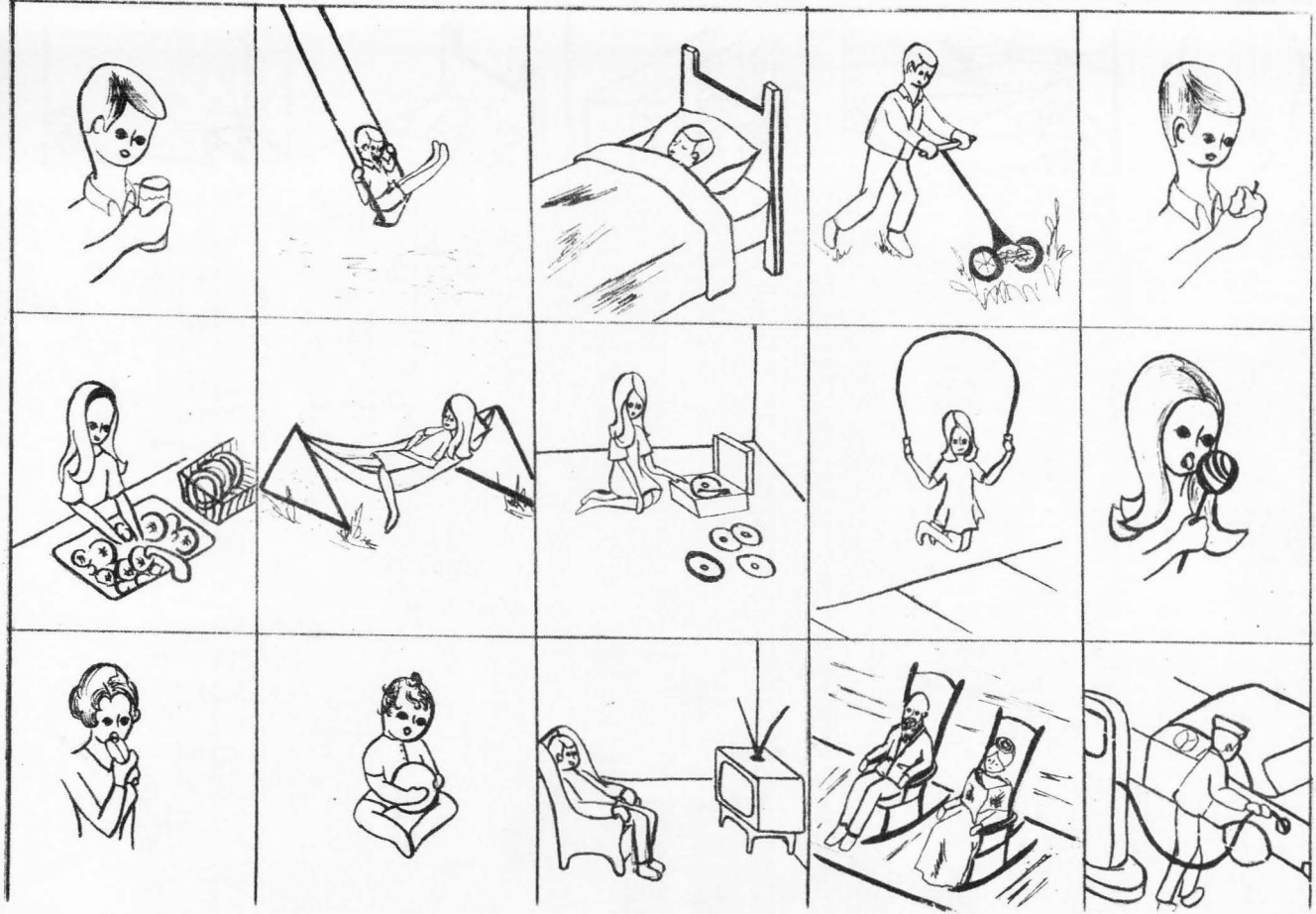
PET-1 (MULTIPLE CHOICE, FIRST EDITION)

