DOCUMENT RESUME

ED 069 189 FL 003 652

AUTHOR	Levin, Lennart
TITLE	Comparative Studies in Foreign-Language Teaching. Goteborg Studies in Educational Sciences 9.
INSTITUTION	Gothenburg School of Education (Sweden).
PUB DATE	72
NOTE	255p.
AVAILABLE FROM	Almqvist & Wiksell Publishers, Gamla Brogatan 26, Stockholm C, Sweden ($9.45)
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	Audiolingual Methods; Behavioral Objectives; Cognitive Objectives; Course Objectives; *Educational Experiments; Educational Objectives; *English (Second Language); Grammar; Instructional Program Divisions; *Language Instruction; *Second Language Learning; *Teaching Methods
IDENTIFIERS	GUME Project; *Sweden

ABSTRACT

This text describes a Swedish research program (GUME Project--the Swedish equivalent of the Gothenburg/Teaching/Methods/English Project) carried out during 1968-71 within the field of second-language instruction and learning. A number of comparative experiments were performed in order to assess the relative merits of two different approaches to teaching grammatical structures in English as a foreign language. The audiolingual habit theory and the cognitive code-learning theory are the principal methods examined in the study. Chapters include a detailed analysis of the GUME Project, experimental procedures and data, statistical analysis, characteristics of 10 experiments, evaluation instruments, and results. A list of tables, several appendixes, and a list of reference materials are included. (RL)

GÖTEBORG STUDIES
IN EDUCATIONAL SCIENCES 9

Lennart Levin

Comparative Studies in Foreign-Language Teaching

ALMQVIST & WIKSELL / STOCKHOLM

LENNART LEVIN

Comparative Studies in Foreign-Language Teaching

THE GUME PROJECT

ALMQVIST & WIKSELL
STOCKHOLM

Printed in Sweden by
Allfoto i Göteborg AB 1972

TABLE OF CONTENTS

List of Tables

List of Figures

PREFACE

The origin of the work to be presented in this book can be traced back to January, 1967, when seminars on foreign-language learning were initiated by Professor Alvar Ellegård, Head of the Department of English, University of Gothenburg. The seminars concentrated on problems of syntax learning, partly because the field was relatively unexplored, partly because new linguistic theories, making syntax the central area of linguistic research, simultaneously seemed to open up new perspectives on established foreign-language learning theories. The seminar found it worthwhile to investigate the tenability of two contrasting theories. Early in 1968 the so-called GUME project (the Swedish equivalent of Gothenburg/Teaching/Methods/English) became established; at the start it joined the now completed UME project at the Stockholm School of Education as a fairly independent cooperative part. During four years of research I have had the privilege of being leader of this project and of continuously receiving stimulating advice from Alvar Ellegård. His competence and good humour has been a great asset to me and to the project as a whole.

Research is never one man's or woman's job. In interdisciplinary research, of which the GUME project is a case, it simply should not be. The investigations contained in the present book would never have come about, had not a number of qualified colleagues and teachers painstakingly constructed numerous English lessons. I would like to take this opportunity to thank sincerely Ingvar Carlsson, Tibor von Elek, Torsten Lindblad, Margareta Olsson, and Mats Oskarsson for their inspiring cooperation and tolerant attitudes towards my often preposterous comments on matters linguistic.

Having moved from one department of education to another, I have had the somewhat unique privilege of receiving support from two professors of education. The personality of Kjell Härnqvist, Head of the Department of Education, University of Gothenburg, is such as to inspire any student of education to carry on within the field. I thank him heartily for his help and encouragement through some ten years. Karl Gustaf Stukát, Head of the Department of Education, Gothenburg School of Education, has closely followed the various facets of the GUME project and my research work connected with it. I am greatly indebted to him for constructive criticism and true fellowship.

In the summer of 1968, when the first GUME part projects were being planned, I had the rare opportunity of discussing research problems in second-language learning with professors John B. Carroll and Michael Wertheimer, USA, at the so-called SOLEP conference (Seminar on Learning and the Educational Process) near Stockholm. The ultimate GUME design proba-

bly lacks some of the sophistication they would have imparted to it, but the project had to be a compromise between the ideal and the possible. I thank them for their help during the early days of the project and for later encouragement in written form.

The investigations were made possible by grants from the National Board of Education, bureau L 4. I appreciate the ideas and the enthusiasm which the members of the bureau always brought with them to the sessions with the GUME staff.

The data were processed at the Computing Center, University of Gothenburg, on IBM 360/65. I owe a great deal to Per Högberg who wrote various computer programs and always provided me with results exactly on schedule.

Computer time was made available by the courtesy of Statskontoret and The University Chancellor's Office.

I should like to express here the appreciation of the members of the GUME project for the help and courtesy extended by Lumalampan Ltd, Stockholm, in matters concerning technical arrangements. We are also very greatful to Skrivrit Ltd., Stockholm, for permission to use copywright materials, to Sveriges Radio for permission to use materials from Skolradio programs, and finally to Skolförlaget Gävle for permission to use and adapt material from the "This Way" series of school books.

Bert Nilsson has corrected innumerable tests, made many a check calculation and assisted me in the daily work in various ways. I thank him heartily.

The pages of my manuscript have all passed through the agile fingers of Kerstin Davidsson, the charming GUME typist. My thanks are due to her.

Behind the figures in this book there are around two thousand pupils, one hundred teachers and their headmasters. I thank them all sincerely for their cooperation.

Finally, I would like to extend my warmest thanks to my wife May who, besides taking care of Karin, Dan, Ulf and Inger, has made me feel as if I never neglected family duties.

CHAPTER 1

INTRODUCTION

This thesis describes a research program carried out during 1968–1971 within the field of second-language teaching/learning. A number of comparative experiments have been performed in order to assess the relative merits of different approaches to teaching grammatical structures in English as a foreign language. During the same period a fierce debate on language pedagogy took place in Sweden. The present research is partly intended to shed some light on problems brought to focus in the course of that debate.

The GUME project

To this day six part studies, similar in design, have been performed within the GUME project. The first five, GUME 1–5, were undertaken at various age levels of the Swedish comprehensive school system whereas the sixth study in chronological order, GUME A, was performed at the adult level. Four of the five investigations at the comprehensive school level were made at the so-called upper stage where the pupils take one of two alternative courses, *sk* ("särskild kurs" = advanced course) or *ak* ("allmän kurs" = easier course). In those cases the studies consist of two parallel experiments, one at each course. Thus, a total of ten comparative experiments will be reported.

Three different strategies of teaching have been compared: (1) *the Implicit method* (Im), which is a kind of structure drill method where no explanations are given to the students, (2) *the Explicit-English method* (Ee), which provides explanations in the target language, and (3) *the Explicit-Swedish method* (Es), which gives explanations in the source language and comparisons with corresponding Swedish structures.

The teaching strategies compared in our studies represent two different types of linguistic theory and two correspondingly distinct lines of teaching methodology. Although the teaching procedures as well as the measuring instruments and technical arrangements varied somewhat between different part studies, the two lines of thinking are reflected in each of our experiments.

The current report

Interim reports giving detailed accounts of the design, procedures and results of the majority of our part studies have been published earlier (see Appendix 1). However, we have felt a need to give a more comprehensive view of the research activities. This for at least three reasons.

Firstly, the experiments form a research program where the successive modifications in design, experimental procedures, lesson content, etc., were caused by experiences made in the course of the project. We feel that the different facets of the research, as well as their interdependence, should be taken into account when the accumulated evidence from the project is interpreted. This is most easily done if the separate experiments are brought together into one volume.

Secondly, it is generally assumed that children and adults learn a second language in different ways and should, accordingly, be exposed to different teaching methods. Although the exact time of "linguistic puberty" is rather unclear, it is obvious that the GUME project includes experimental groups on both sides of this critical point and only a comprehensive description of the project can provide evidence on the relation between age and teaching method.

Thirdly, recalculations of earlier data have been made, partly with the intention of applying techniques not utilized in our earlier analyses, partly with the intention of treating the different part studies analogously as far as possible.

The report comprises a fairly large bulk of data. It is impossible to include, for reasons of space, complete descriptions of the tests, questionnaires and teaching procedures used in the various part studies. We will follow the principle of pointing out essentials and of giving illustrative examples; however, reference will be made to previous GUME reports (see Appendix 1) in order to facilitate checks when we find them necessary or otherwise informative.

Plan of report

In chapter 2 the two foreign-language learning theories alluded to above will be described and discussed. In the same chapter we will treat the concept of teaching method, both in general and with special reference to teaching a foreign language. Here we will also survey some earlier foreign-language teaching methods. The Swedish debate within this area and its relation to the present curriculum are commented on in chapter 3, whereas chapter 4 contains a review of earlier research on the effectiveness of different methods of

teaching foreign languages. In chapters 5–10 various aspects of our experiments are presented; chapter 5 contains a discussion of our considerations in choosing research approach, and chapter 6 describes the statistical techniques used. In chapter 7 a brief historical sketch of the GUME investigations is given in order to provide the reader with an outline of our research activities. Chapter 8 contains a detailed account of the ten experimental samples and an attempt at judging the internal and external validity of our experiments. In chapter 9 and 10 our independent and dependent variables, i.e. the lesson series and criterion tests respectively, are presented. Chapter 11 is an account of the main results of our teaching method comparisons, both with respect to learning effects and attitudes to the various treatments. In chapter 12 some additional findings are presented; a follow-up study is discussed, some results related to choice of course (sk/ak) are presented, and a number of correlations are analysed. Finally, our findings and their eventual implications for further research and pedagogical measures are discussed in chapter 13. Chapter 14 contains a summary of the GUME project and its results.

CHAPTER 2

THEORY AND METHOD IN FOREIGN-LANGUAGE TEACHING

The theoretical dichotomy

Our teaching strategies, which will be described in detail later, approximately correspond to *the cognitive code-learning theory* and *the audio-lingual habit theory* (Carroll, 1965). The two theories disagree on two fundamental points: (1) what language is, and (2) how it is acquired. The two theories and their alleged relevance for foreign language teaching will be discussed in due course; here we only want to draw attention to the fact that the foreign language teaching debate, in Sweden and elsewhere, has displayed a dichotomy of opinion similar to the one represented by the mentioned theories, namely a mentalistic versus a mechanistic orientation. The theories and their methodological equivalents are not entirely new, nor is the methodological controversy. The exposé of foreign language teaching methods in the following section is aimed at illustrating this fact and at putting the methods utilized in our studies in a proper perspective.

Mackey (1965), when summarizing his chapter on the development of language teaching, says:

> "If we now glance back at the development of language-teaching method, we see that it first swings from the active oral use of Latin in Ancient and Medieval times to the learning by rule of the Renaissance grammars, back to oral activity with Comenius, back to grammar rules with Plötz, and back again to the primacy of speech in the Direct Method" (p. 151).

Although Mackey ends his survey here, there are still other "swings of the pendulum" which will become apparent from the following discussion.

Historical sketches of the kind that Mackey makes are to be questioned as scientific documents. The "swing of the pendulum" phenomenon, though acceptable as a pedagogical device, seems a too simplified description of a probably very complex evolution. However, the dichotomy underlying his survey of teaching methods seems to be accepted by others. Rivers (1968), in a similar overview of foreign language teaching methods, also distinguishes between two main streams of thought. For convenience, she terms the representatives of the two groups *formalists* and *activists*. Since her distinction

bears on the methods contrasted in our experiments, we shall quote her at some length:

> "Formalists have mostly relied on a deductive form of teaching, moving from the statement of the rule to its application in the example; activitists have advocated the apprehension of a generalization by the student himself after he has heard and used certain forms in a number of ways – a process of inductive learning. Formalists with a commendable regard for thoroughness have sometimes become too preoccupied with the pedantic elaboration of fine details of grammar, whereas activists have consistently urged a functional approach to structure whereby the student is first taught what is most useful and most generally applicable, being left to discover at later stages the rare and the exceptional . . . . . These divergent attitudes toward various aspects of foreign-language teaching have led to a very different order of priorities in the teaching of the four skills, the formalist tending to value highly skill in reading and accurate writing (especially as demonstrated by the ability to translate), the activist laying emphasis on oral understanding and speaking as basic to fluent reading and original writing" (pp 12–13).

As it appears, the above mentioned division of method is reflected in Carroll's statement that there are to-day two major theories of language learning, the audio-lingual habit theory and the cognitive code-learning theory (Carroll 1965). It should be noted that, in Carroll's opinion, they are not theories in a stricter sense but rather summarizing descriptions of the practices of foreign language teachers and the writings of several theorists. The following quotation from Carroll (op.cit.) serves to illustrate the similarity between him and Rivers as far as the inductive-deductive polarization of teaching strategies is concerned:

> "The audio-lingual habit theory, which is more or less the 'official' theory of the reform movement in foreign language teaching in the United States, has the following principle ideas: (1) Since speech is primary and writing is secondary the habits to be learned must be learned first of all as auditory-discrimination responses and speech responses. (2) Habits must be automatized as much as possible so that they can be called forth without conscious attention. (3) The automatization of habits occurs chiefly by practice, that is, by repetition. The audio-lingual habit theory has given rise to a great many practices in language teaching: the language laboratory, the structural drill, the mimicry-memorization technique, and so forth. The cognitive code-learning theory, on the other hand, may be thought of as a modified, up-to-date grammar-translation theory. According to this theory, learning a language is a process of acquiring conscious control of the phonological, grammatical, and lexical patterns of a second language, largely through study and analysis of these patterns as a body of knowledge" (p. 278).

This strong theoretical dichotomy comes close to what has been characterized as a "paradigm clash" by Katahn & Koplin (1968). The authors discuss two competing paradigms in contemporary psychology which are in fact

related to the two theories discussed above. The differences between them seem to focus primarily on the relative weight that internal information processing events will play in theoretical accounts of behavior, as contrasted with emphasis upon objective description of environmental events. Supporters of each paradim or theory often tend to "see" the domain so differently that arguments pass through the other point of view and do not make meaningful contact. The authors observe that is it usually impossible, *on logical grounds,* to accept one theory and reject the other. The two divergent theoretical positions in the case of foreign-language learning provide the conceptual setting in which the present investigation has been conducted. In our opinion there is no a priori reason for predicting which method (depending on its theoretical background) will come out as the best in an actual learning situation. In statistical terminology, there is no ground for applying one-tailed tests in our method comparisons.

The concept of method

We shall return to the question of theory in connection with foreign-language learning, both from a linguistic and a psychological point of view. First, however, a comment will be made on the concept of method, and thereafter some methods will be considered. Mackey, (op.cit), in his historical survey, mentions fifteen methods, most of which are still in use in one form or another in various parts of the world. What is perhaps more interesting than Mackey's account of the many methods and their characteristics, is his discussion of the vagueness of the concept of method. Such terms as "the Direct Method", the "Natural Method" the "Linguistic Method", etc, are diffuse and inadequate because they usually limit themselves to a single aspect of a complex subject. He suggests that method analysis be made in terms of: (a) selection (b) gradation (c) presentation and (d) repetition of teaching materials. It is through these four inherent characteristics that one may discover how one method differs from another. With the aid of a so-called method profile he tries to quantify method (op.cit, pp 317–318). The profile is elaborate and somewhat difficult to read: its main advantage seems to be that various aspects of the teaching process are treated separately; thus no generalized description of language teaching in all its variety (vocabulary, phonology, grammar, etc) is aimed at.

Casey (1968) also adpoted the method profile idea in order to define method as it relates to foreign-language teaching. His profile is a kind of opinion scale where teachers indicate their position in methodological matters on a continuum ranging from acceptance of the cognitive code-learning theory (–20) to acceptance of the audio-lingual habit theory (+20). Scores in

the middle range, approximating zero, indicate an acceptance of neither theory or a partial acceptance of the two.

In a work paper by Smith (1970) any foreign-language teaching method, whether good or bad, is said to contain the following four elements: (a) presentation (b) explanation (c) repetition and (d) transfer; it is in ordering, emphasis and style of these four steps that methods differ. Incidentally, Smith states that disagreement is especially strong to-day among language teachers about the presentation-explanation or explanation-presentation order. Accumulating evidence is said to support the greater effectiveness of the explanation-presentation order within the population of above average intellectual abilities found in secondary schools, universities, and Peace Corps training.

Wallen & Travers (1967), when discussing the problem of identification of teaching method, state that the variables involved in most studies reflect few of the properties of well-defined scientific variables. The implication is that often no real differences exist in the patterns of behavior manifested by teachers representing different methods. The authors stress that the concept of method may be deceptive, indicating the existence of easily identifiable characteristics of one approach as distinct from another.

> "All too often the unreasonable assumption is made that, because a teaching method has been described, corresponding patterns of behavior can be, or are, manifested by teachers." (p. 467)

An article by Gage (1969) on teaching methods is of limited interest in our present discussion since it deals with teaching method in the most general sense, i.e. as patterns of teacher behavior applicable to all subjects. However, he discusses the problem of concern here, namely

> "the problem of finding ways to compare methods along basic underlying dimensions so that the difference between them can be more clearly identified and their effects can be closely associated with those differences ..... It is necessary to penetrate beneath the global terminology referring to 'methods' to the specifics of teacher and learner behavior for which the terms stand' (p. 1450).

The conclusion seems warranted that most research on teaching methods has had the notorious deficiency of imprecise description of the methods compared, thereby increasing the risk of unjustified generalizations about their relative merits. Bosco and Di Pietro (1970) have pointed to the dangers in treating foreign-language teaching method in a global perspective. The following quotation should serve as a word of caution to anybody planning broad comparisons of teaching methods:

> "We are convinced that research which attempts to demonstrate the superiority of one strategy over any other is misdirected because of the multiplicity of features underlying each strategy and the problem of co-occurrence of features across strategies. Any effective evaluation must be done in terms of *feature* of strategy rather than of strategy considered as a global entity. Although it is theoretically possible that the uniqueness of a strategy depends on a single feature, this will not prove to be the case in practice" (p. 3).

Del Olmo (1968) maintains that if the teaching procedures of a certain method are not presented in full detail, it becomes a method *in abstracto*. He severely criticizes Wilga Rivers' audio-lingual *method* as it is described in her book *The Psychologist and the Foreign Language Teacher* (1964):

> "She has somehow set up an audio-lingual straw man who becomes such a perfect embodiment of the Audiolingual Canon that he is nowhere to be found" (p. 19).

Del Olmo also criticizes the Scherer and Wertheimer study, which will be mentioned later (see pp. 47–48), for incomplete documentation of the differences between the methods compared.