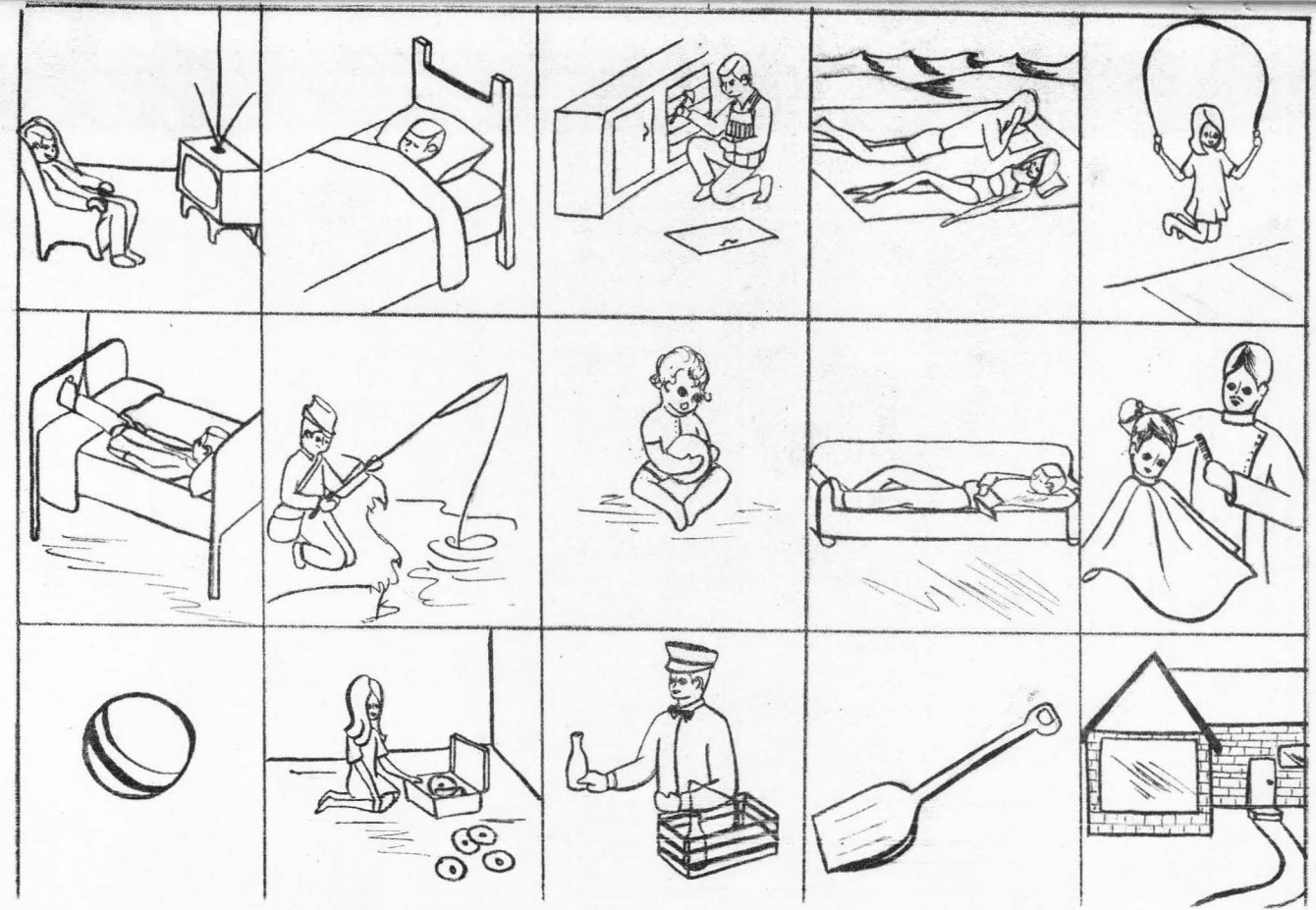
NAME _____

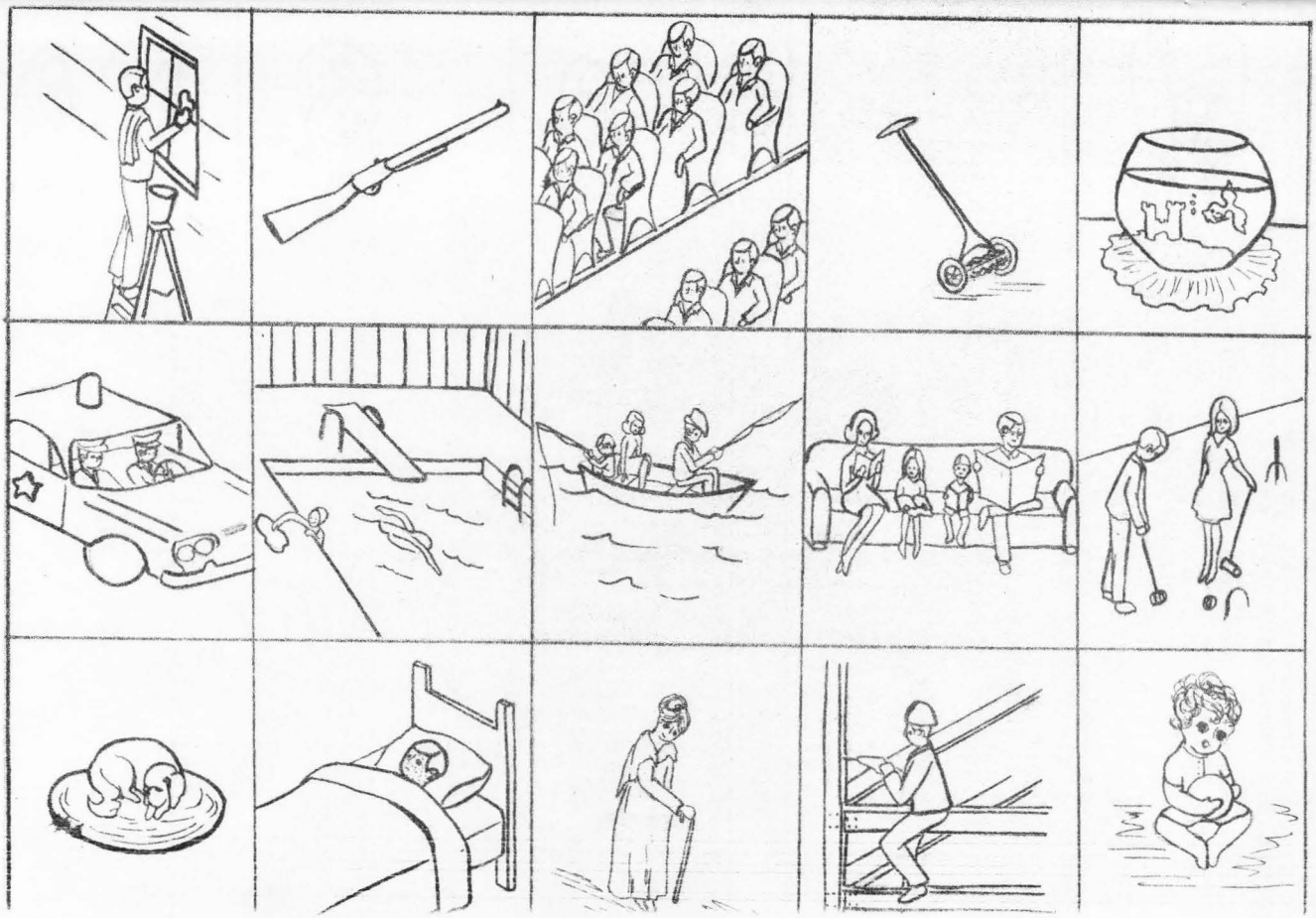
TEACHER _____

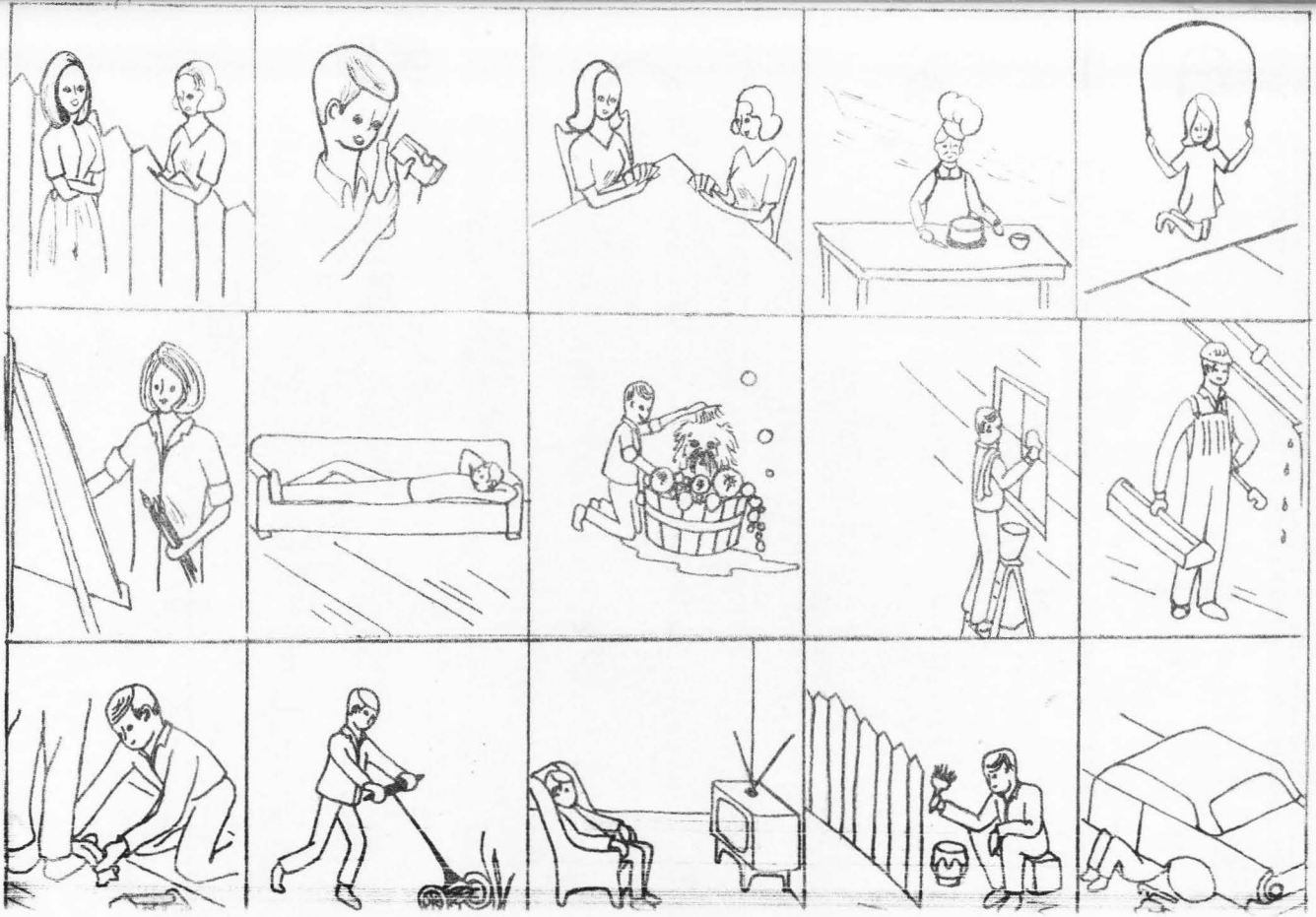
SCHOOL _____

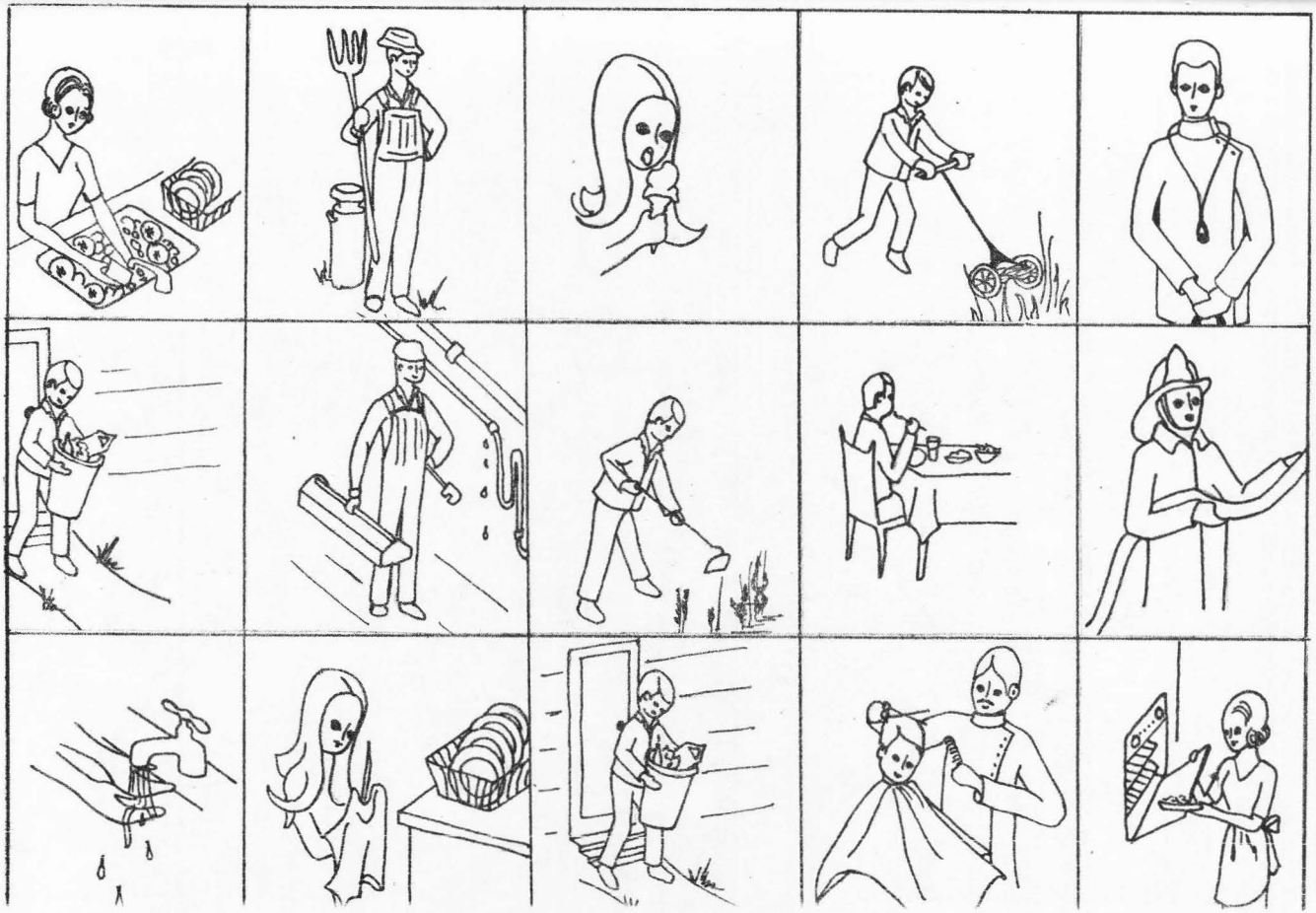
A. GUY LARKINS
JAMES P. SHAVER
BUREAU OF EDUCATIONAL RESEARCH
UTAH STATE UNIVERSITY
LOGAN, UTAH 84321