The problem of concern here has been discussed by Siegel & Siegel (1967) in terms of independence and homogeneity of experimental treatments. A teaching procedure must be independent of others and homogeneous within itself in order to be of use as an independent variable in a comparative experiment. The authors warn against the use of grossly designated methods and urge that the treatments be described in *procedural* terms. Gage (1967) has distinguished between "criteria of effectiveness" and "process" paradigms for research on teaching. Although comparative educational research has mostly been based on the criteria of effectiveness paradigm, there seems to be a tendency now for the process-variety of research to appear. The following studies within the field of second-language teaching utilized various process-oriented techniques in order to achieve precision in the description of teaching procedures.

Jarvis (1968) developed an observation system for classroom foreign language skill activities based on time sampling. The instrument was also used to investigate the teacher's adherence to a certain teaching model. Moskowitz (1968) and Wragg (1970) attempted to adapt the Flanders system of interaction analysis to the foreign language classroom. Hayes and others (1967) developed a plan for language teaching evaluation based on direct observation of teaching in progress. Avoiding the oscurity of the term method they propose a different terminology:

> "Henceforth we shall use the term *(feature(s)* to refer to one or more policies, principles, or procedures viewed independently; we shall use the term *teach-*

> *ing profile* or simply *profile* to refer to a particular array of policies, principles and procedures (features) as they might be found in a given instructional setting . . . . . We also use the term *subprofile* to refer to recurrent variations in detail as they might be found in different classes in the same instructional setting" (p. 23).

The check-list used by the observers contained 324 items; the observation technique can be supposed to be rather time-consuming and exacting. It was used in a survey of 364 faculty members of NDEA institutes, where a strong consensus was obtained in favor of practices that stem from the audio-lingual method.

Obviously the majority of comparative educational research has not fulfilled the demand for adequate description of the independent variables, a fact which has complicated much of the methodological discussion. It is obvious that great caution must be observed in interpreting the results of comparative studies. It is equally obvious that the experimenter, in reporting his results, should describe the teaching techniques as completely and accurately as possible in order to avoid faulty interpretations. Although the designations of the "methods" compared in our studies are neutral in relation to current terminology, there is still a risk that teachers will identify them according to personal preferences. With the hope to avoid this, we shall give as detailed accounts as possible of the treatments used; in one case a whole lesson sequence will be described (see chapter 9).

Some foreign-language teaching methods

With the vagueness of the concept of teaching method in mind we shall now proceed to give a brief account of some of the foreign-language teaching methods appearing in the literature. Our survey will not go further back than to the beginnings of modern practice.

The grammar-translation method was used in most schools toward the end of last century. It is impossible to trace the method back to an originator; it has its roots in the formal teaching of Latin and Greek during the centuries. According to various sources (for instance, Mackey 1965, Rivers 1969, Titone 1968) its main features are: The teaching begins with rules, isolated vocabulary items, paradigms and translation. Vocabulary is divided into lists of words to be memorized but there is little relationship between the vocabulary of successive lessons. Pronunciation is either not taught, or is limited to a few introductory notes. Grammar rules are memorized as units, which often include illustrative sentences. The main defect of the method seems to be the neglect of communication skills. There is a great deal of stress on knowing

rules and exceptions, but the method gives limited training in using the language actively. Rivers (1968) comments that the method is not too demanding on the teacher; when he is tired, he can always set the class a written exercise. In our opinion the quoted characterizations of the grammar-translation method, all being made during the 1960's, tend to become slight caricatures of what probably happened in the classrooms. It is difficult to conceive of a teaching procedure involving no oral practice but rule memorization for its own sake.

Around 1880 Viëtor incorporated descriptive phonetics into a language-teaching method. He severely criticized the grammar-translation method of his day and suggested a new method, based on the spoken language. Knowledge of grammar was to be acquired inductively through the reading of texts. At about the same time Gouin had proposed a method where the element of physical activity was added to the teaching; each sentence was to be acted out while it was being uttered. Both Gouin and Viëtor stressed that sentences to be read should form a meaningful and motivating context and not be taught in isolation. Their ideas combined in a new method, called *the phonetic method* or *the reform method*, which became a source for the elaboration of the direct method. The Berlitz language school, established in 1878, may be viewed as another fore-runner of the direct method. The school offered conversational skill in the foreign language, which the curriculum and the methods of the ordinary schools had failed to give. Berlitz schools for languages, advertising "total immersion" courses where the foreign language is spoken by pupils and teachers from the very first lesson, exist all over the world to this day.

The direct method, created as a protest against the grammar-translation method, was at first quite disorganized. The principles of Viëtor and Gouin were over-simplified in practice, and the method was confused with the various "natural" and "oral" methods which developed simultaneously. "The teacher took the place of the book, had no technique of teaching through actions, and on the whole, did whatever he pleased" (Mackey, op.cit., p. 145). However, at the turn of the century the method began to follow a more definitive pattern. Since the new emphasis was on the foreign language as the medium of instruction, the mother tongue of the pupils was ruled out in the instruction, and understanding of the foreign language was arrived at by demonstration. Grammar was to be learned inductively just as when a child learns his mother tongue, and listening to and speaking the language became primary to reading and writing it. As Rivers (1968, p. 20) has pointed out, the method provided an exciting and interesting way of learning the foreign language through activity, and it proved successful in releasing students from the inhibitions associated with speaking a foreign language. Its main defect, according to Rivers, was that it forced the pupil to express himself too soon

in the foreign language in a relatively unstructured situation; there was not sufficient provision for systematic practice of structures in a planned sequence. As the principles of the direct method spread they were modified in various ways. Grammatical explanations, given in the native language, were introduced to meet the demand for accuracy; translation was included and systematic grammar drills were added. Rivers comments that the various modifications of the direct method are similar to what has been called the eclectic method. A modified direct method for the use in Swedish schools has been outlined by Hensjö (1964).

Even in as brief a review of teaching methodology as the present one, mention must be made of three late nineteenth century linguists whose corporate view is scarcely distinguishable from what is considered by many to be good foreign-language teaching to-day: Henry Sweet, Otto Jespersen and Harold Palmer. Palmer, who developed the work of Sweet and Jespersen into a coherent system during the early 1900's, came in on the crest of the wave of the direct method and brought it back into proportion. Roddis (1968, p. 333 ff) has compared the views of the three in various aspects of foreign language teaching. All three point to the interference of the pupil's native language. As a matter of fact, Palmer placed such interferences at the heart of the language learning problem. The three authors were aware of the problem although they stated the teacher's function in rather general terms. Concerning the idea of habit formation they were more explicit.Palmer and Jespersen were strikingly modern both at the theoretical level and in the exercises and devices they recommended for achieving this end. Sweet argued that language learning must be a "mechanical" process, whereas the term used by Palmer is "habit-forming". A significant statement by Palmer is the following: "Whenever we are distinctly conscious of the words and constructions we are using, we are doing something contrary to nature" (op.cit., p. 336). The "re-shaping" activities suggested by Jespersen are exactly the type of transformation drills utilized in two of our experiments (GUME 3 and 5).

The opinions of the three authors on the role of grammatical explanations are of particular interest here. Sweet, Jespersen and Palmer were unanimous in arguing that example should precede rule and that the pupils should be encouraged to generalize. However, Roddis (op.cit., p. 342) observes that Sweet contradicted his own arguments by claiming that such a principle is impractical since, if the pupil has access to the rule, he will always turn to it first rather than exercise his inductive faculties. The essence of their argument is that example should lead to generalization. According to Sweet the pupils should learn, not directly through rules, but indirectly through examples. In this he anticipated Palmer who held that grammatical knowledge should be unconscious instead of analytic and systematic. All three authors criticize the abuses of grammatical generalizations without prior use of example. A knowl-

edge of the grammar is differentiated from a knowledge of the language as such, and rote-learned paradigms in isolation are condemned as mere rigmaroles.

The similarities between the direct method and the teaching strategies proposed by Sweet, Jespersen and Palmer have been explicated in an article by Darian (1969, p. 545 ff). We shall not prolong the discussion here; suffice it to say that much of modern foreign-language teaching practice, the beginnings of which are often dated to the period between the two world wars (see below), can be traced back to the three mentioned authors.

The direct method, as well as some of the new "reform" methods, reached the United States early in the twentieht century. However, many teachers became disillusioned since the main objective, command of the spoken language, proved unattainable under the prevailing conditions of mass instruction. At that time the majority of American students studied a second language for a period of two years only. It goes without saying that the abilities and ambitions of the average students, not to mention those of the weaker and less motivated ones, did not justify the demands made by the oral use of the foreign language. The need to find solutions to the methodological problems was met by a period of intense experimentation between 1920–1935 (Mackey, op.cit., pp 148–149). It was during this period that the famous Modern Foreign Language Study was performed. The results were summarized in the so-called Coleman report (1929) which started much discussion. The report maintained that the only objective that could be achieved within a short period of learning was the development of reading ability. The effect of the study thus became to spread *the reading method.* In courses where the reading method had been adopted, the study began with a period of oral training. The intention was that the student should be initiated into the sound system by listening to and speaking in simple phrases. The main part of the course was then divided into intensive and extensive reading. One feature of the method was the use of graded texts and readers. This system, although valuable from a pedagogical point of view, gave a false impression of the level of reading achieved (Rivers 1968, p. 24). According to its critics, the reading method produced students who were unable to understand or speak the foreign language beyond the most simple utterances. However, research supporting this critique has not been reported to our knowledge.

These were the conditions which prevailed when America entered into world war II. At that time there was an immediate need within the army to provide officers and men with a working knowledge of various foreign languages. In 1941 the American Council of Learned Societies arranged intensive language programs which were converted, two years later, into the wartime Army Specialized Training Programs, popularly known as the ASTP's. Professional linguists and anthropologists were mobilized to organize the emergency

teaching. Since training was mostly a full-time occupation on the part of the learner, the courses produced substantial results in a relatively short time. *"The Army method"* was supposed to contain the secret of successful foreign language teaching. However, no such method existed. "All that the Army had asked for was results, including a fluent speaking of the language; a variety of methods and techniques were used to achieve these results" (Mackey 1965, p. 149). It should not be doubted, though, that the ASTP's had a common orientation; Hanzeli (1968) has summarized it thus:

> "In these programs, a certain number of basic attitudes or leitmotivs developed quite early: the primacy of speech, language learning as habit formation, de-emphasis of grammar rules, and rejection of translation" (p. 43).

After the war many schools and colleges in America tried to duplicate the techniques favoured in the intensive language courses. But conditions in the schools were not such as to contribute to rapid learning: the classes were too big, the learning time was too limited, and the motivation was often lacking. However, the application of linguistic principles to language teaching as well as the collaboration of specialists in various fields in the production of language teaching materials had come to the attention of the authorities. Carroll (1969) har summarized the actions taken by the authorities during the fifties and sixties to initiate programs of teacher education, materials development, and research in modern-language instruction. The history of foreign-language instruction in the United States for the period 1940–1960 has also been comprehensively treated by Moulton (1961). The development during this period may described as a development towards *the "aural-oral" method*, the term indicating that the main emphasis is on the ability to communicate in the foreign language.

Our brief review of teaching methods up to this point reflects a change in objectives; the more intellectualized rule-learning activities were gradually replaced by methods aiming at acceptable speaking and listening performance. However clear this development may seem in retrospect, we would hypothesize that a number of hardly definable variants of methods existed side by side all the time. Likwise it may be supposed that there was continuously some controversy among the teaching profession about the priority of objectives. The relation between the previously mentioned foreign-language teaching theories (the audio-lingual habit theory and the cognitive code-learning theory) and the methods discussed thus far is not very clear. If a rough, one-dimensional analysis were made, the grammar-translation method would in all likelihood be assigned to the cognitive code-learning theory, whereas the phonetic method, the direct method and "the army method" would be considered representative of the audio-lingual habit theory (the reading method would be more difficult to categorize). However, a more relevant (multi-di-

mensional) analysis, paying regard to language (English, Russian, etc), to aspect of language (phonology, syntax, etc), and to age group (young children, university students, etc) would probably reveal that the relation between theory and method follows no clear pattern. We shall not perform such an analysis here, since some of the methods discussed are little more than historical curiosities to-day.

In the next section the audio-lingual method will receive a separate and more detailed treatment than the methods discussed above. There are three reasons why this method deserves special attention in the present report: (a) it is the dominant foreign-language teaching method to-day, (b) its rationale and procedures are well documented in various sources, (c) it is closely related to the teaching strategies compared in our investigations (with one exception, the explicit method in GUME A). A comment may be appropriate on the relation between the audio-lingual habit *theory* and the audio-lingual *method*. The method is in no sense derived from the theory, but the theory is a kind of summarizing description of existing habit-oriented teaching practices, of which the most widely accepted is the audio-lingual method. We shall presently (p. 34 ff) review some of the critique levelled at the audio-lingual method; this critique will simultaneously provide support for the cognitive *theory*. Although descriptions of cognitive-oriented foreign-language teaching procedures exist (see, for instance, Mueller 1971), no comprehensive treatment of a cognitive *method* has yet appeared, at least to our knowledge.

The audio-lingual method

Nelson Brooks (1960, p. 201) suggested the term audio-lingual as less confusing and more easily pronounced than aural-oral. About a decade later Rivers (1968) gave her view of the status of the audio-lingual method:

> "Interest in the audio-lingual method now extends to every continent. It has been enthusiastically endorsed by some teachers and accepted with reserve by others as has happened with all new approaches to foreign-language teaching. Like all living ideas, it is in a process of evolution, and some of the more controversial of the first proposals are being modified through the experiences of many teachers and students" (p. 36).

Considering the debate in methodological questions that the audio-lingual method has given rise to, one might question the relevance of Rivers' fairly sympathetic picture of the acceptance and evolution of the method. However this may be, we shall now give a somewhat detailed description of the method because of its apparent similarity with some of the teaching approaches compared in the present investigation. Elsewhere Rivers (1964, pp. 12–13) lists a number of sources where the methodological tenets of the audio-lingual meth-

od are put forward, the most well-known among them being Nelson Brooks' *Language and Language Learning* (1960) and Robert Politzer's *Teaching French: An Introduction to Applied Linguistics* (1960).

Rivers notes that "an analysis of these sources shows a remarkable degree of concurrence, indicating that the leaders of the audio-lingual movement have a very clear idea of the objectives, principles, and procedures which they jointly advocate". Here we shall describe the audio-lingual method as it relates to what Moulton (1961) has called "the five slogans of the day" (pp. 86–89):

Language is speech, not writing: In a typical audio-lingual course, the pupil is first trained in understanding and speaking the foreign language; reading and writing come in at a later stage. The exact time for the introduction of graphic material seems to be a matter of opinion. It is stressed that articulation and pronunciation should be as correct as possible. Thus the first or audio-lingual stage is considered to be of the utmost importance for the development of the other two (reading and writing).

A language is a set of habits: At this point a statement by Brooks is informative: "The single paramount fact about language learning is that it concerns, not problem solving, but the formation and performance of habits. The learner who has been made to see only how language works has not learned any language; on the contrary, he has learned something he will have to forget before he can make any progress in that area of language (Brooks 1960, p. 47). The audio-lingual techniques aim at giving the student automatic control of the language by means of pattern practice and structure drills. So-called mimicry-memorization of dialogue material is also intended to serve the purpose of rendering the linguistic behavior habitual and automatic. It is often stressed that language patterns should be learned to the point of "overlearning".

Teach the language and not about the language: This slogan reflects the protest against the grammar-translation method where grammar rules and their exceptions were studied in abundance. It is apparent from statements made by prominent audio-lingualists that although analogy is preferred to analysis, rules in the form of generalizations are accepted. However, the explanations or generalizations always come *after* the structure has been thoroughly drilled. According to Politzer (1961, pp. 5–6) "rules ought to be summaries of behavior". The following quotation from Brooks (1960) illustrates the same view:

> "It would be naïve to propose that in formal education we should not provide our students with useful rules of grammar. But such rules should not be very numerous and should be stated in language that makes the matter clear not only to someone who already knows but also to the learner who does not yet know. In general, they should be given to the student after he has had

sustatined practice in using the structure the rule refers to, and the amount of class time devoted to their consideration should be minimal" (p. 142).

A language is what its native speakers say, not what someone thinks they ought to say: The application of this principle is supposed to take place in the construction of audio-lingual teaching materials, where the examples are chosen from ordinary speech rather than artificially constructed to illustrate certain points of grammar.

Languages are different: The view is held that the major difficulties for the learner are the points where the native and foreign languages differ most radically. By means of contrastive analysis these points are identified, and then the audio-lingual materials are planned so as to give special drilling at these points. In the case of dialogues, native language versions of idiomatic expressions are often given in the text.

Theoretical foundations of the audio-lingual method

The above presentation of the audio-lingual method has been made mainly in terms of techniques and procedures. We shall presently discuss its theoretical foundations, both psychological and linguistic, but first a quotation by Valdman (1970) on the differences between the audio-lingual method and the direct method will be given. We consider his comment of importance since the two methods are often discussed, at least in Swedish debate, without the necessary distincion between them.

"(To state it differently), in the direct method emphasis is placed on the production of sentences that have content, with the acceptance of the calculated risk of pronunciation inaccuracy and grammatical error, while in the audio-lingual approach emphasis is placed on accuracy and well-formedness, with the acceptance of the risk that, in early stages of instruction at least, students will manipulate utterances relatively devoid of content" (p. 309).

The linguistic roots of the audio-lingual method are to be found in the twenties and thirties when structural linguists began to view language as a system or a functioning means of communication. Bloomfield was the dominant linguist in the new movement. Oriented towards the behaviorist school of psychology, he rejected mentalistic interpretations of learning in favor of a mechanistic approach (see, for instance, Chastain, 1969, pp. 98–99). The affinity between audio-lingual procedures and the following statement by Bloomfield is apparent:

"The command of a language is not a matter of knowledge: the speakers are quite unable to describe the habits which make up their language. The com-

mand of a language is a matter of practice, and language learning is over-learning: anything else is of no use" (Bloomfield, 1942, p. 12).

The audio-lingualists thus hold that only procedures which call forth behavior in the learning situation will develop the behaviors desired. Politzer has stated that the behavioristic school was the one that contributed most significantly to the development of modern language teaching in the 1940's (Politzer 1964, p. 149). It is obvious that the audio-lingual method is closely related to Skinnerian behaviorism. For instance, the mimicry-memorization and pattern drills are the practical outgrowth of Skinner's principle of successive approximation. In his famous *Verbal Behavior* (1957), Skinner sets forth his view on language and language learning; in doing so he introduces a unique conceptual apparatus, including such terms as the mand, the tact, the autoclitic, etc, which all stand for various verbal operants. In general, Skinner's book is an attempt to deal with the basic facts of language within a stimulus-response framework.