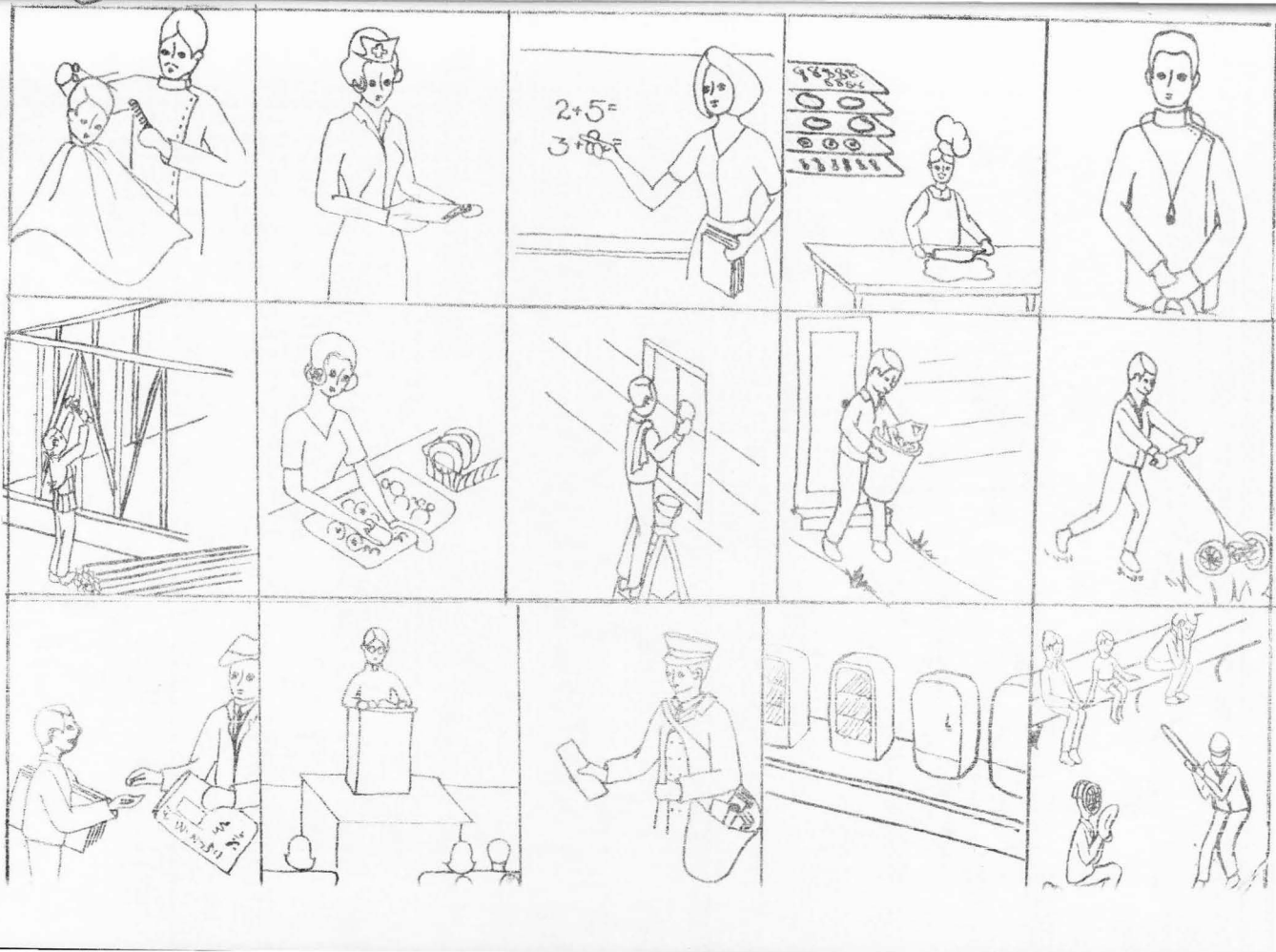


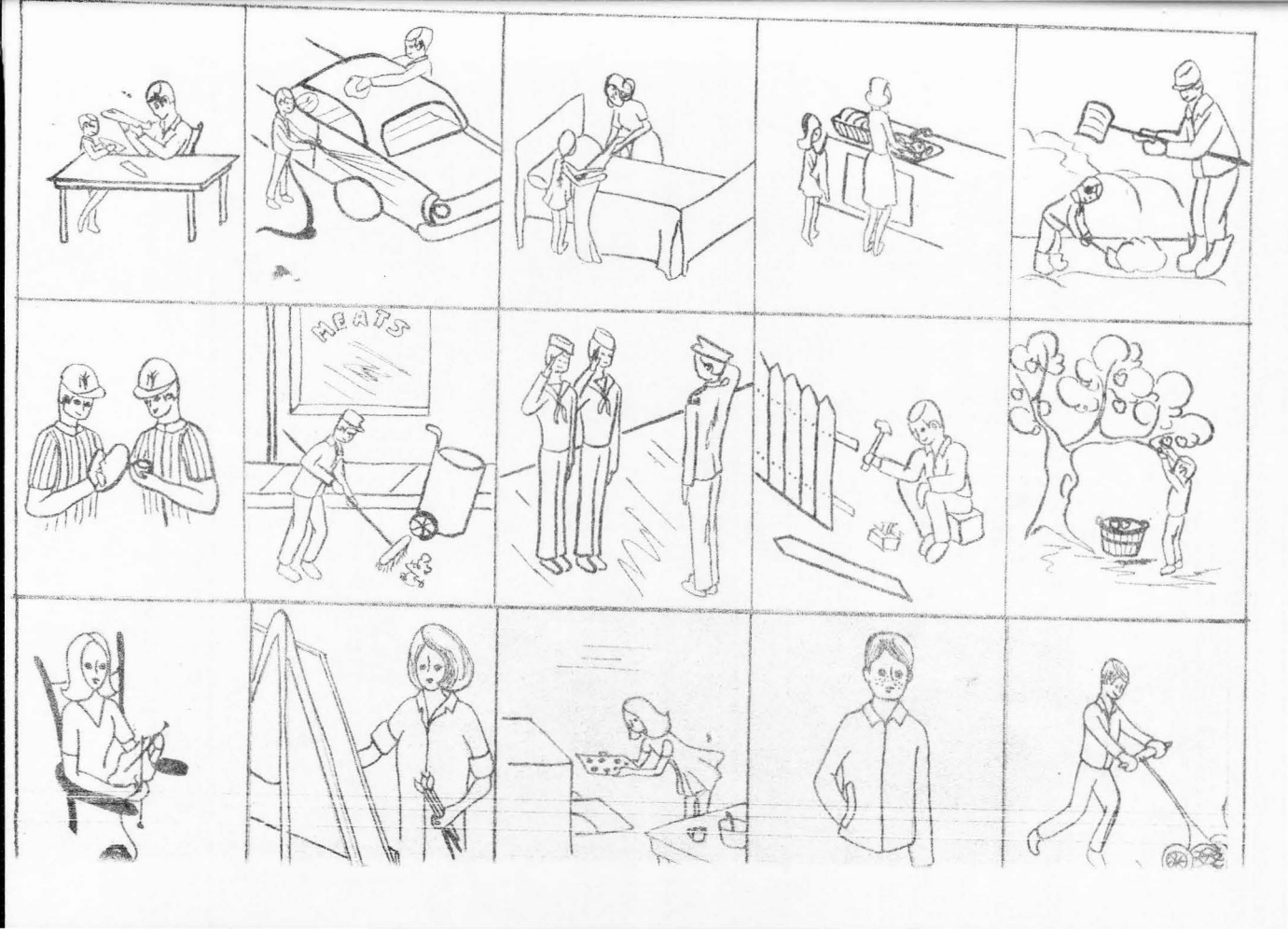


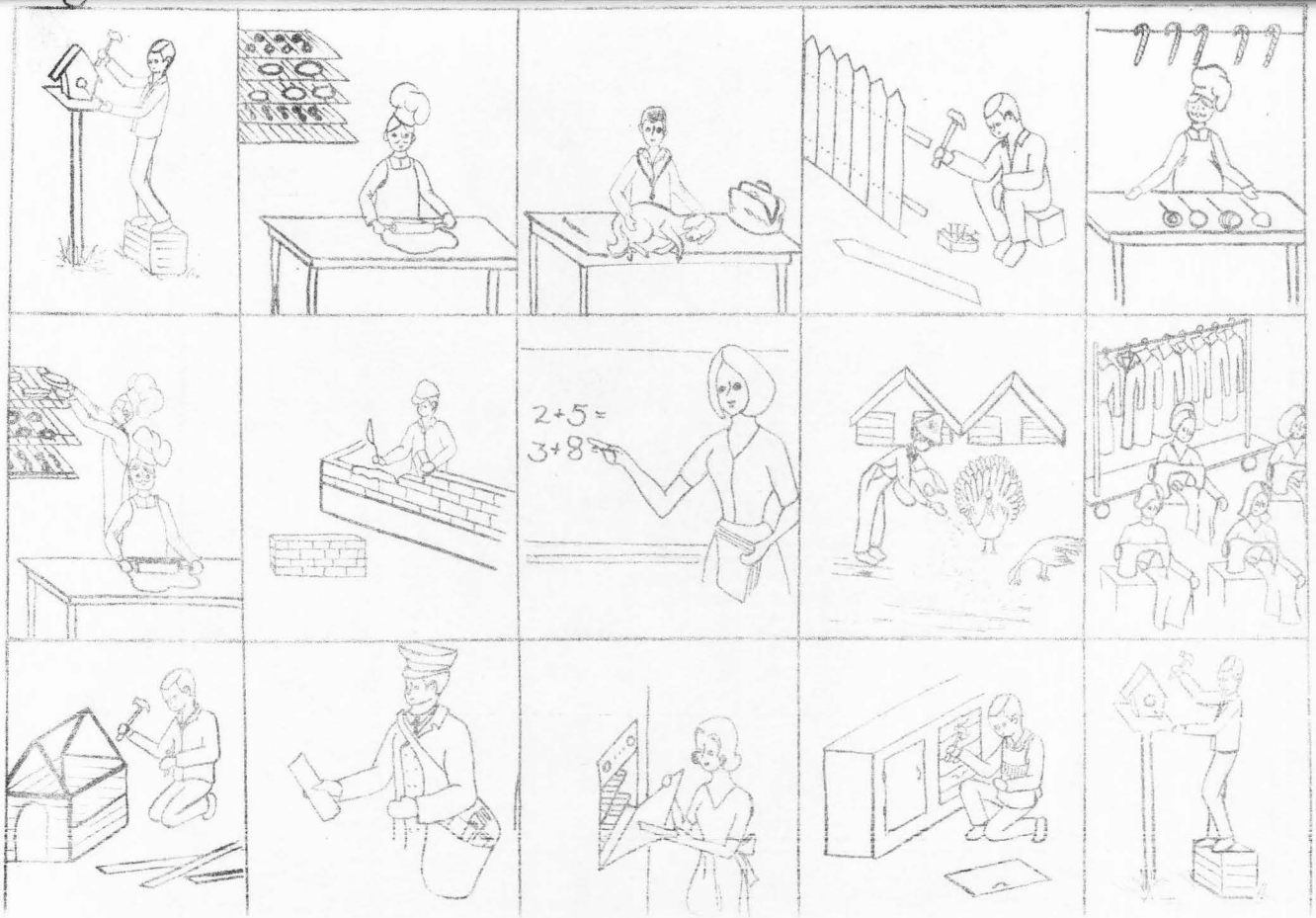


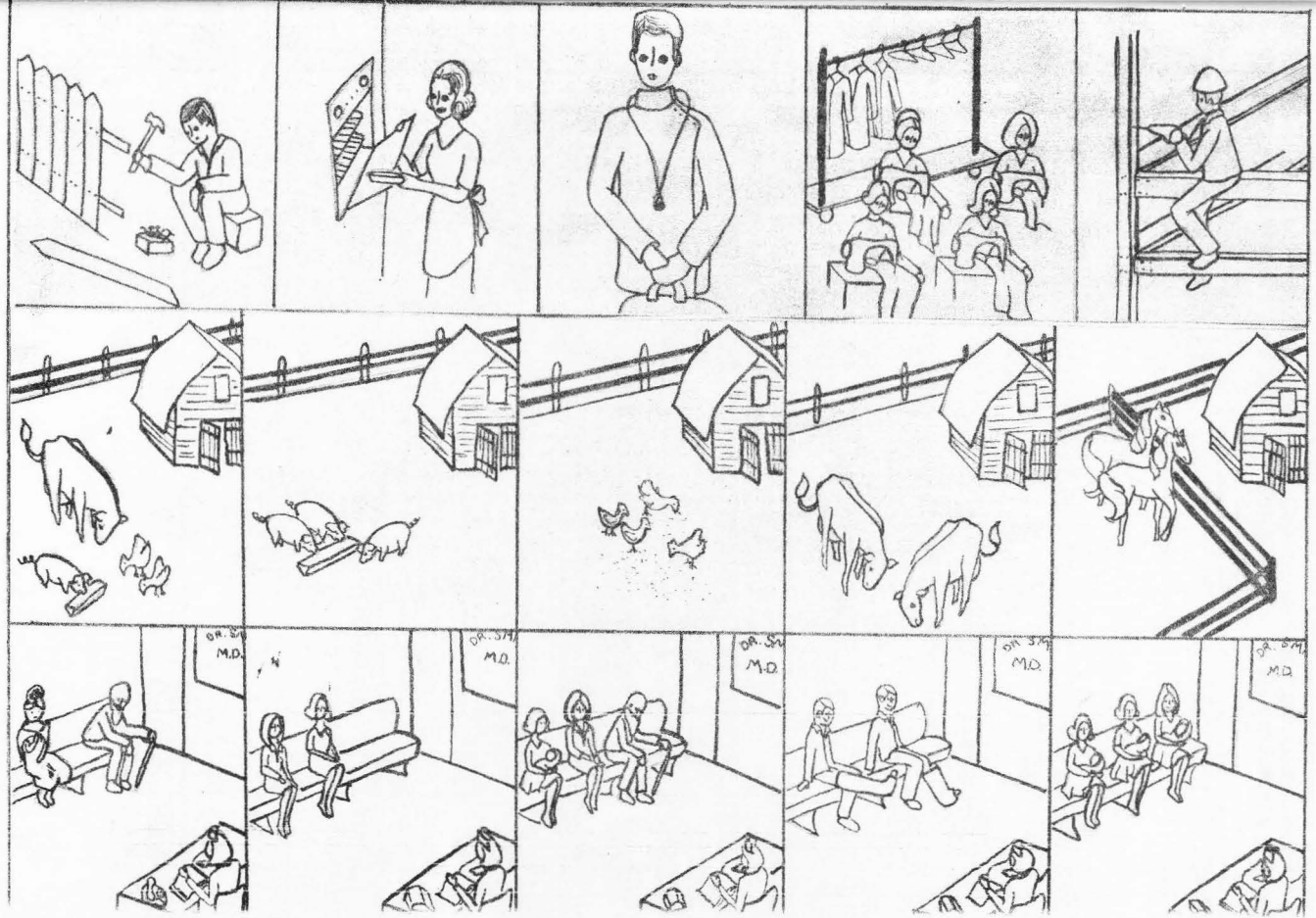


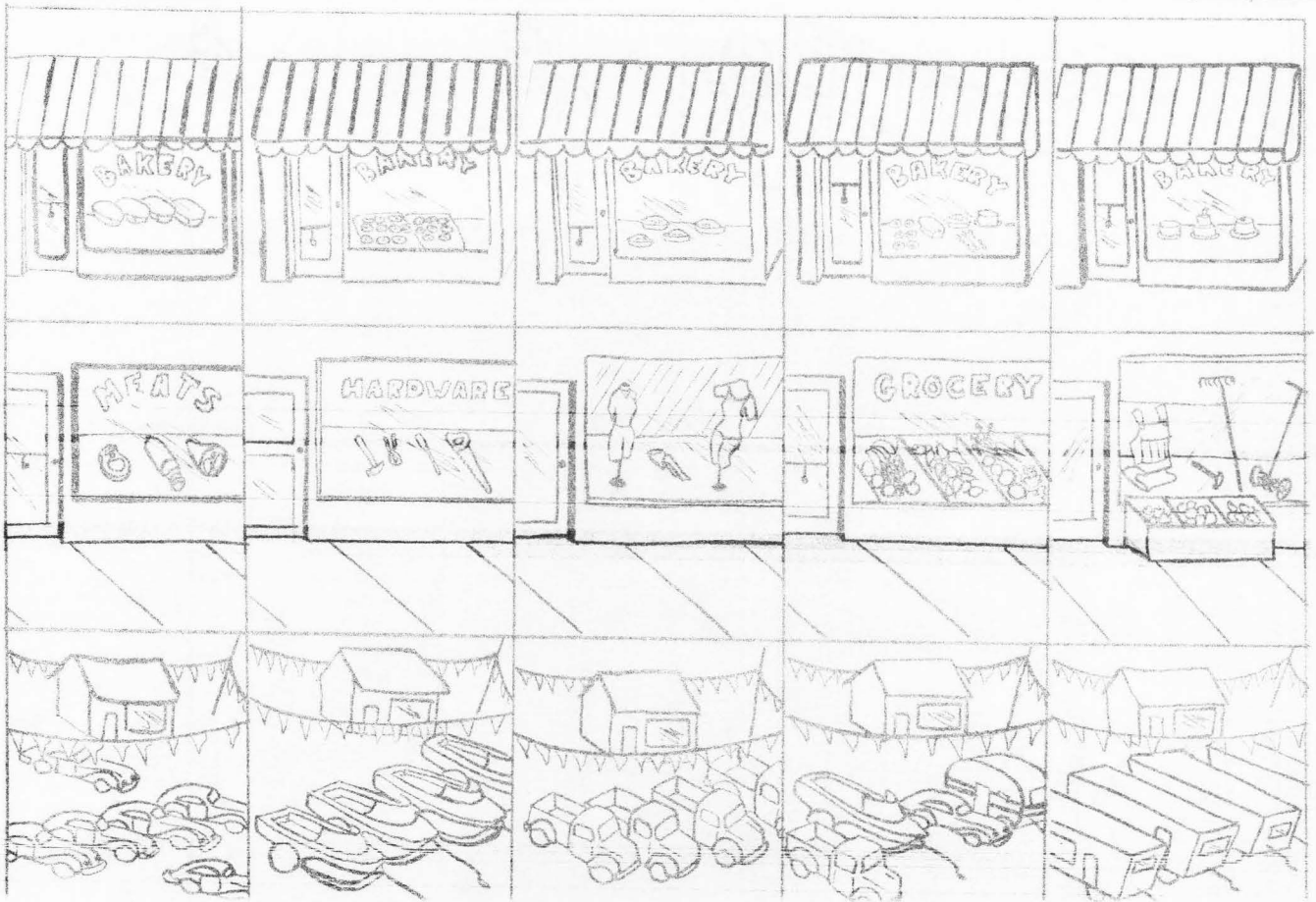