Bosco and Di Pietro (1970) have attempted to trace the psychological and linguistic framework of some current instructional strategies, among them the audio-lingual method. There seems to be no one-to-one relation between teaching method and psychological or linguistic theory:

> "While it appears that no current instructional strategy is built *exclusively* and *directly* upon a single, well-defined psychological or linguistic system, the conceptual framework of current theoretical systems has served nonetheless as a general point of orientation for instructional practice" (p. 5).

The authors' analyses are summarized in the following grid:

		Strategies		
		GT	DM	AL
Psychological features:				
1.	Functional	–	+	+
2.	Central	+	–	–
3.	Affective	–	+	–
4.	Nomothetic	+	–	+
5.	Idiographic	–	–	–
6.	Molar	–	+	–
7.	Cyclic	–	–	–
8.	Divergent	–	–	+
Linguistic features:				
1.	General	+	–	–
2.	Systematic	–	–	+
3.	Unified	–	–	–

Key:
GT: grammar-translation
DM: direct method
AL: audio-lingual
+: presence of feature
–: absence of feature

The authors find that the audio-lingual method is functional, nomothetic and divergent as far as its psychological features are concerned; linguistically it is termed systematic. That it is functional means that the learner is expected to produce sentences in the foreign language in order to meet specific communication goals. Non-functional strategies attach greater significance to the learner's capacity to understand linguistic structure than to his facility in using the language actively in concrete communication situations. That the audio--lingual method is nomothetic means that priority is given to the shaping of generalized behavior. The authors exemplify by an active-passive transformation (incidentally, a kind of exercise that is used frequently in the present investigation); the presentation of the two sentences simultaneously is considered explicit enough for the pupils to be able to understand the rules underlying the transformation. By explicit the authors obviously do not mean verbalization of the rule. That the audio-lingual method is divergent means that the various skills, phonetic discrimination, listening comprehension, oral expression, reading comprehension, etc, are isolated and treated separately. In non-divergent strategies, an undifferentiated or global view of language is assumed. The audio-lingual method is non-central, i e it does not stress the understanding of "general orientation schemas" but rather the shaping of habits of efficient performance. The method is non-affective, i e the teacher should not concentrate on *intensity* of response but rather on quantitative and repetitive techniques. The method is non-idiographic, i.e. the instruction does not give much room for expressional spontaneity but concentrates on memorization of key sentences or the manipulation of drills. The audio--lingual method is non-molar, i.e. it does not concentrate on gross functional patterns but isolates them into small elements in the effort to achieve precision. Finally on the psychological side, the method is non-cyclic, i.e. the pupil is supposed to "overlearn" any point before moving to the next; in a cyclic approach the pupils become gradually familiarized with it by returning to it at different intervals in the course of instruction.

The scales or continua that Bosco and Di Pietro use for their description of methods range from the more reductionistic psychology represented by behaviorism to the molar orientation found in the gestaltpsychology (op. cit, pp 7–8). It is obvious that the audio-lingual method is more closely linked to the former. As it appears, the authors' analysis is in good agreement with Moulton's "five slogans" presented earlier in this chapter.

On the linguistic side, the audio-lingual method is systematic, which means that matters of language structure are consistently covered according to some organizational scheme. It is non-general, i.e. generalizations about grammatical structures are not made by reference to grammatical rules of a general nature; they are drawn from observations of a language's particular structure. Finally, the audio-lingual method is non-unified, i.e. the learner is not kept

aware of the underlying grammatical model.

Rivers (1964) has discussed a number of theoretical assumptions underlying the audio-lingual method. We shall comment here on the two which have had the greatest influence on our own research:

1. Foreign-language learning is basically a mechanical process of habit formation.
2. Analogy provides a better foundation for foreign-language learning than analysis.

The first point, that language learning is habit formation, has been strongly stressed by Brooks (1960):

> "Pattern practices make no pretense of being communication, but they take the learner through the types of behavior that must be automatic when he does communicate. Pattern practices are to language in action what practice exercises in any skill are to meaningful performance in that skill" (p. 142).

Brook's very strict behaviorist position is that association between the stimulus word (or phrase) and response should continue to the point of automatic performance. Following Skinner (1957), two corollaries of the first assumption are that (a) habits are strengthened by reinforcement, and that (b) foreign-language habits are formed most effectively by giving the right response, not by making mistakes. As Rivers (1964) has pointed out "the audio-lingual techniques seem to meet this situation adequately, as they provide plenty of opportunity for the student to use foreign-language responses in the classroom situation and to receive the reinforcement of acceptance and comprehension" (p. 33).

The second assumption is of particular interest in the present investigation; actually, it is closely related to the main hypothesis of our studies. We have earlier (pp. 29–30) given two statements by Brooks where he recommends a sparse use of rules or generalizations. Politzer (1961) has expressed the same idea:

> "What the student needs is a perception of the analogies involved, of the structural differences, and similarities between sentences" (p. 15).

Palmer's views on the value of grammatical explanations, expressed as early as 1921, are still representative of modern audio-lingual thinking:

> "Nearly all the time spent by the teacher in explaining why such and such a form is used and why a certain sentence is constructed in a certain way is time lost, for such explanations merely appease curiosity; they do not help us to form new habits, they do not develop automatism. Those who have learnt to use the foreign language and who do use it successfully have long since

> forgotten the why and the wherefore; they can no longer quote to you the theory which was supposed to have procured them their command of the language" (Palmer 1921, p. 57).

Thus, according to the audio-lingual view, conscious attention to the critical features of a grammatical structure will interfere with the fluent use of it. This proposition is probably the one that has caused the most servere controversies.

Critique of the audio-lingual method and theory

Skinner's *Verbal Behavior* was severely criticized in a review by Chomsky (1959). Not only Skinner's attempts to extrapolate from bar-pressing behavior of animals to complex linguistic behavior, but also this treatment of linguistic phenomena in ordinary behavioristic terminology is objected to by Chomsky; one example will clarify the devastating character of his critique:

> "It seems that Skinner's claim that all verbal behavior is acquired and maitained in 'strength' through reinforcement is quite empty, because his notion of reinforcement has no clear content, functioning only as a cover term for any factor, detectable or not, related to acqusition or maintenance of verbal behavior. Talk of schedules of reinforcement here is entirely pointless. How are we to decide, for example, according to what schedules covert reinforcement is 'arranged', as in thinking or verbal fantasy, or what the scheduling is of such factors as silence, speech, and appropriate future reactions to communicated information?" (p. 154).

To our knowledge Skinner has never answered Chomsky's critique, at least not in written form. Chomsky's own views on language and language acquisition were first presented in *Syntatic Structures* (1957), where his so-called transformational grammar, a very formalized linguistic theory, was also advanced. According to Chomsky, the most obvious and characteristic property of normal linguistic behavior is that it is stimulus-free and innovative; he has also referred to this property as "the creative aspect of language use" (Chomsky 1965). In learning his native language, the child is functioning as an implicit inductive scientist. He collects data from his environment in the form of linguistic utterances he hears, classifies them into various grammatical categories, and constructs rules in producing new utterances. The system the child develops is not static but subject to revision as new linguistic data become available in the course of development. This "language acquisition device" ("lad") in the child is supposed to be largely innate, a view which Chomsky shares with others (see, for instance Lenneberg, 1964). Chomsky's transformational grammar is divided into two levels, a surface structure level

and a deep structure level. This division has given rise to the hypothesis that imitative-repetitive drills, however systematic, will never go beyond the surface structure, and that an explicit verbalization of underlying structures, resulting in conscious control of transformational mechanisms in the structure under consideration, will result in better learning and greater ease in generating new sentences. Carroll (1966) refers to this as a fact: "In learning a skill, it is often the case that conscious attention to its critical features and understanding of them will facilitate learning" (p. 105).

In Sweden, the opposing theories of Skinner and Chomsky have been analyzed by Ellegård (1968), who hypothesized that a cognitive-oriented method would promote better learning than a method in line with the audio-lingual habit theory.

In his writings Chomsky makes a distinction between the learner's *competence* and *performance*. Whereas, according to Chomsky and other transformational grammarians, association, imitation, and generalization are sufficient to establish performance of specific verbal acts or behaviors, insight into the acts performed is necessary to render competence. Competence is viewed as the learner's ability to use his linguistic knowledge adequately in novel situations, to produce utterances he has never produced before. Jacobsson (1968) has questioned the relevance for language learning of what he calls "the transformational gospel". In his view the concept of competence refers to an ideal, non-existing speaker; when the linguist or psychologist is to draw inferences about a learner's competence, he is always forced to do so by means of data collection, i.e. by observing acts of performance. According to Jacobsson, the conceptual framework of the generativists is perhaps more attractive aesthetically, but it rests on fragile grounds, namely the hardly definable concept of "intrinsic competence" (Jacobsson 1968, p. 371).

Spolsky (1966) discusses competence and performance in terms of "knowing a language" and "language-like behavior". In his article he criticizes programmed foreign-language instruction for having adopted a narrow Skinnerian theory of learning:

> "(A theory of language learning) must go beyond the establishment of a number of language-like behaviors to the establishment of a linguistic competence similar to that of the native speaker. Perhaps this goal is ultimately impossible, but to accept the Skinnerian model is to give up any hope of achieving it" (p. 127).

Saporta (1966), discussing Chomsky's generative grammar and its applications to second language teaching, criticizes the behavioristic tenets as inadequate for explaining language acquisition. His views on the question of learning grammatical rules is particularly relevant here:

> "To say that new sentences are produced by generalization or analogy is of little help unless one can make explicit how a learner selects precisely the correct analogy. The ability to accept *I eat fresh fish* and to reject *I eat fresh well* implies command of an abstract grammatical rule, a rule which distinguishes *I eat fish* from *I eat well* and, incidentally, which makes the distinction without appeal to the accoustic signal. In short, the correct generalization implies knowledge, perhaps unverbalized, that nouns and not adverbs may be modified by adjectives, and that *fish* and not *well* is a noun. No amount of hand waving will obscure the fact that this is what has to be learned, and the appeal to generalization is vacuous since it presupposes knowledge of precisely what it is that is to be learned. On the other hand, having made this point explicit, we are no nearer understanding what the most efficient way is of learning it" (p. 87).

Barrutia (1966) states that in language learning not only sets of responses, but also some form of internal "strategies or plans" have to be learned. Having learned these strategies or plans seems to be synonymous to having achieved "competence", "knowing a language" or having gained "insight". According to Barrutia, this is facilitated by explanation of the grammatical rules. He abandons the position that "no grammatical rule is ever necessary" (p. 163) and compares placement of the explanation before and after the drills. Both procedures are said to involve certain disadvantages from a learning point of view, and Barrutia therefore suggests what he calls "a prudent eclecticism", putting the explanation between the drills. Incidentally, this is the strategy mainly adopted in the present investigation.

The plans and strategies proposed by Barrutia also seem to be related to the "metaplans" of Miller, Galanter, and Pribram (1960). These are "plans to generate plans" of grammatical usage, and when the appropriate moment comes "they can be projected into an infinite variety of unforeseen situations" (p. 178).

Jakobvits (1968), when discussing the implications of psycholinguistic developments for the teaching of a second language, is very explicit on the problem of concern here:

> "Rules that the child discovers are more important and carry greater weight than practice. Concept attainment and hypothesis testning are more likely paradigms in language development than response strenght through rote memory and repetition" (p. 101).
> "The teaching of such (explicit) verbalizations therefore ought to facilitate foreign language acquisition" (p. 105).

It should be noted that, according to Jakobvits, verbalizing a grammatical relation can take two forms; one of the kind that can be found in a grammar book including technical terminology, and one which is a kind of generalization expressed in any convenient way using whatever terms are available to

the individual, whether technically correct or not.

Mowrer (1960), when discussing his so-called revised two-factor theory and the concept of habit, holds that responses, in the sense of overt, behavioral acts are never "learned" and thus not dependent on quantity or reinforcement (p. 386). If Mowrer's theory is correct, there is a risk that intensive drill in the classroom will cause boredom rather than increased learning (cf Rivers 1964, p. 39). Tolman (see Hilgard 1956, chapt. 6) also refuses to accept the idea of reinforcement as strengthening or establishing a habit. To him reinforcing the right response represents confirmation of the hypothesis or expectation of the learner. According to Tolman, it is necessary to give the pupil practice in using foreign-language phrases successfully in a variety of situations, but he also warns against too much reliance on practice as a method of building up habits. Continued practice after a response has been learned tends to fixate a particular response, making it harder for the pupil to vary it on future occasions. This observation by Tolman corresponds to the often-heard criticism of the audio-lingual method that it runs the risk of producing "well-trained parrots" (Rivers 1968, p. 46).

The audio-lingual proposition that habits are most successfully formed by giving the right response, has been criticized by various researchers. Jakobovits (1968) notes that the fluent speech of most native speakers does not consist totally, or even in the majority of cases, of well-formed sentences. He holds that the requirement to utter exclusively well-formed sentences would seriously hinder the fluency of most native speakers. He continues:

> "The logical implication of this observation would be that no language teacher should ever force his pupils to use only well-formed sentences in practice conversation whether it be in the classroom, laboratory or outside. This conclusion is not as odd as it may seem at first sight. After all, children seem to acquire the competence to produce well-formed sentences despite the semi-grammaticality of the adult speech to which they are continually exposed" (p.107).

Cook (1969) has expressed a similar view in her comparison of the conditions of first and second language learning. She observes that a child's errors in connection with learning the native language are usually considered "cute" by the environment whereas, in the case of second language learning, the pupil's mistakes are considered "dangerous". She argues that language learning necessarily passes through hypothesis testing where errors represent incremental rather than decremental learning:

> "If the second language learner is to proceed by a series of makeshift hypotheses, he too must be allowed to err (in terms of native competence) so that he can test his hypotheses and abandon those that are unsuccessful" (pp. 210-211).

Fodor (1966, p. 112) states that "imitation and reinforcement, the two concepts with which American psychologists have traditionally approached problems about language learning, are simply useless here". He makes this strong statement when discussing how a child learns the correct base structure for a certain type of sentence.

Our last reference on the question of language learning as habit formation will be a pronouncement by Chomsky at the Northeast Conference on the Teaching of Foreign Languages, 1965:

> "It seems impossible to me to accept the view that linguistic behaviour is a matter of habit, that it is slowly acquired by reinforcement, association and generalization, or that linguistic concepts can be specified in terms of a space of elementary, physically defined 'criterial attributes'. Language is not a 'habit structure'. Ordinary linguistic behaviour characteristically involves innovation, formation of new sentences and new patterns in accordance with rules of great abstractness and intricacy. This is true both of the speaker, who constructs new utterances appropriate to the occasion, and of the hearer who must analyze and interpret these novel structures. There are no known principles of association or reinforcement, and no known sense of 'generalization' that can begin to account for this characteristic 'creative' aspect of normal language use".

Some comments on the theoretical controversy

One may ask, when facing the contrasting opinions discussed above, if the theories advanced are equally tenable and, if this is not the case, which one seems most promising for generating hypotheses concerning optimal foreign-language learning. In the first place it is interesting to note that a number of authors have voiced scepticism about the relevance of any present theory for predicting proper language teaching procedures.

Chomsky (1965) claims that both psychology and linguistics are in a state of "flux and agitation" and that neither discipline has achieved a level of theoretical understanding that might enable it to support a technology of language teaching. Carroll (1965) considers present theory of foreign language learning to be at a rudimentery stage; in his opinion there exists no proven theory to account for all the phenomena that we can observe or even the phenomena that we can predict or control (p. 278). Anisfeld (1966) states that at the present stage of development of psychology applications to the classroom situation can be accomplished only by a superficial treatment of psychological subject matter and an over-simplified analysis of the nature of the problems involved.

Some recent reports have testified to the "flux and agitation" observed by Chomsky. James (1969) notes that transformational grammar has provided

great insights for applied linguistics, whereas Johnson (1969) describes it as a complete failure as far as language teaching is concerned. Warhaugh (1969), in reviewing "the state of the art" for the Center for Applied Linguistics, states that the theory of foreign-language learning is characterized by "uncertainty".

Crothers & Suppes (1967) discuss the relevance of psychological and linguistic theory for foreign-language learning in connection with their comprehensive study of learning Russian phonemes, words, and sentences. In their opinion, no existing psychological or linguistic theory can account for any substantial portion of the systematic details of language learning. Their comment on the opposition between the behavioristic and cognitive approaches is worthy of note:

> "The thesis that we want to defend about the apparent conflict between behavioristic and cognitive theories is that much of the conflict is apparent rather than real. When the theories are formulated in a mathematically sharp fashion and in terms that suffice to deal with the details of any substantial body of experimentation, then a surprising amount of agreement in formal structure is to be found, in spite of the rather different terminology used" (p. 7).

Similarly, Carroll (1971) argues that the opposition between rule-governed behavior and habits is a false one. The individual's linguistic habits, in so far they conform to the habits of the speech-community of which he is a member, may equally be looked upon as rule-governed behavior.

Considering the strongly opposing opinions in foreign-language theory and practice, it is only natural that a tendency towards eclecticism has been noticed in some authors. Hanzeli (1968) suggests a theory which takes both habits and rules into account. Rivers (1968), in expressing her middle-of-the-road position, states that there must be a constant interplay in the classroom of learning by analogy and by analysis, of inductive and deductive processes. Gagné (1965), though not participating in the present controversy, makes some interesting observations. According to him there is a case in foreign-language learning both for a deductive approach, utilizing rules in a fairly traditional way, and an inductive approach where the student is left to draw inferences on his own (p. 194).

Carroll (1971) suggests what he calls a meaningful synthesis of the two theories - a cognitive habit-formation theory (!). According to this theory, there is a place in foreign-language learning both for presentation of "the facts of the language" and formation of habits.

It makes intuitive sense to believe that each of the theories has unique advantages. It also makes sense to believe that these advantages are differentially related to such things as the objective of language teaching, the age and ability of the learner, and the particular aspect of language to be taught. In

our opinion the method-objectives and method-individuals interactions have been notoriously neglected. For instance, concepts such as linguistic competence or linguistic performance have been discussed without the necessary consideration to whether vocabulary or syntax learning was concerned, whether young children or university students were to be taught and whether the main teaching objective was ability to translate or listening comprehension. To this may be added that the concept of method has often been treated globally and vaguely, which has further obscured the discussion. These inadequacies may have contributed to the impression that existing theories are not sufficiently developed for predicting proper classroom practices. We would argue, however, that if specific variables are selected for study - instead of treating the teaching process in a global perspective – there is a good probability that research will prove parts of each theory to contribute to methodological advancement. On the other hand it is hardly probable that a complete foreign-language teaching methodology can be derived from one single theory.