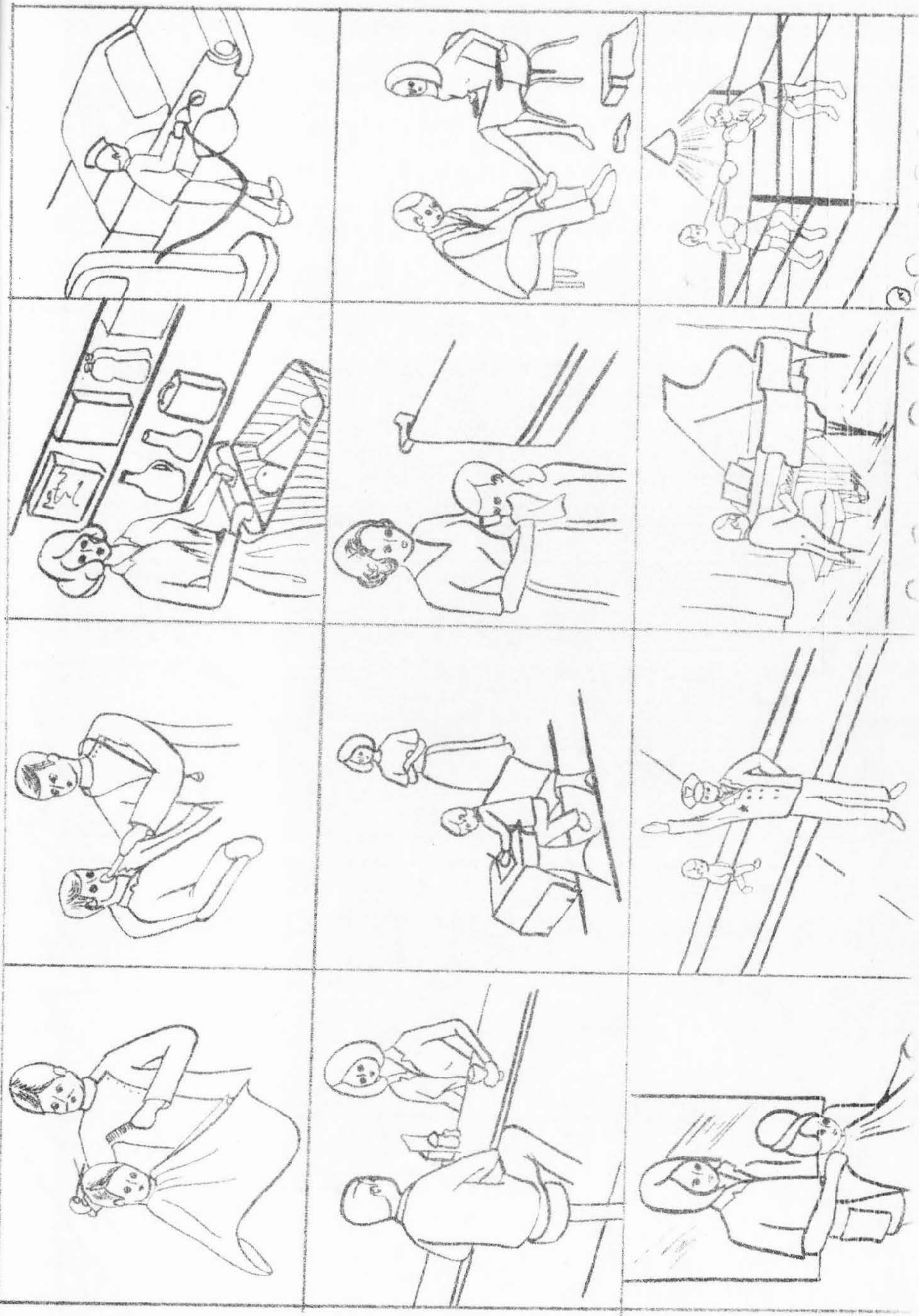

2+5=
3+

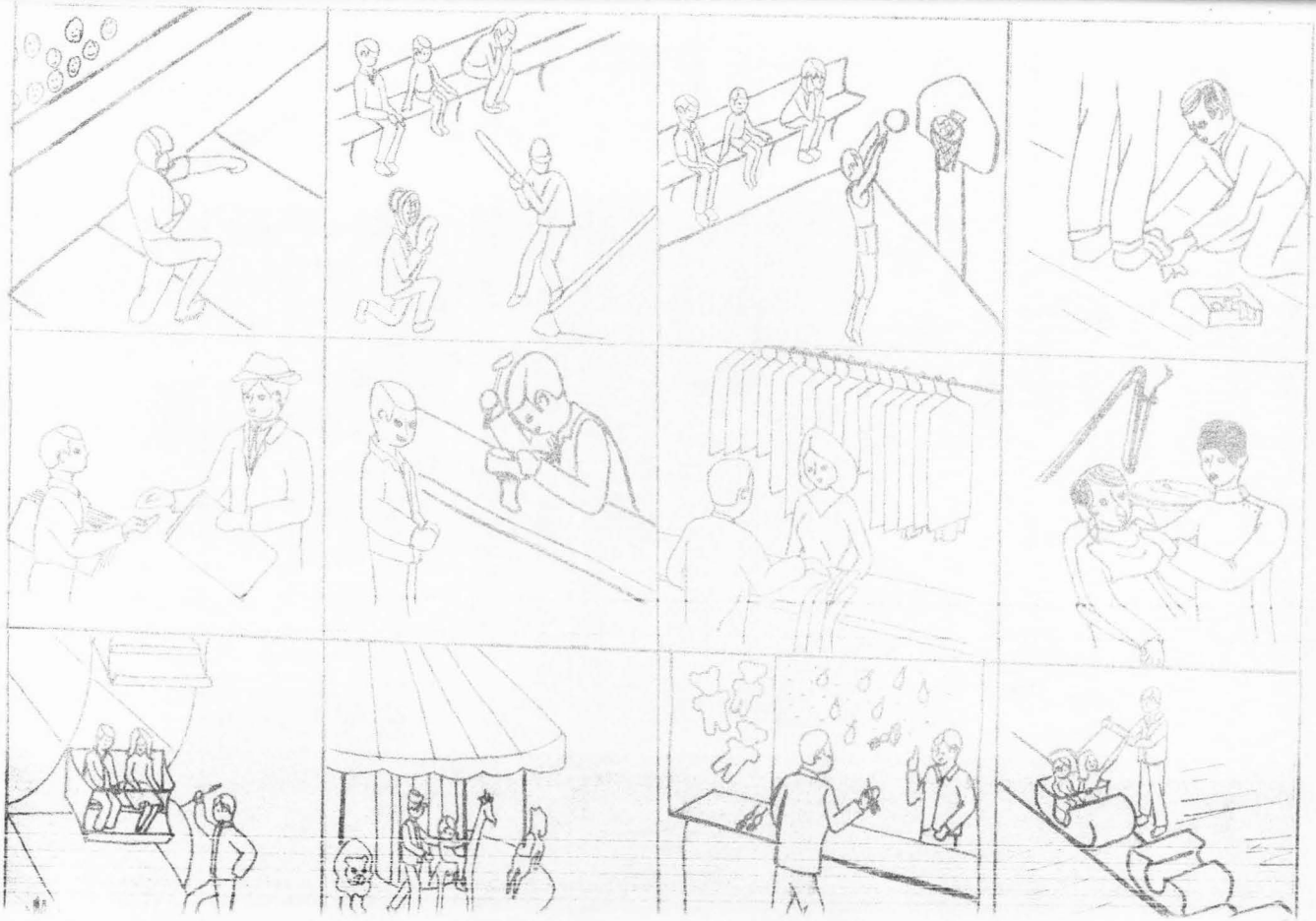

HEATS

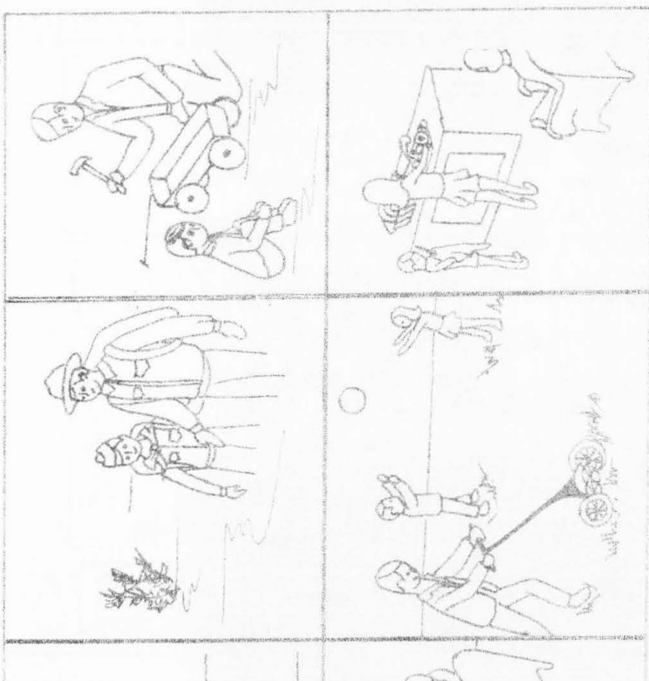