A single study such as the present one cannot aspire to investigate "learning a foreign language" in all its variety. Any project is necessarily limited with respect to the linguistic phenomena that it treats and the characteristics of the individuals that it is directed towards. However, the accumulated evidence from such studies, provided their independent and dependent variables as well as their experimental samples have been adequately defined, will hopefully increase our knowledge about foreign-language learning in a more general sense.

CHAPTER 3

THE SWEDISH CURRICULUM AND SWEDISH DEBATE

Curriculum

Since the curriculum has been interpreted differently by different linguists and teachers, a brief discussion of its recommendations with respect to the learning of grammar may be in order.

The official curriculum for Swedish schools at the compulsory level (Lgr, Läroplan för grundskolan) sets down both goals and recommended methods for the teaching of English and the second foreign language (French or German). Until the autumn term of 1970 the curriculum of 1962 (Lgr 62) was still in force. Since then, however, it has been replaced by the 1969 version (Lgr 69) with its Supplement in English (abbreviated Lgr 69: II Eng).

In Lgr 62 is stated (pp 197-198) that grammatical knowledge is a means to understand and use the foreign language and not an end in itself. The pupils should not be burdened by unnecessary (sic) analyses and rules but learn the grammatical structures by systematic drills of different kinds. The teaching of grammar should be limited to frequent and important structures. It is emphasized that insight into grammatical patterns is essential both in order to prevent misunderstanding of spoken language and texts read and in order to express oneself in the foreign language. The study of grammar should be cyclical, i.e. a certain structure should be commented on repeatedly and in greater depth only after the pupils have become acquainted with it. The teacher is recommended to introduce new grammatical structures with great care; several unknown structures should not be presented during the same lesson; a new structure, when introduced, should always be imbedded in well-known vocabulary. It is, according to Lgr 62, advisable to use the Swedish language when grammar is being discussed, if no real clarification can otherwise be attained. Before the explanation or rule is formulated, the pupils should have heard several examples of the pattern in question.

The teacher is also advised to make the pupils formulate the rule on their own; this kind of inductive teaching is supposed to train the pupils' power of observation as far as linguistic phenomena are concerned. It is stressed that the grammatical structures should be exercised in the foreign language. However, oral translation from Swedish into the target language is not excluded when practising grammatical points.

Elsewhere (p. 194) it is stated that the teaching should be conducted in the foreign language as much as possible. Listening and speaking skills are said to be of especially great importance in the case of English, the first foreign language for Swedish pupils.

Lgr 69:11 does not mention translation into the foreign language as a means of promoting a functional control of grammar. The insight which the pupils acquire about the structure of the target language is said to be arrived at first and foremost by systematic drilling. Generalizations should come in late and preferably be formulated by the pupils themselves which proves "that the pupils have reached insight *through* the exercise" (p.14). There should be at least ten (sic) examples of the pattern in question on each instructional occasion. Overlearning is considered necessary for a lasting command of the language. If the Swedish language is used for observations on grammar, which, according to Lgr 69 11:Eng, is permissible in rare cases, no comparisons with Swedish usage should be made. The following statement is made on the use of rules: "Every grammatical rule *must* (italics ours) be formulated with English as the starting-point." The writer of the recommendations also contends that if *some-any* are translated this will give rise to a mixing of them which might be avoided it they were practised separately, which in turn will make confusion impossible since the two words, in a given context, exclude each other.

Some comments of a more general nature are made (p. 4): By using pictures or objects, the teacher can make his language teaching more concrete; in this way, it is maintained, verbal explanations become more or less superfluous. It is also stated that there is a dependence between Swedish and English which has an inhibitory effect on the learner. The teacher is advised to free the pupils from this dependence, which is best done by letting them listen to and speak the foreign language as much as possible.

The two curriculum versions, Lgr 62 and Lgr 69, obviously have the same main objective: training and development of the practical, or communication, skills. In our opinion Lgr 62 may be looked upon as a proponent of an eclectic method which might perhaps be placed "slightly to the left of the middle" on a contiuum reaching from habit-formation to cognitive code-learning. It should also be apparent that Lgr 69 has a more clear orientation towards the mechanistic schoolof language acquisition and should, accordingly, be placed further to the left on the same continuum. However, in our view both versions give the teacher a fairly free choice of method, given that the main objectives are not obviated.

Debate

In several Swedish daily newspapers and scientific journals there has been an intense, and at times rather aggressive, debate on foreign language teaching

matters during the last few years. Although the most persistent theme during the debate was the merits and deficiencies of the language teaching method recommended in the authorized curriculum, a number of different topics have been covered: the (alleged) low standing of the pupils in second languages at the comprehensive school, the gymnasium, and the university levels, the question of mono- or bilingual glossaries, the effect of various frame factors on language teaching (size of classes, undifferentiated classes, lack of teaching materials, etc.), the question of translation or no translation, the university reform and its consequences for the training of foreign language teachers, etc. Most of the debate evidence has been collected in two books, one by Ellegård & Lindell (1970), and one by Edwardsson (1970), the latter containing continuous comments by the author on the various contributions.

The most recent debate or, rather , series of debates, lasting from 1968 and onwards, had its predecessors. Actually, a debate in *Pedagogisk Debatt* in 1959 may looked upon as an expression of new trends in language teaching methodology, trends which have been questioned by one side in the recent debate. In the following ten-year interval there appeared debates as well as single contributions showing a great similarity with the most recent debate; the perhaps most noteworthy contribution is Holmberg's article »Educators or Drill-Sergeants?" in *Moderna Språk* (1965). Most of the debate evidence during this period was listed in the bibliography of one of the earlier GUME reports (Lindblad 1969).

Alvar Ellegård, one of the sponsors of the GUME project, started the latest debate by proposing a re-thinking in teaching methods considering new findings in linguistics and psycholinguistics and in comparative research. According to Ellegård, a method promoting insight by conscious attention to the structural features of the language would be superior to the direct method suggested by the curriculum. Although there were opinions pro and con, most of the teachers participating in the debate sided with Ellegård; in fact there was a vast debate majority in favor of a teaching method fostering "insight". A dramatic demostration of this opinion was an address signed by 2001 language teachers at the gymnasium level and handed over to the Minister of Education. In it the teachers stated that the results of the foreign language instruction had deteriorated rapidly during the last years. They blamed the situation on the monolingual instruction recommended in the curriculum for the gymnasium (Lgy 1965). In ten points they made it clear what changes they wanted in future. They desired the prescribed methodology to include features from the traditional method as well as from methods created more recently. Not only should teachers in foreign languages but also those in Swedish endeavour to give the pupils grammatical insight appropriate to the different age-groups. The grammarbook should partly build on contrastive analysis and the rules should be in Swedish. The oral instruction should be sufficiently

backed up by written exercises, and translation from and into the foreign language should be used as an instructional means whenever it was considered to be to the purpose.

It should be noted that some of those who came out in defense of Ellegård had obviously misunderstood him, interpreting him as advocating a traditional grammar-translation method. Representatives of the Board of Education, in defending "the official" method, accused Ellegård of misinterpreting the curriculum. In their view, the curriculum recommends a modified direct method which does not preclude the use of the mother tongue, nor is it loose in its formulations concerning the need of solid knowledge of grammar and vocabulary. Ellegård replied that, however wrong his own interpretation of the curriculum might be, the language teaching profession at large had the impression that the recommendations are strongly biassed towards a pure direct method, and tried to teach accordingly. Some debaters put the blame on officials at the Board of Education, accusing them of advocating, at teachers' meetings and the like, a method without support in the curriculum.

One argument on the "official side" was that Ellegård had disregarded the main objectives of foreign-language teaching at the comprehensive school level (ability to comprehend spoken language and to speak it without inconvenience), objectives which were said to be attainable only by an essentially monolingual method.

During the debate reference was often made to research results, particularly those of the Pennsylvania project (the first year results) and the Swedish UMT investigations (see the following chapter). The two sides of the debate apparently had sharply diverging views on these results and their implications for foreign-language teaching. Apart from this it may be stated that the debate was often based on little, if any, empirical evidence. "Teacher experience", "traditional pedagogy", etc., were the authorities quoted as support for one opinion or the other.

Despite the fact that the methods discussed were seldom very strictly defined by the debaters, it seems safe to conclude that none advocated an extreme direct method or, with perphaps one exception, an extreme grammar-translation method. In fact, the general tendency during the debate has been described by one observer as "a struggle towards the middle".

As Ellegård & Lindell (op.cit., p. 182) have pointed out, representativity of opinion is hardly obtained by means of a free debate. However, the general tendency during the debate, i.e. the preference of most participants for a method emphasizing grammatical insight and explanations in the mother tongue, was confirmed in an inquiry performed by the UMT project (Hall 1969).

Towards the end of the debate the head of the Board of Education officially stated that the curriculum had largely been interpreted too strictly and nar-

rowly; the teachers were free to choose method within the general frame of language teaching objectives. It was also regretted that information from the Board of Education to the teaching profession had perhaps not been exemplary. After this official pronouncement the discussion abated.

It is impossible to predict what effect the debate will have in future. Many participants urged that comparative research be initiated in order to investigate the prevailing method and its theoretical foundations. Research of this kind is under way, witness the present investigation. Apart from this it may be supposed that the debate has had a generally wholesome influence, fostering a more balanced view on foreign-langauge teaching methodology.

CHAPTER 4

EARLIER RESEARCH ON THE EFFECTIVENESS OF FOREIGN-LANGUAGE TEACHING

Carroll (1967), in reviewing comparative research within the field of foreign-language teaching up to around 1960, is rather pessimistic about the scientific merits of it:

> "Rigorous experimental design has been largely absent from such studies; instead, simple group comparisons have been made at various stages of training, with hardly any use of control measurements. The results of these studies have been largely inconclusive" (p. 1066).

The large-scale investigation by Agard and Dunkel (1948) is, according to Carroll, a case in point. Their study was a broad survey where the results from a variety of high schools and colleges using either traditional or "new-type" methods, or both, were forwarded to a central office for statistical analysis. The authors reported that (a) few students in the aural-oral programs were able to attain "spontaneously" fluent speech in one or two years time" (p. 288) and that (b) the experimental groups had consistently superior pronunciation compared to conventional groups but lagged in reading proficiency (pp 287-288 and 289 respectively). Carroll's critique of the study implies that few penetrating measurements were made, that exact controls were lacking and that the tests used were not as reliable and valid as might be desired (Carroll, op. cit., p. 1067). However, elsewhere Carroll (1969 a) has noted that the study aroused an interest in comparative research within the field of foreign-language learning:

> "Ever since the Chicago Investigation of Second-Language Teaching (Agard and Dunkel) there have been studies that attempt to show what kinds of student achievements can be expected from the audio-lingual as compared with the grammar-translation approach" (p. 869).

Smith and Berger (1968), reviewing related research in connection with their own comparative study, end their survey by stating:

> "By 1964 no sufficiently realistic and generalizable research had been undertaken to shed light on specific questions of modern foreign language instruction facing the American secondary school: which strategy (or laboratory system) works best when translated from a specific local small scale setting into the larger reality of numerous secondary schools?" (p. 10)

A similar view of the general value of comparative studies in foreign-language teaching before 1960 was voiced by Scherer and Wertheimer (1964) at the time when they were planning their own comprehensive investigation:

> "A rigidly controlled large-scale experiment which would yield clear-cut data was therefore still needed. If we could find ways to measure all the separate skills of language proficiency – and perhaps to assess other psycholinguistic characteristics of a speaker's ability in a foreign language – we should be in a position to draw some definite scientific conclusions about the relative merits of the two methods. More important, it would become possible to specify in just what ways, if any, an audio-lingual approach is superior to a traditional one" (p. 12).

Accepting the above quotations as valid judgments of the status of earlier research, we shall limit our own review to studies from the mid-sixties and onwards. As a further limitation, we shall choose only references where theoretical problems considered in the previous chapter are dealt with.

Scherer & Wertheimer (op.cit.) performed a two-year experiment at the University of Colorado, comparing an audio-lingual and a traditional method; the subjects were students in a college German program. Although the project staff had planned to use a matched-pairs design, this strategy was abandoned for various reasons and the subjects were assigned by simple randomization to the two teaching methods. The experimental (A-L) group received an initial period of purely audio-lingual training before it was given any training whatsoever in reading or writing, whereas the control group was given reading material from the start; in the latter, "traditional" group, grammatical analysis was frequently used and grammatical terminology taught. The authors state that uniformity of teaching procedures was obtained through weekly meetings and conferences with individual teachers and by visiting consultants who checked that the experimental teachers adhered to their respective teaching strategies. The authors used a variety of criterion measures: a six-skill battery including listening comprehension, speaking, reading, writing, German-to-English translation and English-to-German translation. In addition, they administered tests and questionnaires which measured various psycholinguistic and motivational factors.

At the end of the first year of instruction, the audio-lingual group was found to be far superior in listening and speaking skills, whereas the traditional group surpassed in reading and writing. For administrative reasons the two groups had to be merged in a common group of instruction during the second year of instruction. At the end of the second year the differences between the two groups had largely disappeared in the case of the passive skills, i.e. listening and reading, but there were still differences in the active skills: on the average, the audio-lingual group had better speaking fluency and the traditional group did better in writing. On the whole, however, the differences

between the groups were so small as to warrant the conclusion that it does not make much difference whether the audio-lingual or the traditional method is used. It appears, though, that the audio-lingual method produced more desirable attitudes towards speaking the foreign language.

The Scherer & Wertheimer study has been extensively reviewed and discussed. Critique of it has mostly been concerned with the fact that few students completed the two-year study (N = 49), that the two groups were not kept separate during the second year and that the experimental teachers were not strictly enough instructed to follow a certain teaching pattern. Caroll (1965) has pointed out that no precise formulation of the relevant theories underlying the two teaching methods was made; Scherer and Wertheimer were merely concerned with the general comparison of two widely used methods of teaching as they understood them. However, as we have stressed earlier, the strength of the study, at least in comparison with earlier ones, lies in the rigorous controls and the creative use of various criterion measures. The main conclusion that may be drawn from the study is perhaps that what is learned is exactly what has been emphasized in the instruction; no mysterious transfer takes place between the various skills.

Politzer (1967) investigated the effect of presence versus absence of explanations and, in cases where they were given, their proper order of presentation in relation to the grammar drills. More precisely, he dealt with the much debated question whether the explanation (a) should precede the drill (b) should be given after some material has been introduced (c) should be given at the end of the drill as a so-called generalization, or should be given at all. The hypotheses were tested with six French and six Spanish drills, the drills being administered to four French and four Spanish classes. Drills were recorded on tape in such a way that the differences in the use of the explanation constituted the only differences between the four treatments. Treatments were rotated among the classes. Written tests including transformation items constituted the criterion measures and dependent variable of the experiment. Aptitude measures (Modern Language Aptitude Test) were used as covariates in order to establish the relation of aptitude with treatments and in order to adjust criterion measures for aptitude differences by analysis of covariance. The results of both the French and the Spanish study indicated that differences between school classes were more important than treatments. Statistically significant differences between treatments were obtained only in two out of twelve experiments; in those experiments treatment (a) and (b) proved superior. Politzer concludes that the independent variable under investigation – position of or absence of explanation – does perhaps not have the importance attributed to it in some of the current pedagogical discussion.

Wohl's (1967) study is a small-scale experiment comparing two methods of teaching English as a second language. The subjects were girls in one school

class in the first year of the secondary division of a small private school. The experiment was a matched group design with both groups taught by the experimenter. The independent variable was presence/absence of analysis of the grammatical pattern. The experiment lasted three months, during which time there were five English lessons a week; materials were specially prepared by the experimenter. There were two pre-tests and four post-tests. No statistically significant differences were obtained between the groups on the criterion test. The author comments that there were no adverse effects suffered by the experimental group in their having learned some grammatical abstractions.

McKinnon (1965) performed an experiment in which he taught third-graders various sentence patterns of the Motu (New Guinea) language. Three teaching method variables were compared for effectiveness. In method one, the children practised imitating recorded sentences (prompting). In methods two and three visual referents were provided. Method two children also imitaded (prompting), but method three children composed the utterance appropriate for each visual situation before hearing the model sentence recorded (confirmation). The design used also made possible comparison of two procedure variables, one inductive and one deductive. For the inductive procedure no instructions were given, for the deductive procedure simple directions pointing out the features of the pattern were given. The results showed method three to be superior, i.e. the method in which active practice in sentence construction was aided both by pictures representing the meaning of sentences and by grammatical explanations that allowed conscious application of rules.

Lim (1968), in a study more or less modelled on McKinnon's, investigated the same type of variables. In the experiment third-graders were taught four sentence patterns in Malay, practising individually with a Language Master during the two-week duration of the experiment with each class. The most clear-cut results were obtained in the case of prompting versus confirmation, the latter proving to be the superior method. The confirmation method differed from the prompting method in that the pupils were more active: they produced their own version of a Malay structure before the Language Master pronounced it; in the prompting method the pupils just repeated the structures produced by the Language Master. The deduction-induction variable produced no main effects on any of the criterion measures used in the study. It should be noted that both teaching procedures utilized explanations of the syntactic features to be learnt; in the deductive method they were given at the beginning of the practice, in the inductive method they were given halfway through the practice session. The author concludes that in the usual classroom situation it seems to make little difference at what point of practice the grammatical rule is given.

Casey (1968) performed an ex post facto investigation of two methods of

teaching English as a foreign language in some Finnish secondary schools. A number of teachers were identified by Casey's method profile mentioned in a previous chapter (pp. 20–21 above); pupils who had been taught by teachers with positive scores on the profile made up the audio-lingual group, whereas pupils taught by teachers with extreme negative scores constituted the cognitive code-learning group. The pupils in the two groups were matched on a number of variables. All pairs of pupils were significantly different in only one respect, namely the methods index. The investigator constructed a series of tests including both the aural-oral and written aspects of the language; the battery supposedly did not favour any of the methods. On the oral test battery, pupils in the experimental group, who had studied under teachers using mainly an oral approach, recognized more phonemic distinctions, had better pronunciation on selected phonemes, constructed more complicated oral dialogue than the control group, but in no case was the difference statistically significant. In the tests of written two-way translation, the pupils in the control group performed better; however, the difference was not significant. There is thus a striking similarity between these results and those of Scherer & Wertheimer: though the method differences are generally small, there is a tendency for better learning of the elements that have been emphasized in the instruction.

Chastain (1968) reported a study undertaken to investigate the relative effectiveness of the audio-lingual approach and the cognitive approach in teaching introductory Spanish classes at the college level. Although the experiment proper lasted one year, the author has given an account of the standing of the two groups after two years, the subjects being mixed during the second year with students who were not part of the study (Chastain 1970). The students were randomly assigned to one or the other of the methods; various checks showed that the two groups were equal in all relevant background variables. The two instructors engaged in the experiment switched hours at the beginning of the second semester in order that as many students as possible might have both instructors during the course of the first year. The classes met four times a week in the classroom for fifty minutes each. It is apparent from Chastain's description of teaching materials and procedures in the two groups that they were treated according to the audio-lingual and cognitive tenets respectively. It should be noted, though, that the cognitive code-learning method as practised in the study was not the traditional method. There was a minimum of translation, and there was a great deal of oral work in class. The students of both methods were assigned tasks as homework, a dubious feature in experiments of this kind. Four criterion tests covering listening comprehension, speaking, reading, and writing, were administered at the end of the first year. The analyses and results indicate that significant and consistent differences were found in reading where the

results favoured the cognitive group. The audio-lingual group was significantly superior in one aspect of speaking, imitative ability. No-significant differences in listening comprehension and writing favoured the cognitive group. Somewhat astonishingly, the author interprets the results as clearly favouring the cognitive code-learning theory. In our view, the statistical evidence rather points in the same direction as that of the earlier studies presented in this chapter.

As was mentioned above, most of the experimental subjects continued to read Spanish a second year. They were intermingled with other students taking the same course, and they were not treated or in any way controlled by the investigator. However, at the end of the second year they were given the indentical criterion tests as were administered at the end of the first year. The intention was to find out if the differences from the first year still prevailed; in other words, did the treatments have any lasting effect? There were no significant differences between the audio-lingual and cognitive students at the end of the second year. Chastain concludes that neither method is uniformly better for all students in all language skills and speculates on a synthesis where the best of both methods be combined.

Torrey (1969) compared three methods of teaching grammatical patterns contained in simple Russian sentences. She constructed what she called two microlanguages intended to illustrate two different abstract linguistic categories. Although her study was of the laboratory variety, she "purposely retained several characteristics of real language lessons in order to provide as much realism as is compatible with an experimental method" (p. 360). The methods compared were a drill method designed to induce learning of generalized patterns, a "grammar" method providing explicit, intellectual knowledge of the grammar rules, and a third method combining the features of the two previous methods. Her criterion tests included free recall, English-to-Russian translation, a Cloze test, and a test of memory span.

In the case of free recall no differences were obtained between the methods. However, on all the other tests the drill method proved statistically superior to the grammar method; the combined methods group performed better than the grammar group but not as well as the drill group. Torrey's experiment seems to support pattern practice but, as Carroll (1966) has pointed out, it probably demonstrates that active practice in sentence construction is better than no practice of any kind.

Mueller (1971) describes a study of five different two-semester courses in French taught at the University of Kentucky from 1966 to 1969. Three of the courses implemented the audio-lingual theory and two of them employed cognitive code-learning principles. The student body in each of these courses was essentially similar as measured by the Carroll-Sapon Modern Language Aptitude Test. The MLA Cooperative Tests of listening, reading and writing were administered at the end of the various courses. The results indicate that

the cognitive code-learning courses were significantly above the national norms with respect to listening and writing, whereas the audio-lingual courses were below or near the norms. The audio-lingual courses had a relatively high rate of attrition; the author interprets this loss of enrollment as diffidence or dissatisfaction with the audio-lingual procedures. The results, favouring a cognitive approach, should be cautiously interpreted since they refer to relatively poorly controlled survey studies and not experiments in a stricter sense.

In Sweden some comparative studies within the field of second-language learning have been carried out during the last few years. We shall comment on those having a bearing on our own investigation.

Werdelin (1968) compared the value of external direction and individual discovery in a language learning situation; although the study was concerned with vocabulary learning, which has not been our concern, the methods compared have a resemblance to those in our experiment. One group was told the principles of the Arabic alphabet and applied it on examples; a second group was given most examples first, then told the principles, and given additional exmples; a third group was given examples only. The three groups had been selected at random from seven eight grade classes and matched with respect to scholastic achievement, line of study, and sex. The performance of the groups was measured by two tests, one of which was a transfer test; the tests were administered immediately after the experiment and after two weeks in order to measure retention. The results were not in favor of any particular method of instruction. There was a tendency for the students who were told the principles ("the cognitive" group) to be somewhat superior to the other groups in learning the foreign alphabet. On the other hand, the students who had to discover the principles from examples (the "drill" or "inductive" group) proved significantly superior in situations involving transfer and retention over two weeks. The author comments:

> "The aim of educational research must be to look for a general law or rule, but we are still far from it. From what we can find from this experiment, there is not much difference between the methods applied to this material" (p. 251).

Sjöberg & Tropé (1969) performed an experiment in a similar vein. However, in their case the inductive-deductive contrast concerned the learning of a grammatical rule, a problem of greater relevance to our study. The grammatical problem investigated was the use in English of the ing-form after preposition where simple infinitive is used in Swedish. Two methods of instruction were compared: in one group the pupils were told the grammatical principle and allowed to practise it on a number of examples, in the other the pupils were given the practices only. Two examiners gave the instructions alternately in the two groups. The experimental groups consisted of forty-five pairs of

sixth-grade pupils matched according to sex, line of study, and intelligence, The criterion tests included one test of positive and one test of negative transfer, the latter indicating that the pupils adopted the grammatical rule even where it was not applicable. On four of five criterion sub-tests as well as on the positive transfer test the rule group excelled the non-rule group significantly. However, since this was the case also on the negative transfer test, suggesting that the rule group applied the grammatical rule mechanically without really understanding its significance, it is difficult to interpret the results in a meaningful way. Parts of the test battery were administered five weeks later in order to measure retention. At that time all the previous differences were levelled out; in fact, there was a tendency for the rule group to have forgotten comparatively more than the group which had been taught with the aid of examples only. The authors conclude that no certain method of teaching the problem in question can be recommended on the basis of the experiment.

Lindell (1971) has summarized the research activities of the UMT project (Undervisnings/Metodik/Tyska – Teaching/Methodology/German) in Malmö, Sweden. The project has dealt with German in the seventh form, i.e. the first year of the second foreign language. Of particular interest here are the comparisons of different teaching procedures. According to the author (p. 65), the experiments were not planned to test the validity of different grammatical models, but the attention was originally directed towards the effects of various types of language laboratories; however, problems concerning the role of grammatical explanations were investigated. In one experiment a comparison was made between a group which was given grammatical explanations and a group which was presented the materials in a structured manner but without any explanations. The grammatical point taught was the present tense of *sein* which has no counterpart in Swedish. The results were clearly in favor of the group provided with explanations. Although no method comparisons proper were performed in cases where the structural differences between German and Swedish are smaller (accusative or personal pronouns and accusative of nous), the results of various studies in the language laboratory indicate that grammatical analysis or explanations are not necessary for significant learning to occur. The author ends his account of the studies with a clarifying discussion of the relation between teaching method on one hand and the structural differences between the first and second language on the other.

In another experiment Lindell (op.cit., p. 45 ff) investigated the proposition, mentioned in connection with the presentation of the audio-lingual method (see p. 29), that it is beneficial not to introduce any graphic material during the early stages of foreign-language learning. In the experiment ten school classes were taught introductory German; an illustrated reader was used during the experiment which lasted ten lessons. Five of the school

classes used the intact reader, whereas the remaining five had copies with the pictorial material but without text. The independent variable of the experiment was thus presence or absence of text. The various drills and exercises were not dependent on the text. A large number of criterion measures were used. Only in a test of written production was a significant difference between the methods obtained; the pupils taught with text excelled. However, in all the remaining variables except one there was a tendency for the text group to surpass the group without text; when the sub-test results were added to a total, the overall difference between the two groups was strongly significant.

Lövgren (1966) compared the effects of mono-lingual and bilingual word-lists in learning German. Although vocabulary learning is of limited interest in the present investigation, the source language/target language problem is of major concern. Six school classes at the gymnasium level took part in the experiment. Two reading texts were constructed and both were provided with two word-lists, one mono-lingual and one bilingual. No pretests were used, but the two experimental groups, each including three school classes, were rotated among treatments. This investigation, which has been much discussed and in some quarters strongly criticized, clearly indicates that the bilingual word-lists produced better learning.

We shall conclude our review of related research by discussing, at some length, the largest undertaking in recent years in the field of foreign-language learning, namely the so-called Pennsylvania project (Smith &Berger 1968, Smith & Baranyi 1968, Smith 1969 a). When planning the present investigation we had access to the first Pennsylvania report; we have tried to take its techniques and general procedures into consideration, which, of course, has not been possible, or even desirable, in all cases.

Three teaching strategies were being compared:

TLM Traditional Method

FSM Functional Skills Method

FSG Functional Skills +Grammar

The intact school class was the experimental unit. Class assignment was random only across the two skills methods. In the case of TLM, only teachers who had expressed a preference for that strategy were assigned to it. The assignment procedure is thus a potential source of error since it is possible that teacher preferences reflect belief in the strategy, which may breed more enthusiasm for the work and hence encourage better results.

The objectives and characteristics of the three teaching strategies were defined by a select panel of modern language educators, among them Robert

Lado, Stanley Sapon and Albert Valdman. The traditional method is obviously very traditional, at least according to Swedish standards, which is demonstrated by part of the description of TLM:

> "Use of native tongue in the classroom predominant. Target language not to be used for purposes of communicating instructions or information to students .....
> Grammar:
> 1. Analysis before application.
> 2. Language organized into word lists, paradigms, principal parts, rules.
> 3. Analysis in depth of grammatical structures .....
> General orientation of traditional program is academic and intellectual" (Smith & Berger 1968, p. 19.).

FSM is of the audio-lingual variety:

> "The functional skills are taught by means of the dialogue and its associated activities. There is opportunity for extensive student practice in both listening and speaking in the target language. Vocabulary is learned only in context while formal prescribed grammatical analysis is avoided" (op. cit., p. 21).

In the authors "list of criteria" describing FSG it is difficult to detect what distinguishes it from FSM. The only difference we have found which might provide sufficient stimuli for the teacher to behave differently is the following:

> "Pattern drills are supplemented by explicit instruction in the appropriate grammar" (op. cit., p. 23).

Considering this diffuse difference between FSM and FSG one might venture to say that the experiment is in reality a comparison between one very traditional and *one* audio-lingual teaching method.

Analogously three laboratory systems were defined by the above-mentioned panel. However, since our main concern here is with the comparisons between teaching strategies, we shall not include the language laboratory part of the study in our review.

Both German and French classes were included in the study but only beginners in the respective language were concerned. Pupils in grade 8, 9, 10, and 11 made up the experimental population, which enabled an investigation of the optimal age to start second-language learning (within the age limits given). The experiment was planned as a four year follow up. The pupils were, compared to Swedish circumstances, a very select group; only 17-20 % take a foreign language in Pennsylvania. The high selection of the Pennsylvania group is also apparent from the IQ's: 113,5 for the French and 115,1 for the German group. The original (=first year) population consisted of 104 school classes (61 French, 43 German) from nearly as many schools, repre-

senting a wide geographical variation within the state of Pennsylvania. The teachers were all willing to participate in the experiment. Each one had at his disposal a detailed instruction covering his teaching strategy and/or laboratory treatment; the teachers also attended periodic workshops. A most important control of the teacher variable was exercised by so called field consultants who were expected to visit each project classroom about twice a month, discuss the teachers' experiences and advise teachers and administrators of forthcoming project activities.

> "Teachers deviating *markedly* (italics ours) from the assigned strategy/laboratory system were dropped from that assignment and from the project" (op.cit., p. 30).

No special course material was constructed but the teachers were free to choose one out of five (French) or one out of four (German) textbooks. A minimun pensum to be covered per time unit was established (if a class did not manage to cover this pensum it was cancelled from the statistical computations). On the other hand no maximun pensum was established; thus different classes could (and did!) cover different amounts of text.

The pupils were tested extensively three times a year by a number of criterion tests covering various aspects of linguistic performance and attitudes towards the teaching procedures. The teachers received a proficiency test for teachers, and their attitudes towards the teaching strategies were assessed both before and after assignment by means of semantic differential opinion scales.

The design applied was Campell & Stanley's experimental design no. 10, the nonequivalent control group design. The statistical techniques used for comparing treatment effects were mainly analysis of covariance and variance.

The results at the end of *the first year* showed that (a) "traditional" students exceeded or equalled "functional skills" students on all measures, (b) student attitudes were independent of the strategy employed, (c) "functional skills" classes proceeded more slowly than "traditional" classes, (d) there was no relation between teacher scores on the proficiency test and the achievement of their classes in foreign language skills.

Of the original 104 classes, 62 remained in the study throughout *the second year* of instruction. The major conclusion after the second year were that (a) no significant differences existed among strategies on all skills except reading (TLM>), (b) student opinion of foreign language study inclined to the negative throughout instruction, independent of the teaching strategy employed, (c) within the functional skills strategies, the specific teaching materials used proved to be a decisive factor in producing learning effects, (d) neither teacher experience in years or graduate education nor scores on the proficiency test were related to mean class achievement.

During the third and fourth years attrition was considerable; in the case of

the French group, too few pupils remained in the Traditional treatment for meaningful comparisons to be made with Functional Skills classes. Because of the high attrition strict controls were precluded; the third and fourth years should be looked upon as a follow-up of the experimental instruction rather than as a controlled study (Smith 1969a, p. 22). Nevertheless, the following conclusions were drawn from *the third year* results in the German group:

> "Traditional' students equalled or significantly exceeded the achievement of 'Functional Skills' students on the MLA Cooperative Classroom Listening and Reading Tests" (p. 41).

Complete data extending over the full four-year period were obtained on 92 students, 72 German and 20 French, i.e. 2 % of the original experimental group. The German students were quite evenly distributed among the three strategies: TLM: 27, FSM: 24, FSG: 21. This sample permitted the computation of an analysis of covariance using the pre-experimental Modern Language Aptitude Test as a covariate. No significant differences were found to exist at the end of *the fourth year* between the three teaching strategies in the two criterion variables, a listening and a reading test.

Returning to the first year results, we want to stress that the superiority of TLM was largest and statistically significant on the MLA Cooperative Tests (reading, vocabulary, grammar, total). What is noticeable about this test battery is that it consisted of an outdated version (1939–41) that had been reprinted for the purposes of the study. The description of the tests makes it clear that they have an academic orientation that obviously puts TLM at an advantage. During the second year of instruction the 1939–41 versions were replaced by modern variants, and the differences between TLM and FSM/FSG vanished. Considering the type of measuring instruments used in the study the results become almost self-evident and suggest that, in spite of all "lists of criteria", the instructional objectives had not been defined concretely enough, nor had test items been constructed which corresponded to defined objectives. The use of the 1939–41 version of the Cooperative Tests was at best intended to give the Traditional method "a fair chance".

As was pointed out earlier the teachers could choose between four or five text-books or materials. Although it was argued that the situation approximated the real school setting where a large amount of material is available, this is extremely unsatisfactory from an experimental point of view. (A check showed that within the school districts involved in the study, twenty-seven different sets of texts and instructional materials were utilized). Furthermore there were no restrictions on how much text could be covered per time unit. The text materials chosen as well as rate of progress are thus possible sources of variation in the Pennsylvania study. During the first year, progress in the Traditional classes was almost three times (!) as great as in the Functional

Skills classes. Above that the TLM text material was found to contain a large vocabulary. Valette (1969) has shown that even the more modern variants of the MLA Cooperative Tests demand a considerable range of vocabulary; thus it is not surprising that TLM should surpass the Functional Skills methods.

One possible explanation of the considerably faster rate of progress in TLM could be the fact that those classes only had teachers who sympathized with the method.

Considering the above limitations in control and design, we feel that the first year results should be interpreted with great caution. At the time the findings were first presented, they were taken, in some quarters, as clear evidence of the inferiority of modern audio-lingual teaching procedures. Noticing that practically all differences between the three teaching strategies had vanished after two years of instruction, noticing further the authors' own great cuation as regards the third and fourth year results, it becomes evident that the Pennsylvania project has not provided a clear answer to the question of which foreign-language teching method is superior.

The Pennsylvania project has become extensively reviewed and debated. The reader is referred to the December issuse of *Foreign Language Annals*, 1969, and the October issue of the *Modern Language Journal*, 1969, for detailed accounts of various aspects of the study. Here we shall briefly comment on some of the criticisms expressed.

In the first place, it is interesting to note that different qualified researchers have contrasting opinions on the general outline and design of the study. Wiley (1969) considers the design and its implementation to be exemplary in comparison to other evaluation studies because of its attempt at random assignment of units to treatments and because of its use of the classroom mean as the statistical unit of analysis. Similarly, Carroll (1969) comments that it is one on the few large-scale studies that has well observed the canons of scientific educational research. On the other hand, Aleamoni & Spencer (1969) hold that the study, while professing to be an experimental design, falls in the category of *ex post facto* research. According to the authors, the project is unwieldy and unmanageable, and no valid conclusions can be drawn on what effect the classroom conditions may have on student achievement.

Several reviewers agree that there was no clear distinction between the three methods, nor were the observation scales used for describing classroom activities constructed so as to make control of adherence to method by teachers possible (Carroll 1969b, Clark 1969, Valdman 1969). Scrutiny of the observers' ratings reveals that in all likelihood the TLM students used oral language more than they were supposed to.

Otto's (1969) review is negative towards those aspects of the study which regard the teachers and the part they played. He contends that the orienta-

tion sessions for teachers did not provide exemplary models of effective teaching behaviors for each strategy, that they were no work-shop sessions (which was what was needed), that assistance and supervision was not sufficiently provided, and that the teacher's manual was poorly organized.

Valette (1969) comments that the project results are almost out-dated before they have been disseminated. Her argument is that, in 1969, the distinction between "traditional" and "audio-lingual" is losing some of its relevance because the new traditional texts – the "third generation" texts in Valette's terminology – make creative use of dialogues and pattern drills whereas the "second generation" audio-lingual texts pay attention to formal grammar. In her review, Valette mentions one feature which most reviewers have touched on, namely the fact that the complex findings of the Pennsylvania project have been oversimplified and misinterpreted in various press releases. Stressing the disservice such journalism does to both the project personnel and the foreign-language teaching profession as a whole, she urges anyone really interested in the results to read the full reports.