2+5=
3+8=

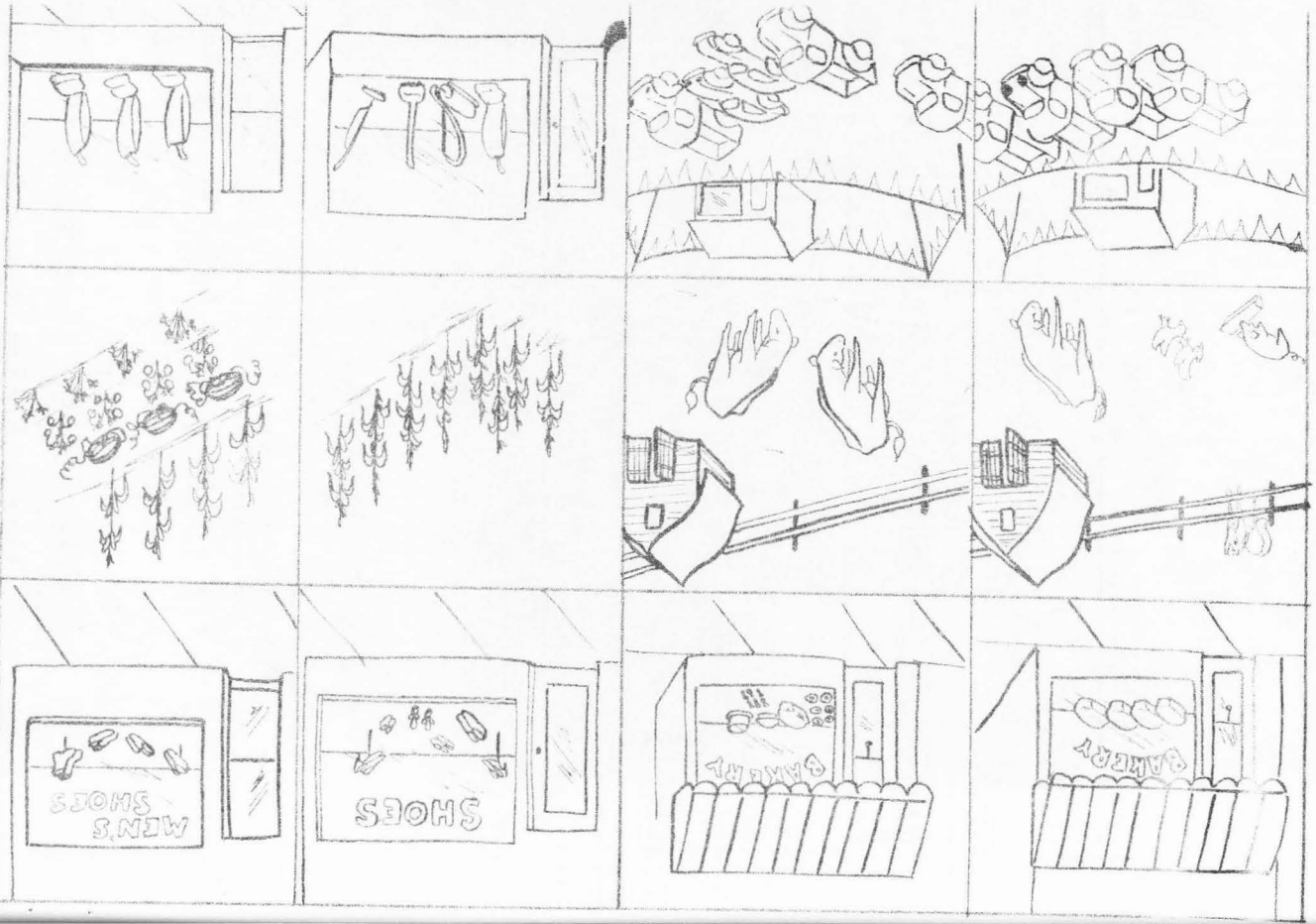

DR. SM
M.D.

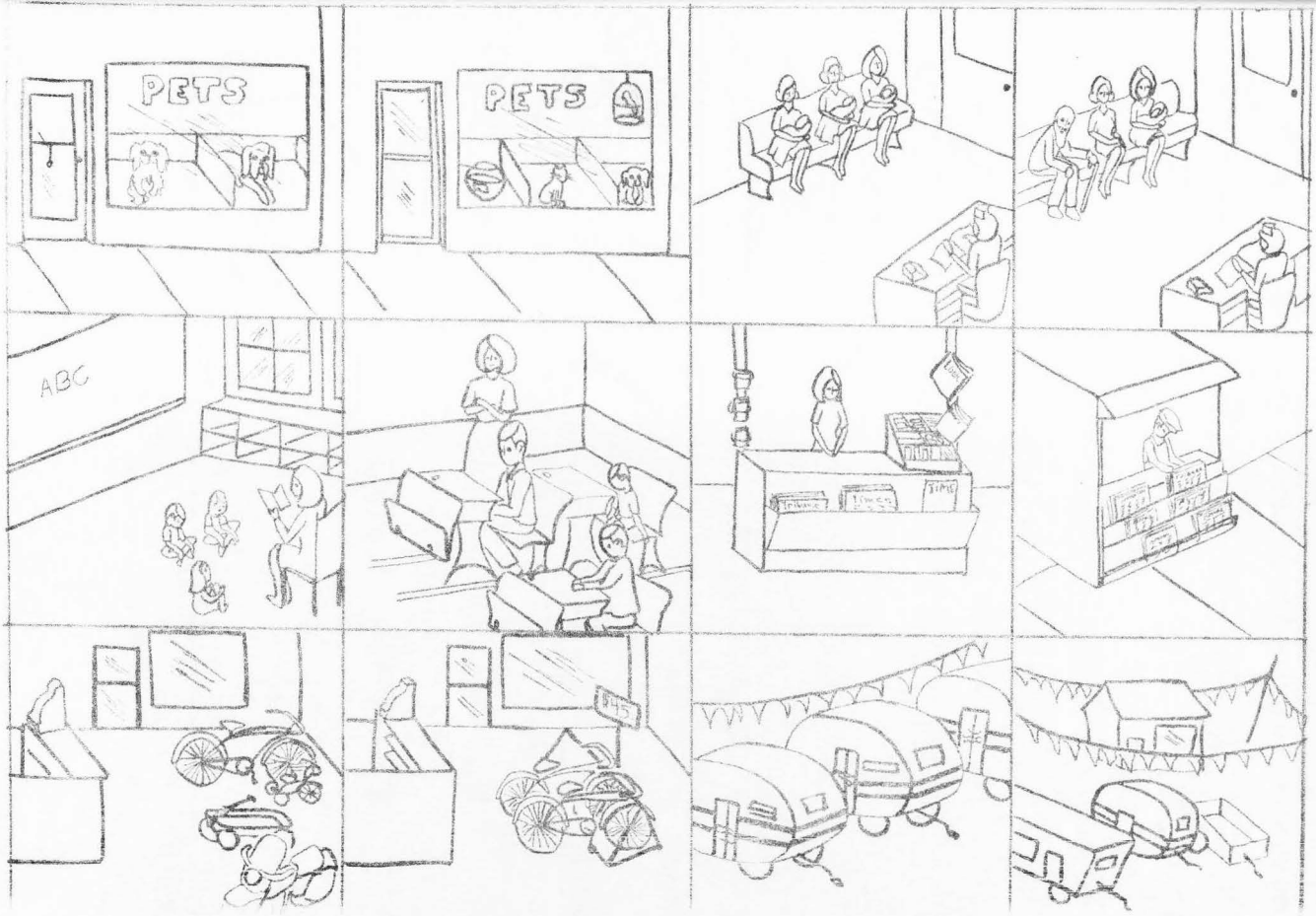

BAKERY
BAKERY
BAKERY
BAKERY
BAKERY
MEATS
HARDWARE
GROCERY








MENS
SHOES
SHOES
BAKERY
BAKERY


PETS
PETS
ABC