The Pennsylvania experiment illustrates the difficulties involved in controlling the many variables at work in a broad field study. As was mentioned earlier, we had the advantage of planning our own investigation with the first Pennsylvania report available. Although GUME is an experiment on a smaller scale and in logistic matters should not be compared to the Pennsylvania study, its main objectives and experimental design are similar. Direct similarities and differences, in so far as they can be judged as interesting, will appear on a comparative reading of the respective accounts. In our opinion two essential differences are (a) the much stricter control of the stimulus (teaching) situation that was achieved in GUME by the elimination of one source of error, namely the variation in teacher behavior, and (b) the more specific nature of the independent variables in the GUME project.

The present chapter includes a review of related research from 1964 and onwards. The decision not to include earlier studies was based partly on the fact that they have been reviewed elsewhere, partly on the opinion, forwarded by several authors, that the results of earlier foreign-language teaching research have been largely inconclusive. Of the investigations discussed in this chapter, considerable length was devoted to the Pennsylvania study, a large-scale comparative research undertaking with which the present investigation has much in common.

The investigations reviewed represent a large variation with respect to research models, teaching procedures, age groups, foreign languages taught, etc. The feature which most of them have in common is the comparison

between some kind of cognitive-oriented and some kind of habit-oriented approach.

Although the outcomes of the various investigations display no clear-cut and general pattern, they may be interpreted as slightly biassed in favor of a cognitive-oriented approach. However, several studies demonstrate that what is learned is precisely what has been emphasized in the instruction. That is to say, in cases where the criterion test measures over-all performance by means of different sub-tests, the various groups tend to excel on those parts which correspond to the contents of the instruction. On the other hand, clearly significant over-all differences between methods stressing different aspects are rare. Where such a tendency is found, it has ordinarily vanished at the time of the retention test. The results of various studies indicate that, where grammatical explanations are used, their position in the teaching sequence is of little importance.

CHAPTER 5

THE GUME PROJECT – DESIGN CONSIDERATIONS

Comparative experiments – pros and cons

The present project or, rather, research program, consists of a number of comparative experiments in a field setting. Since we have been concerned, in a broad sense, with evaluation of pre-produced teaching materials, evaluation models besides the one used by us may seem equally relevant. For instance, formative or summative procedures aimed at gauging the teaching materials in relation to stated objectives or some absolute standard (Stake. 1969) might have been resorted to. However, the actual choice of evaluation model should be made in relation to the general character of the research undertaking. In this section we shall describe the character of the GUME project as we see it, thereby also trying to motivate our choice of the comparative experiment as our research instrument.

A distinction is occasionally made between conclusion- and decision-oriented research (Wiley, 1969). The former is performed so that the investigator may draw conclusions about the phenomenon he is studying. Conclusions, however, are tentative by nature and may be modified as more evidence is accumulated. Decision-oriented research, on the other hand, is performed to gather evidence which will be used for generating decisions about actions to be taken. Wiley gives the example of a school superintendent who cannot wait for twenty-five years of accumulated evidence before deciding whether to purchase a language laboratory. If he does so, he will really have decided against it (p. 209).

Cronbach & Suppes (1969), in distinguishing between the two concepts, make the following statement:

> In a decision-oriented study the investigator is asked to provide information wanted by a decision-maker: a school administrator, a governmental policy-maker, the manager of a project to develop a new biology textbook, or the like. — The conclusion-oriented study, on the other hand, takes its direction from the investigator's commitments and hunches. The decision-maker can, at most, arouse the investigator's interest in the problem. — The aim is to conceptualize and understand the chosen problem. (pp 20–21).

If we follow this distinction, which is a distinction in respect of initiatives and basic commitments and not in respect of potentialities for educational improve-

ment, then the GUME project is obviously a conclusion-oriented undertaking. Ideas and hypotheses among the project members have steered the planning and execution of the various experiments; no institution or body has required the project to produce certain varieties of foreign-language teaching materials.

Härnqvist (1972), discussing the two kinds of research, maintains that both can be of utmost importance for educational policy and practices and should be supported each in its own right but perhaps with different methods of resource allocation.

In a paper by Alkin & Johnson (1971), containing comments on the research and development program sponsored by the Swedish National Board of Education, the authors similarly distinguish between conclusion- and decision-oriented research. The former is said to be concerned primarily with achieving a better understanding of a particular set of phenomena whereas the latter is said to be directed toward the solution of a particular problem (p. 3). The two concepts thus seem to be fairly synonymous to basic and applied research respectively.

Summative evaluation of an instructional program should be viewed as a decision-oriented activity the purpose of which is to facilitate a rational decision with respect to the particular program. The product of summative evaluation is expected to be a set of descriptive statements about a single program or about the relative merits of two or more programs (Schutz, 1968). Since the GUME studies have obvious similarities with summative evaluation, concerned with comparisons of end-of-course post-test scores as they are, we want to make the following comment: We do not regard our investigations as instances of program evaluation in the ordinary sense; that is, they do not represent summative evaluation as the concept is generally understood. As will be made clear later (see p. 113 below), our English lessons do not represent complete lessons to be conducted in the ordinary classroom. As a matter of fact, we have not investigated, in a general sense, methods of teaching English as a foreign language, not even methods of teaching English at a certain age level. What we have tried to investigate is methods of teaching certain grammatical structures in English at certain age levels with the main object of finding out if specific variables, although complex ones, facilitate learning. Thus, what we aspire to is to reach some conclusions, however tentative, about the relative merit of specific variables related to specific theories of foreign-language acquisition. We do not aspire to make any decision about which of the alternative series of lessons (or instructional packages, if that term is preferred) should be chosen in order to reach some general objective. The GUME project is, then, a case of conclusion-oriented research; we think this term is to be preferred to basic research since the latter may be associated with research of a laboratory kind, concerned per-

haps with variables that do not have the remotest relevance for the current activities of the ordinary classroom.

Stake (in Wittrock & Wiley, 1970, p. 281) places summative evaluation, formative evaluation, and instructional research (of the variety that GUME represents) on a continuum. The basic difference when going from one end of the continuum to the other is a matter of how much the results can be generalized. Summative evaluation does not permit generalization beyond the particular package. Formative evaluation, which is done within the development of an instructional package, leads to revision and extensions of that package and provides a basis for limited generalizations (although they still pertain to a specific package). Instructional research, on the other hand, is concerned with relationships that hold for a large number af packages.

There has been a good deal of controversy about the value of the comparative experiment as a research tool. Cronbach (1963) proposed two basic types of studies to accomplish the goals of summative evaluation, one of which was the educational comparative experiment. However, as the following quotation makes clear, his point is that it is difficult to implement valid comparative experiments:

> "The aim to compare one course with another should not dominate plans for evaluation. To be sure, decision makers have to choose between courses, and any evaluation report will be interpreted in part comparatively. But formally designed experiments, pitting one course against another, are rarely definitive enough to justify their cost. Differences between average test scores resulting from different courses are usually small relative to the wide differences among and within classes taking the same course. At best, an experiment never does more than compare the present version of one course with the present version of another. A major effort to bring the losing contender nearer to perfection would be very likely to reverse the verdict of the experiment" (p. 676).

In line with Cronbach's view Anderson (1968) argues that, in a comparative experiment, a no-difference result has very low social utility since it cannot facilitate consumer decisions.

Counter-argument on this matter is apparent in an article by Scriven (1967) where the principles of formative and summative evaluation are discussed and where Cronbach's "despair over comparative studies" is optimistically contradicted:

> "If we have really satisfied ourselves that we are using good tests of the main criterion variable (and we surely can manage that, with care) then to discover parity of performance is to have discovered something extremely informative. No difference is not 'no knowledge' " (p. 67).

Thus, according to Scriven, the comparative field study has a definite, though

by no means unlimited, place in evaluation. Cronbach's (1963) second alternative in evaluation studies is one in which comparison is avoided, not with specified goals or objectives, but with another group. His approach includes systematic observation, process analysis and collection of item data rather than test scores. Scriven's point that any measurement of performance incorporates a built-in comparison is also stressed by Wiley (1970):

> "The trouble with the 'time trial' study (Cronbach's term) is that one is almost always interested in a comparison with some other objects, for if one were not, a decision would not need to be made. And given that a comparison is necessary, the constancy of conditions becomes extremely important and is difficult to guarantee without the important concomitants of a comparative experiment". (p. 263).

In decision-oriented research the role of the comparative experiment is thus somewhat controversial. In conclusion-oriented research, on the other hand, where the purpose is to test hypotheses and where specific variables – not complete packages – are being investigated, the comparative study seems to be a natural and frequently used design alternative. Klausmeier (1968), in discussing various research and development strategies, states: "The prevalent form of basic research is the controlled experiment and its variants" (p. 1). As soon as the purpose of the research is to elucidate causal relationships the controlled experiment is considered appropriate by many authors (see, for instance, Wardrop, 1968 and Stanley, 1969). These views reflect a conception of group comparisons as fundamental in experimental research; the demand to relate a particular effect to one independent variable makes them necessary.

A number of critics have pointed out that comparative experiments have yielded non-significant differences between the teaching methods compared. Nachman & Opochinsky (1958), giving a number of examples to illustrate this point, state:

> "Reviews of teaching research have *consistently* concluded that different teaching procedures produce little or no difference in the amount of knowledge gained by students" (p. 245, italics ours).

Stephens (1967) and Grittner (1968), discussing the last half century's comparative research, state that almost no knowledge has been achieved about the relative superiority of one educational strategy over another. However relevant these observations may be it is difficult to take them as valid argument against comparative research. Part of the inconclusive results may be explained by the fact, pointed out by Wallen and Travers (1967), that many investigations have dealt with methods in terms of broad categories, the effects of which have been to cancel each other out.

Whether comparative experiments should be resorted to depends on the purpose of the evaluation. In decision-oriented undertakings, where the objective is to establish the relative value of one or more particular products in relation to specified goals, the more intensive process studies suggested by Cronbach seem relevant. However, in conclusion-oriented research, where the aim is to obtain knowledge about the effects of particular variables, comparative studies are legitimate tools provided the variables have been clearly defined and are educationally relevant.

Since we regard the GUME project as a case of conclusion-oriented research and since we have chosen the comparative study as our research method, we shall conclude the present section by stating that the research design used in our experiments corresponds to Campbell and Stanley's (1967) "design 10", The Nonequivalent Control Group Design. In comparison with this design, our own contained no control group in the traditional sense; GUME 1-5 included three experimental groups (Im, Ee, Es) and GUME A two groups (Im, Es). However, the traditional sense of the term "control group" lacks generality:

> "Thus the traditional notion that an experimental group should receive the treatment not given to a control group is a special case of the more general rule that comparison groups are necessary for the internal validity of any scientific research" (Kerlinger 1970, p.306).

Aptitude-treatment interaction (ATI)

The present project was originally planned with the intention of investigating, by two-way analyses of variance, interaction between teaching method and various levels of "intelligence". Although we have no *a priori* assumption about the overall efficacy of the various teaching strategies compared, our hypothesis is that the inductive-oriented Im method will facilitate learning relatively more for pupils of low intelligence, whereas the deductive-oriented explanation-methods will provide better alternatives for pupils of high intelligence. Support for this hypothesis may be found in, for instance, Wilga Rivers (1968, p. 48).

There has been an intensive search for ATI in recent years. A distinction is usually made between ordinal and disordinal interaction, and it is the latter type that has attracted the researchers' special interest. An interaction is defined as disordinal only when the differences between alternative treatments at two levels of a personological variable (IQ, for instance) are both significantly non-zero and different in algebraic sign. Bracht (1970) made a survey of 108 studies which were designed so as to permit the computation of

interaction. Of all these studies only 5 demonstrated the existence of disordinal interaction. Of these five, four were obtained in cases where the personological variable was of a specific or factorially simple nature. The author comments:

> "Despite the very large number of comparative experiments with intelligence as a personological variable, no evidence was found to suggest that the IQ score and similar measures of general ability are useful variables for differentiating alternative treatments for subjects in a homogeneous age group" (p. 638).

Cronbach & Snow (1969, p 193) comment that there are at present no solidly established ATI relations even on a laboratory scale and no real sign of any hypothesis ready for application and development. The type of interaction study that the authors propose is one in which alternative treatments are developed from a conception of the abilities which are relevant to successful performance in the alternative treatments.

It should be noted that the GUME treatments were not developed with this more subtle ATI concept in mind; at the present time we are not prepared to hypothesize which specific foreign-language teaching variables may be differentially related to which specific personological variables. However, in our case where different levels of one general personological variable ("intelligence") are related to relatively specific treatments, we consider it to be of interest to investigate interactions, although in a somewhat tentative manner.

Age-treatment interaction

The title is somewhat pretentious considering the fact that within the GUME project it is only possible to investigate, in a fairly exploratory way, the relation between age and method. We have no continous age distribution extending through our series of experiments; the levels represented in our studies are 13 years, 14 years, 15 years, and adults respectively. However, in so far as dissimilar main results are obtained between the various levels, this may give rise to further hypotheses concerning age as an independent variable in foreign-language acquisition. We shall make a few comments on the purported differences between child and adult with respect to foreign-language learning.

Wolfe (1967) introduced the notion of "linguistic puberty", stating that it provides a natural linguistic dividing line between child and adult. However, like most authors he does not try to fix any particular age when this puberty is supposed to occur. According to Wolfe the adult, because of his linguistic maturity, profits more from a method utilizing deductive rather than inductive procedures.

A number of authors are unanimous in stressing the need to provide the adult learner with generalizations or explanations of grammatical structures (see, for instance, Fries 1945, p. 29, Agard & Dunkel 1948, p. 282, Nida 1957, p. 41, Ausubel 1964, p. 422, and Rivers 1969, p.75). Ausubel also argues that the mediational role of the native language should be utlized rather than avoided in the teaching of adults.

According to some authors (see, for instance, Saporta 1966) the language learning ability of the child is an inborn mechanism which is lost as the child matures. According to others there is little ground for this hypothesis. Bolinger (1968) has called it mere superstition, Newmark & Reibel (1968) and Reibel (1967) argue that the language learning capability is qualitatively the same in the adult and the child, and Carroll (1971) expresses doubts about a "critical period" and decline in language acquisition ability during the middle school years.

In view of these contrasting opinions we shall make comparisons between the main results obtained at the various age levels. It would have been a great advantage if additional levels had been included in our experiment. However, for exploratory purposes the present material may suffice to indicate substantial differential effects of our methods at the various age levels.

The "zero-point" problem in research on foreign language teaching

When a student begins to study a foreign language, he usually starts something relatively new. From a research point of view this is a great advantage because a natural "zero-point" is thereby given. This at least theoretical advantage, pointed out by Carroll (1967), is not at hand in the GUME project since the students were in their third (GUME 4), fourth (GUME 1-3), or fifth (GUME 5) year of English; in the adult group this background of formal training varied from 0 to 3 years (see p. 105 for further comments). It is reasonable to assume, and it is also confirmed in a Swedish investigation (von Mentzer 1970, p. 52), that the variation among students as regards proficiency in English is large. This variation is controlled statistically in the GUME project by analysis of covariance (to the extent that it is measured by our pretests). One might venture the guess, however, that in a comparative study such as the present one, where the students have had two to four years' teaching before they enter the experiment, the *amount* of treatment must be fairly large if differences between treatment effects are to be detected.

In GUME 1-3 the treatment proper consisted of six lessons (excluding a preparatory lesson), which might be judged as very little, but it was what the resources permitted. In order to counterbalance the shortage of time we chose to make the teaching strategies distinctly contrastive and, in certain

respects, extreme. Thus the students in the explanation-groups were given grammatical explanations for 9 minutes (out of 30) each lesson, which probably is more than any teacher would consider optimal. Against the background of the short lesson series it was considered necessary to give the treatment variable (the explanations) emphasis by giving it disproportionately long time each lesson. Although this procedure is defensible in conclusion-oriented research, we are well aware that problems will arise in generalizing the results to the ordinary classroom situation.

It may be hypothesized that the rather long explanation time used in GUME 1-3, instead of giving the treatment variable a fair chance to "break through", might work in the opposite direction, creating lack of concentration and boredom in the pupils. Therefore, in the GUME 4 and GUME 5 studies a different strategy was adopted; no restriction was put on length of explanation time, but the explanations were made "optimal". By this somewhat pretentious term we want to indicate that explanations were introduced when they were judged relevant. As it appeared, this strategy had the effect that the explanations usually became shorter (than in GUME 1–3) and that the Ee and Es explanations could, and did, vary in length. The length of the grammatical explanations in the adult study are in line with those of GUME 1–3, i.e. they are of substantial length so as to form a distinctive contrast with conditions in the non-explanation group.

A relevant summary of the present discussion seems to be that it is difficult to predict the effect of the explanation-time-variable for counteracting the lack of "zero-point" mentioned by Carroll. Of course, one way to circumvent the problem would have been to utilize a foreign language not encountered by the students or to use nonsense-syllable materials, but these courses of action were never contemplated. It is to be supposed that, in the present project, the absence of a zero-point operates against revealing true treatment differences, if any.

Replications in educational research

The ultimate goal of our research is to gain some knowledge about the impact of a specific independent variable (explanations vs non-explanations in teaching grammatical structures in a foreign language) on pupils' acquisition of the foreign language. By "knowledge" is implied that we hope to be able to generalize our results. However, as Kerlinger (1969) has pointed out, generalizations in educational research are very probabilistic in nature. As a means of providing stronger evidence, thus making generalizations more probable, he suggests replications of experiments; they are particularly urgent in cases where random samples cannot be obtained.

Our experiments may be viewed as a series of replications. In fact, the same type of study was performed with different samples in different places, with different measuring instruments and different experimental manipulations. The modifications which were undertaken from study to study will be presented in due course. The replications should also serve the purpose of increasing the external validity of our inferences since, if our independent variable is one of consequence, replications under somewhat different conditions should produce similar results (cf Wiley 1969).

CHAPTER 6

STATISTICAL TREATMENT

The sampling and statistical units

As is often the case in educational research, it was not possible to assign pupils randomly to treatments. For administrative reasons intact school classes had to be used in our experiments; the school class, not the individual pupil, is thus the sampling unit. This perennial problem has been commented on by several authors. Wiley & Bock (1967) argue that the classroom is the proper sampling unit because the pupils in a common classroom share a lot of influences, such as the physical aspects of the classroom, distractions due to discipline problems, the time of day of the class etc.; thus the pupils' performance on outcome measures is interdependent, and their scores cannot be considered uncorrelated.

> "In the language of experimental design, it is the classroom and not the pupil which is the experimental unit. Thus, it is the classroom mean rather than the score of an individual pupil which is the fundamental datum of the experiment. Correspondingly, it is the number of classrooms and not the number of pupils which determines the number of degrees of freedom available in the data" (pp. 355–356).

The authors make a case for experiments with a limited number of school classes, provided blocking on schools is possible. As will be shown later (table 2, page 94), blocking on schools was possible only to a very limited extent in our experiments.

Wardrop (1968), reviewing several curriculum evaluation projects, states that their major limitation was the fact that the statistical analyses were based on individual students' performances when the experimental unit should have been defined as the classroom. Raths (1967), supporting this view, stressed that the experimental units should be the smallest units of students in the study to which treatments have been assigned randomly and which have responded independently of each other for the duration of the experiments. At a recent symposium on the evaluation of instruction, the proceedings of which are summarized by Wittrock & Wiley (1970), there was some controversy on the matter of which unit is to be considered the most meaningful from a psychological and teaching point of view (see p. 281 ff).

Fletcher (1968) used fictitious data including a limited number of strongly deviant observations in order to illustrate the difference in results when the statistical anlyses were performed on group data and individual data respectively. He applied both a parametric and a non-parametric test and found that, in both cases, the analyses performed on group data permitted rejection of the null hypothesis at the .01 level whereas the anlyses performed on individual data did not. In interpreting Fletcher's findings it should be noticed that the treatment effects were caused by a few deviant subjects in each group. He concludes:

> "It seems, therefore, that in the absence of truly homogeneous subjects, and without sufficient knowledge concerning actual treatment effects, the most meaningful interpretations will follow analyses in which the individual subject's score represents the experimental unit. This writer sees nothing but danger in using group means (or any summary statistic) as the experimental unit in statistical analyses – parametric or non-parametric." (p. 160)

Glass & Stanley (1970) discuss deviant scores from a somewhat different angle. In their illustrative example, consisting of two treatments and two intact groups in each treatment, two "trouble-makers" influence the behavior of the groups to which they were assigned to such an extent that the respective groups learn nothing whatever; the example serves to illustrate how interdependence of responses influences the results in an extreme case. Assuming that interdependence of responses exist, the analysis should be performed on group means; however, in the example quoted the analysis would not be worth the bother because of the small number of degrees of freedom. According to the authors, *independence* is the crux of the matter. Although the assumption of homogeneous variances and normality between the replications of an experiment are easily tested,

> "... the researcher will usually be faced with the task of making a considered judgment of the degree of independence of the replications rather than the task of applying a particular statistical test. His judgment must be based on an intimate knowledge of the dynamics of the experimental setting" (p. 506).

Lumsdaine (1967) treats the problem of concern here in experimental situations where it is important to rule out the possible effect of "group dynamics" influencing the results. He suggests different ways of assigning subjects randomly to treatments even in cases where intact groups are used. However, in cases where it is not feasible to achieve this, he states that the group mean should be used as the statistical unit of analysis.

We think it is reasonable to assume that, in ordinary teaching, the pupils' responses are partly interdependent. However, as the instruction becomes more pre-produced, it is equally plausible to assume that the pupils respond more independently. This has probably been the case in the GUME project

where the lessons were "canned", leaving practically no room for the teacher to influence the pupils. From this particular point of view, then, it would be of minor importance which unit, the class or the individual, were used as the unit of analysis in the present study. However, the question must still be answered with respect to error variance for classes and individuals respectively. Tatsuoka (1969) equates sampling of classes with cluster sampling without a subsequent second-stage subsampling (random sampling of individuals from the clusters).

> "Adequate analysis is possible only when the investigator is fortunate enough to be able to assign a substantial number of classes (say, ten or more) *at random* to each treatment condition. He may then use the class means as the basic observations and essentially follow the usual analysis-of-variance or multiple regression procedures" (p. 480)

Kerlinger (1970) is less explicit on the number of classes required:

> "If a *fairly* large number of classes are selected and assigned at random to experimental and control groups, there is no great problem" (p. 316, italics ours).

Carroll (1969 b) regards it as a sort of unwritten rule of thumb in educational and psychological research that there should be a minimun of about twenty observations within a group in order for the experiment to have sufficient power to reject the null hypothesis in a reliable way. If this were an absolute requirement, and if the school class mean had to be used as the unit of analysis, sixty classes would have been the minimum requirement in all GUME experiments except one (GUME A), an unwieldy number considering the administrative work involved and the resources in personnel and money available for the project. Thus, when the researcher has a very limited number of classes at his disposal, he *is* in an awkward position.

The alternative usually suggested is to check the similarity of the treatment groups on available background variables; if the groups do not differ, there being at least no evidence against the assumption that the sampling procedure produced no bias, analyses applicable to simple random samples may be used. It should be noted, though, that performing the analyses on individual scores will increase the risk of a type I error because the estimate of error variance may be smaller than it should be – how much depending on how far from random assigment the composition of the school classes was (cf Tatsuoka 1970, p. 480).

Among our experiments only GUME 4 (9 classes per treatment) approximates the sampling requirement discussed; the remaining experiments contain 2, 3 and 4 classes per treatment. We shall follow the procedure of comparing the treatment groups on various concomitant variables and, in case

where there is no evidence against the equality assumption, perform the analyses on individual scores. In a number of cases, however, we shall also perform the corresponding analyses on school class means in order to compare the two sets of results; considering the current dispute on the sampling and anlysis unit problem, we think these comparisons may be of interest *per se*.

Measures of progress

In comparative experiments where the subjects have been given a pretest as well as a posttest, the raw difference score between the two measures is often used as a measure of progress. According to many experimenters, such a measure makes intuitive sense as a measure of change in performance or gain in skill. A difference score may also be considered to reflect a construct such as "learning ability" on a certain task. As compared to various types of adjusted scores, the meaning of a raw score is also more easily communicated to an audience with limited experience in statistics. However, several authors have pointed out the unsatisfactory psychometric properties of raw difference scores.

Du Bois (1962) noted that unevenness in scale units may be critical when the numerical operation is subtraction, as in the computation of a difference score. He further claims that crude gain, defined as posttest less pretest score, is practically always negatively correlated with initial score. It can be shown that for the correlation of crude gain and initial score to be positive the following inequality must hold: $r_{12}s_2 > s_1$, in which the subscripts 1 and 2 represent initial and final score respectively (op. cit., p. 79). However, his contention that the standard deviation of the final score is seldom greater than the standard deviation of the initial score is debatable (cf Anastasi 1958, p. 194 ff). In chapter 12 we shall present a number of correlations of the kind discussed here. As a possible improvement to raw difference scores DuBois proposes residual, or regressed, scores.

Campbell & Stanley (1967) observe that "that most widely used *acceptable* test is to compute for each group pretest-posttest gain scores and compute a *t* between experimental and control groups on these gain scores" (p. 193, italics ours). However, the authors add that randomized blocking on pretest scores and the analysis of covariance are usually preferable to simple gain-score comparisons.

Kerlinger (1970), discussing what he calls the "classical design" of research, simply states:

> "Usually, the difference scores $Y_a - Y_b = D$, are analyzed, a simple and efficient procedure" (p. 309).

Cronbach & Furby (1970) emphatically argue against the use of raw gain scores. According to the authors, such measures lead to fallacious conclusions, primarily because they are systematically related to any random error of measurement. The authors suggest an improvement in the form of a multiple regression procedure by which true scores are estimated. Similar arguments are presented by Cronbach & Snow (1969). The authors warn against the use of gain scores as the dependent variable in an experiment on instruction.

> "Basically the pretest score is an aptitude and should be treated along with other aptitudes. The raw posttest score is the proper dependent variable" (p. 14).

In each of our experiments we have calculated raw gain scores. However, simultaneously we have made use of dependent variables more in line with those proposed by Cronbach and others. The different types of dependent variables used will be presented in connection with the discussion of design alternatives (see the next section).

To conclude the present discussion we shall introduce a second type of progress score which we assume will have certain advantages over raw gain scores. This measure is simply the ratio between the individual pupil's actual raw gain score and his possible raw gain score. The index is expressed as a percentage, thus:

$$\frac{\text{Actual raw gain} \times 100}{\text{Possible raw gain}} = \%$$

The measure will hereafter be referred to as Actual/Possible Progress (A/P P). If a pupil scores very high on the pretest, there is not so much room for progress because of ceiling effects. The A/P P index takes care of this, giving more weight to progress at the upper end of the scale. DuBois (1962) pointed out that high raw gain values are not wide enough as compared with units in the median position; it is precisely this deficiency that the A/P P index is intended to remedy. An example will clarify the point:

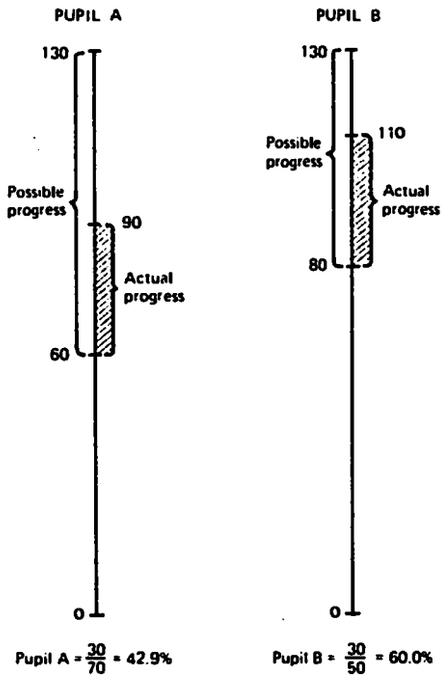


Figure 1. Illustration of two hypothetical Actual/Possible Progress (A/P P) scores.

Pupils A and B have been given the same test (containing 130 items) on two occasions, as pre- and posttest. Pupil A has 60 points on the pretest and 90 on the posttest, pupil B 80 on the pretest and 110 on the posttest. The improvement of both these pupils is thus 30 points in terms of raw gain scores. Possible improvements for the two pupils are 70 (130-60) and 50 (130-80) points respectively. Their A/P P scores as computed by the above formula become 42.9 (%) and 60,0 (%) respectively; thus pupil B has made the greater progress according to this index.

The A/P P index will be used parallel to raw gain scores for purposes of comparison.

Three alternative treatment models

In order to increase the precision of what would otherwise be a completely randomized analysis of variance design, experimenters often make use of

concomitant variables. The following three techniques are ordinarily utilized:

(1) The block design (treatment-by-levels design)

(2) Analysis of covariance

(3) Analysis of variance of difference scores.

Where (1) is used, the levels are defined along the scale values of the concomitant variable, and subjects within levels are assigned to treatments at random. This assignment to the various treatments is usually made in the same proportion for the different levels in order to simplify the analysis. It should be observed that no blocking in this sense was achievable in the GUME experiments. As has been mentioned earlier, the school class is the sampling unit; thus classes, not individuals, are randomly assigned to treatments. However, since information was available about the pupils in various control variables (e.g., IQ and pretest), it was possible to make a subsequent assignment of pupils to different levels of the control variable. In actual practice we divided the experimental groups into three thirds according to their standing on the control variable. Since the treatment is fixed for each individual, the procedure in all likelihood brings about varying numbers of pupils in the different cells. This procedure, which for convenience may be called "pseudo-blocking", was applied in order to reach at least a tentative answer to the question of interaction between IQ (as defined by our test) and treatment.

Analysis of covariance provides a second alternative by which a potential source of error variance may be controlled. Federer (1955) suggests the following rule to experimenters:

> "If the experimental variation cannot be controlled by stratification, then measure related variates and use covariance" (pp. 483–484).

By *related* is understood that the gain in precision, relative to one-way analysis of variance, is greater the higher the correlation is between the covariate and the dependent variable. In our experiments, where stratification in the strict sense was not possible, we have consistently resorted to analysis of covariance. In order to find out which set of covariates predicted the dependent variable maximally, a step-wise multiple regression procedure was applied.

The third type of design used to increase precision is analysis of variance of difference scores; it is sometimes referred to as the method of differences. Some properties of difference scores were discussed in the preceding section. The method of differences is probably most frequently used in cases where X and Y may be considered parallel forms of a test; this is the case in the GUME project where we have also used the method.

The three types of design just mentioned have been compared with respect to their precision by several authors.

Federer (1955) preferes (1) to (2), although he makes no systematic comparisons.

Outhwaite & Rutherford (1955) give empirical evidence which suggests that when the number of replicates per treatment is approximately equal to the number of treatments, a modified Latin square design is more efficient than analysis of covariance.

Gourlay (1953) compared (1), (2), and (3) and concluded that analysis of covariance always results in the most precise experiment.

Cox (1957) made an extensive study of various techniques for employing concomitant information in an experimental design. He concluded that (1) is more advantageous when $r < .60$ and that (2) becomes appreciably better than (1) only when r is as large as .80 or more. He also noted that (1) is reasonably efficient for any form of smooth regression, not just for linear regression. However, if the distribution of X is leptokurtic, the efficiency of (1) is lowered due to the end blocks having units with widely discrepant values of X.

Feldt (1958) compared pairs of designs based on constant N. He notes that for $r < .40$ design (1) results in approximately equal or greater precision than (2), for $r \geq .60$ the advantage is in favor of (2). When $r < .20$ and there are small values of N, neither (1) nor (2) yields greater precision than a completely randomized design. It is interesting to observe Feldt's comment that "the marked superiority of covariance occurs for values of r which are rarely encountered in educational and psychological experiments" (p. 347). Design (3) was found to have clearly less precision than either (1) or (2); unless a substantial correlation exists between the control and criterion variables, (3) results in considerably lower precision than the completely randomized design. In discussing his findings Feldt states that, beside precision, consideration should be given to other factors, such as simplicity of design in communicating the results, the extent to which valuable supplementary information may be derived from one or another of the designs, and the effect of possible departure from the assumptions on which the designs are based. The assumption of linear regression may constitute a restriction on the usefulness of (2). Feldt concludes that the less stringent assumptions of (1) more than compensate for the relatively small advantage in precision which may obtain for (2).

Thus, none of the three techniques seems to be superior under all circumstances. For comparison's sake all three will be used in presenting the GUME results. We do not aspire to any strict methodological investigation involving computation of comparative indices, but rather a parallel presentation of (1), (2), and (3) which will provide some empirical evidence on whether the three techniques produce similar results. Considering the fact that, in the GUME

experiments, (1) was achieved by what we have called "pseudo-blocking" and the circumstance that (3) seems to be a generally inferior technique, we will attach greater *a priori* importance to the analyses of covariance. In the following section we will further comment the latter technique as it was applied in our experiments.

Comments on the analyses of covariance

As was stated previously (p. 73) we shall perform all analyses of covariance on individual scores and, in cases where the number of degrees of freedom is not too limited, on class means. Elashoff (1969), in discussing the assumptions underlying covariance analysis, points at two major interpretation difficulties when individuals are not randomly assigned to treatments. First, there is the probability that some original bias between the treatment groups is still present in the adjusted scores because the effect of a disturbing variable was overlooked. This is equal to stating that the treatment groups were not randomly selected originally but differed with respect to some variable related to the dependent variable in the experiment. In the GUME experiments there is one type of selection present; part of our school classes represent the easier course in English, part of them represent the advanced course in English. However, the two courses will be treated as separate populations in our analyses of covariance. There is no reason to believe that the pupils were not randomly assigned to classes *within* each course. A certain variability in socio-economic variables is known to exist between school districts; bias in these respects between the treatment groups in thus a potential "disturbing variable". However, it can be easily checked and need not be overlooked. The second problem pointed out by Elashoff is that extrapolation may be needed when the x variable shows real differences among treatment groups. In order to check this eventuality we shall compare the covariate means of the treatment groups by analysis of variance.

Following the previously (p. 74) mentioned recommendation by Cronbach & Snow (1969) we shall use the posttest scores as the dependent variable in our analyses of covariance. In order to find out which combination of concomitant variables (of which all were unaffected by the treatment) showed the highest multiple correlation with the posttest, a step-wise multiple regression procedure was applied in each experiment (see Draper & Smith 1966, chapter 6). The procedure starts by selecting the highest zero-order correlation with the criterion and then proceeds by selecting the remaining variables in order as they contribute to an increase in R. The results of these analyses are presented in Appendix 5. Two facts are apparent from the table: (a) in 8 cases

of 10 the pretest showed the highest zero-order correlation with the posttest (b) adding one or more variables ordinarily increases the multiple correlation very little. In view of these facts we have decided to perform analyses with only one covariate, in all cases the pretest. It should be noted that in those cases (GUME 2 ak, and GUME 3 ak) where the pretest was not selected first in the multiple regression procedure, its correlation with the posttest was identical or close to identical with that of the variable selected.

Analysis of covariance procedures are based on the assumption that the concomitant variable, x (in our case the pretest), is measured without error. Although the reliabilities of our pretests are relatively high (see Appendix 3) they still contain errors or measurement. Lord (1960) suggested a method for correcting analysis of covariance when the control variable is fallible, but his test is limited to only two treatment groups. Härnqvist (1968) has extended the method to be valid for more than two groups. Essentially, the method is based on estimated true x scores. As Härnqvist has demonstrated (op.cit., pp. 54-55), the regression for the true y scores on the true x scores is equal to the corresponding regression for the observed y scores on the observed x scores divided by the reliability coefficient of the x variable; the reliability of the y variable does not enter into the correction. The general effect of the correction is to increase the slope of the common within-groups regression line. Depending on the position of the various treatment means relative to the regression line, the correction can bring about differences between groups that did not exist in the observed means *or* it can delete differences that did exist between the observed means. The figure below is a simplified illustration of the point. In case (a) the steeper slope of the corrected line implies that the differences between the observed and the expected means usually decrease. Thus in this (fictitious) case the differences between the adjusted means of the high group (Im) on one hand and the two low groups (Ee, Es) on the other become smaller. In case (b) the correction has the opposite effect, i.e. the differences between the high and low groups tend to increase.

In a recent investigation of school performance in relation to various background variables, Svensson (1971) has applied this correction technique. The author also gives the computational procedures in an illustrative example based on his own data (p. 153 ff).

Porter (1967) developed a correction method based on more than one fallible covariate. He suggests that estimated true scores be used as a covariate when the reliability of the covariable is between .5-.9. For lower reliabilities he found that the agreement between the F distributions and their theoretical counterparts was not very good.

Thistlethwaite (1969) compared conventional analyses of covariance with analyses based on estimated true scores in an actual quasi-experimental study. The reliabilities of the so-called press scales used as covariables varied from

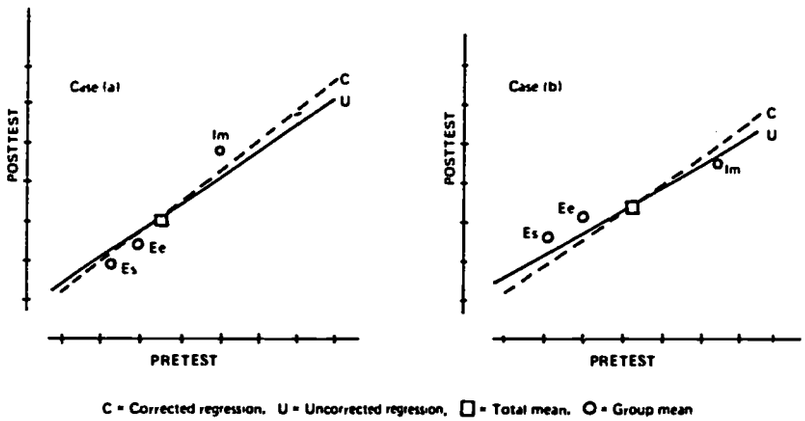


Figure 2. Two fictitious cases illustrating possible effects of unreliability correction on differences between treatment means.

.48 to .87. There was considerable correspondence between the fallible and true score analyses. The author concludes that conventional covariance analysis seems to provide fairly robust, though slightly inflated, estimates of the significance levels which would be obtained with appropriate correction for measurement errors in the covariable.

It might be argued that in a study such as the present one significant differences between treatments are of greater interest than no-difference results. Thus analyses based on estimated true scores might preferably be limited to those cases where it could be predicted that the correction procedure would *increase* differences between treatments. However, we have followed the procedure of computing both analyses in all cases. Considering the high reliabilities of the pretests (see Appendix 3) the corrections will not drastically affect the results obtained by conventional analysis. One exception to this may be GUME 5 ak where the pretest reliability is only .59.

We shall conclude this section by briefly commenting on some of the assumptions underlying covariance analysis. It is obvious that, in each of our experiments, the x variable is statistically independent of the treatment effects; the x variable was measured prior to the administration of treatments, and treatments were assigned to groups at random. The assumption of linearity between x and y has been checked by inspection of x-y scatter plots for each treatment group; no departures from linearity have been discovered. Homogeneity of regression is a necessary requirement for the investigation of treatment effects. As Feldt (1958) has pointed out, heterogeneity of regression in covariance analysis implies the presence of interaction in a treatment x levels design. A quotation from Cronbach & Snow (1969) is to the point here:

"The finding of significant heterogeneity that novice investigators usually view with distress really signals the possibility of ATI, and should be examined further with that in mind" (p. 22).

As has been stated earlier we will perform two-way analyses of variance in order to investigate interaction effects. These will of course be of particular interest in cases where the corresponding analyses of covariance have demonstrated heterogeneity of regression.

Covariance analysis is based on the assumption that the distribution of adjusted scores for each treatment is normal. Atiqullah (1964) has pointed out hat non-normality in the distributions of y's has little effect on the F-test when the distributions of x's is normal. Examination of the x distributions has shown that they are approximately normal in all cases.

Statistical treatment of opinion scales

So far we have discussed possible experimental outcomes only in terms of learning effects. However, the attitudes of both pupils and teachers to the teaching procedures are manifestations which ought to be investigated as reliably as possible. Some kind of measurement or control of non-cognitive outcomes is an important aspect of any evaluation study. By means of questionnaires we have obtained the opinions of pupils and teachers on various aspects of the treatments. The questionnaires consist of two parts, one with fixed-alternative answers and one with open answers. The questions of the fixed-response variety include 3-, 4-, and 5-point scales intended to measure attitudes towards technical as well as general procedural matters. In each questionnaire we have lumped together a number of questions which we think reflect the students' generalized attitude towards the teaching procedures as a whole. The results obtained on these questions will be presented in chapter 11.

The pupils' attitude scores thus form a second type of dependent variable (besides the posttest scores) in our experiments. The outcomes on the two types of dependent variables will be compared; if the superiority of one particular method, in terms of learning effects, is supported by sympathetic attitudes towards the same method, its case is particularly strong.

The pupils' responses to the questions of the open answer type will, for reasons of space, receive only brief mention. Similarly the teachers' questionnaires will only be briefly commented on; these questionnaires were administered in order to obtain information for further refinement of the lessons (the teachers did not take an active part in the teaching, the lessons being "canned").

Survey of statistical analyses

In the present chapter we have given our rationale behind the various analyses performed. Below we shall present a summarized description of the analyses and the order in which they will appear (in chapter 11).

ANALYSES OF MAIN TREATMENT EFFECTS

Separate analyses will be presented for the two courses, sk and ak, in cases where the total age group contains both (i.e. GUME 1-3, and 5). For all experiments the analyses will be performed with the individual posttest score as the unit of analysis. These analyses will be made in a conventional way on the one hand, and with correction for covariate unreliability on the other. For each experiment two one-way analyses of variance will be performed and the results will be compared with those obtained in the analyses of covariance. One analysis is based on the raw difference score between posttest and pretest, the other is based on the A/P P score where the raw difference is related to the theoretically possible difference. All analyses will be performed on individual scores. As a second control, main treatment effects will be investigated by an analysis of variance (two-way), in which case the experimental sample will be divided into three roughly equal levels according to pretest scores. These analyses will also be utilized in searching for interactions (see below). Analyses with the school class means as the unit of analysis will be performed in cases where the number of degrees of freedom do not rule out this possibility (i.e. GUME 1 sk, GUME 2 sk, GUME 3 sk, GUME 4, GUME 5 sk and ak). In all cases the pretest will constitute the covariate and the posttest will make up the dependent variable. In a limited number of cases (GUME 1--3) a retest was also administered one month after the experiment; in these cases the retest will form the dependent variable and the pretest, the covariate.

INTERACTION ANALYSES

In each experiment, two two-way analyses of variance are performed, both with the posttest as the dependent variable. In one analysis, strafication is made according to scholastic aptitude ("intelligence") scores, in the other according to pretest scores. These calculations will provide opportunity to investigate aptitude *x* treatment and achievement *x* treatment interaction respectively. In both analyses the sk and ak samples will be pooled.

ANALYSES OF ATTITUDES

The effects of treatments on the pupils' attitudes towards the teaching procedures will be measured by means of a non-parametric test. In GUME 1-5 a k-sample test will be used, in GUME A a two-sample test.

CHAPTER 7

OVERALL DESCRIPTION OF THE EXPERIMENTS

Our experiments, including the preparatory and data treatment stages, are dispersed in time over a four-year period. Although practically the same hypotheses are continuously being tested and although the designs applied are very similar, differences in various respects do exist between the experiments. The present chapter is intended to clarify these differences and, more generally, to provide an overview of the research activities within the GUME project. Essential details of each experiment will be presented in tabulated form (table 1, p. 91), and the chronological sequence of the experiments will be given in graphic form (fig. 3, p. 92).

Independent variables

The independent variables, i.e. the teaching procedures as they were adopted in the lesson series, will be discussed in detail in chapter 9. Their main characteristics are stated below in order to provide the necessary background.

THE IMPLICIT METHOD (Im).

This method, based on the habit formation theory, is a relatively "pure" audio-lingual method, strictly systematized but with no explicit formulations of either what the drills are about or how the problems should be solved. The pupil's attention is directed to the crucial features of the sentence by way of analogy or contrast, and the systematized drills are supposed to result in a subconscious assimilation of the rules. The Swedish language is not used on any occasion. It is clear that in this exclusive use of the target language the Im method has a facet in common with the original direct method. However, it is also evident that it owes its heavily structured drills as well as the dialogues to the audio-lingual method. In sum, the implicit method is an inductive approach in which the pupil is left to draw what conclusions he can from the drills. We believe that the majority of the teachers, rightly or wrongly, consider the Im method to be the one coming closest to the method suggested by the Curriculum.

THE EXPLICIT METHODS (Ee AND Es)

Both our explicit methods would fall under the cognitive code-learning category which stresses intellectual understanding of what one is doing. The pupil is made consciously aware of the functioning of the language by verbalized generalizations and explanations about what he has just heard, spoken, read, or written. It is worth pointing out that no grammar rules in the old sense are given, no rules for the pupils to learn, but there are just explanations of and comments on what the pupils are doing in the drills. The explicit methods (in GUME 1-5) are not to be compared with a grammar-translation method; in fact, a large part of the time is taken up by structure drills, the same as in the implicit method. The mixing of structure drills and generalizations is in line with the deductive-oriented, modified audio-lingual approach described by Rivers (1968).

There are two variations of the explicit methods. The first version, the Explicit-English method (Ee), gives the explanations in English. The second, the Explicit-Swedish variety, uses the Swedish language. The explanations in English and Swedish are, however, not merely translations of each other, as the Swedish version also includes comparisons with the corresponding Swedish structures.

It should be noted that two of the approaches, Im and Ee, correspond to the intentions as expressed in the Curricula for Swedish Schools (cf. chapter 3). Although the Curricula do not forbid the giving of explanations or even rules in Swedish, it is clear that the Es approach is least in accordance with the methodological intentions of the Curricula.

To sum up:

the Implicit method (Im) corresponds to an inductive-oriented audio-lingual method without generalizations.

the Explicit-English method (Ee) corresponds to a deductive-oriented audio-lingual method with generalizations in the target language.

the Explicit-Swedish method (Es) corresponds to a deductive-oriented audio-lingual method with generalizations in the source language (Swedish) and comparisons with Swedish structures.

These are the independent variables compared in GUME 1-5, i.e. the experiments performed at the comprehensive school level. In GUME A, the part project at the adult level, only Im and Es are compared. In the latter experiment the Es method differs from the previous Es methods in that it comes closer to a traditional grammar-translation method; further description of the approach will be given on page 125.

Project history

In the present section a chronological account will be given of the various part projects and their main characteristics.

The first three experiments, GUME 1–3, may be regarded as one unit. They were planned simultaneously and performed in close succession. GUME 1 got under way early in October, 1968, and GUME 3 was finished in March, 1969 (see fig. 3, p. 92). The experiments were performed in grade 7 where the pupils are approximately 14 years old. The experimental sample in each part study consisted of 18 school classes, 12 of which represented sk (the advanced course) and 6 ak (the easier course); the ratio of sk/ak classes was intended to reflect the actual proportion of pupils taking the two courses. Within each project and course the classes were randomly assigned to teaching strategies. Thus each of the three part projects consists of two experimental samples (sk and ak) which will be treated separately in our analyses; in all GUME 1–3 involve six parallel experiments. The 54 school classes chosen for GUME 1–3 represent considerable geographic variation within the Gothenburg area. GUME 1 utilized classes from the western and central parts of Gothenburg, GUME 2 classes from the central and nothern parts and GUME 3 classes from the north-western and eastern parts as well as Mölndal, a town bordering on Gothenburg.

Within each of the part projects, one specific area of English syntax known to cause Swedish pupils great trouble, was chosen for investigation. The distribution of grammatical problem areas among the part projects is as follows:

GUME 1 The do-construction
GUME 2 The some/any dichotomy
GUME 3 The passive voice

This choice of specific area of investigation is thus the main difference between the three projects. Except for this choice the three part projects should ideally be identical. In reality it is difficult to make any statement on the exact degree of similarity between the teaching procedures in the three projects. The coordination and constant exchange of ideas between the program constructors and the present writer was intended to achieve this end as far as possible. However, viewing the different studies as replications of each other, slight differences in procedural matters should be persmissible (cf. p. 69). The reader is also referred to the Appendices of the GUME 1–3 reports for detailed accounts of the independent variables.

Each teaching strategy within each part project consisted of 6 lessons, each lesson lasting 30 minutes. In the explicit classes explanations and analyses

took 9 minutes per lesson. The time allotted to explanations was taken from the drills and exercises.

All the lessons were recorded on tape. The pupils listened to the "canned" lessons by audio-active headsets with induction receivers. In the ordinary classrooms telephone wires had been installed to create a magnetic field. This arrangement, a simple sort of language laboratory, could also be supposed to facilitate concentration. Three assistants provided the instruction and transported the necessary material (headsets, tape-recorders, projectors, teaching equipment). The assistants were university students without teaching experience and their sole function was to start the tape and hand out the booklets containing the lesson material. They did not intervene in the instruction proper, nor did the teachers, who were present purely as observers and guardians of law and order in the classroom.

After the pretest but before the actual experiment started the pupils received a pre-teaching period, i.e. a short lesson aimed at teaching the pupils how to handle the earphones and how to do the oral 4-phase drills; it was also intended as a test of the equipment. With minor variations, the experiments (including pre-, post-, and IQ-testing as well as administration of questionnaires) took four weeks to finish. Five weeks later a retest, identical with the pre- and posttests, was administered. GUME 1–3 were the only part projects where retesting was done.

GUME 4 and GUME 5, the next two part projects in chronological order, were planned and executed simultaneously; the experiments proper were performed during April, 1970. In order to investigate age groups different from those of GUME 1-3, GUME 4 was undertaken in grade 6 where the pupils are approximately 13 years of age, and GUME 5 in grade 8 where the pupils are approximately 15. The two methodological experts who constructed the GUME 4 lesson series are identical with those responsible for GUME 1 and GUME 2 respectively; the expert who constructed the GUME 5 materials is identical with the person responsible for GUME 3. The main differences between GUME 4 and GUME 5 will be clear from the following description.

In GUME 4 the duration of the experiment was doubled in comparison with GUME 1–3; the three lesson series (Im/Ee/Es) thus consisted of 12 lessons each. The explanation time (in Ee and Es) also differed from those practised in GUME 1–3 where approximately 1/3 of the lesson time was taken up by explanations (see p, 68 for a discussion of this topic). Despite the fact that there was no deliberate attempt to keep the length of the Ee and Es explanations equal, they nevertheless became almost identical in this respect (see report GUME 4, p. 38).

In GUME 1-3 the three part projects concentrated on one syntactic structure each. In GUME 4 the pupils were exposed to a wider range of grammati-

cal problems (see p. 118). Considering the length of the lesson series, this greater variety of content was thought necessary in order to motivate the pupils.

As in the previous studies, the lessons of GUME 4 were tape-recorded. The ordinary classroom teacher administered the lessons which implied handing out booklets, starting the tape-recorder, and supervising the pupils. The teachers were not supposed to give any help of a linguistic kind. The pupils did not use headsets with earphones (as in GUME 1-3) but in each classroom extra loudspeakers were installed to provide optimal listening conditions. One modification was made with respect to the teacher role: in order to let the live teacher control pupil activities with respect to the *oral* parts, the teachers were carefully instructed to activate the pupils into repeating after the tape and to indicate, by pointing, etc., which of the pupils should answer a particular question. This participation by the teachers was thus intended as a check on pupil activities and should, if carried out according to instructions, be almost identical among the teachers.

In grade 6 no division into sk and ak courses in the subject of English has yet occurred. The class-teacher system is still prevalent, which means that practical problems (disturbances in research schedule because of unforeseen circumstances, etc.) can be more easily solved than in classes at the upper stage (grades 7-9) where a number of teachers will be affected by such changes. Among a surplus of teachers willing to participate in the GUME 4 experiment, 27 classes were chosen among those using a particular textbook (Ashton-Olsson, "Hands up") and showing the greatest conformity in a number of characteristics (number of pupils, boys/girls ratio, age of teacher, etc.). All the classes are from Gothenburg though with a large over-representation of classes from the northern and western parts of the city.

GUME 5 utilized school classes from the eastern parts of Gothenburg and from Mölndal. Since classes participating in the previous years' experiments (GUME 1-3) had to be excluded from GUME 5, it proved difficult to enlist a sufficient number of classes; two ak classes from municipalities in the county of Bohuslän, situated some 10 Swedish miles north of Gothenburg, had to be included. In all 12 sk (the advanced course) and 12 ak (the easier course) classes were included in the investigation; within each course the classes were randomly distributed among the three treatments.

In grade 8, where GUME 5 was undertaken, the pupils have been divided into sk and ak for two years. It may be argued that in grade 8 it would have been better to concentrate on one of the courses, trying to optimize the teaching materials for that course rather than making something intermediate and perhaps non-optimal for both. However, in the light of the curriculum (supplement English, p. 145), where it is stated in so many words that the goals for sk and ak are the same, it becomes of great interest to investigate if one and the same teaching procedure can function in both courses.

In GUME 5 the lesson series consisted of 6 lessons. The syntactic structure taught in GUME 5 is the passive voice, and the lesson material is to a certain extent identical with the material of GUME 3. However, since a higher form (grade 8) was chosen for the GUME 5 experiment, the grammatical content as well as the lesson material was enlarged. As in the case of GUME 4, no attempt was made to equalize the length of the Ee and Es explanations, but the two treatments differed only by two minutes in total (see report GUME 5, p. 38).

The teaching conditions were similar to those of GUME 4. That is to say, the pupils listened to tape-recorded lessons without the use of earphones, and the teachers were to lead chorus reading with gestures and give the right answers in the free conversation exercises.

In both GUME 4 and GUME 5 it was originally planned that the criterion test should be administered as a retest at the beginning of the following term, i.e. when the pupils were just starting grade 7 and 9 respectively. However, for the results to be interpretable it would have been necessary to control the teachers for an unduly long period of time, preventing them from teaching the structures dealt with in the project. Since it was considered unrealistic to control the teaching process thus, the retention test was dropped.

The GUME A project was undertaken at the Gothenburg Municipal School for Adults (Göteborgs vuxengymnasium), and the experiment proper was performed during a two-month period towards the end of the 1970 autumn term (see fig. 3, p. 92). The experimental sample consists of the entire adult group taking the 7th grade course that term. The subjects' background in English, although it varied with respect to years of formal training, was such as to warrant placement of the individual subject in grade 7. Only two teaching approaches were compared, namely Im and Es. It should be noted that the Es method more closely approximates a traditional grammar-translation method than do any of the explanation methods of the previous experiments. (see p. 125).

Two teachers developed preliminary versions of teaching materials for adults (one teacher made the Im lessons, the other, the Es lessons). These lessons were tried out in connection with a pilot study during the spring term of 1970. The revised versions, which were used in the actual experiment, consist of ten 40-minute lessons each.

At the beginning of the autumn term of 1970 the entire adult group taking English (the 7th grade course) was organized into six groups. During this term three of the groups were taught by one of the teachers mentioned above, and three of the groups were taught by the other. (In order to facilitate development of materials and to achieve control over the experimental conditions the two teachers, though not members of the regular staff, took teaching positions at the Gothenburg Municipal School for Adults for the whole year of 1970).

Before the experiment started and the pretest was given a written diagnostic test and a listening comprehension test (PACT) were administered to the six groups. Analyses of variance revealed that there were no significant differences between the groups (detailed information is presented in a GUME report by von Elek & Oskarsson, 1972). During the experiment the teacher who had constructed the Im materials administered the Es lessons, and vice versa. The lessons, presented by tape recorder, were accompanied by projector transparencies and printed teaching materials. The role of the two teachers was simply to start the tape recorder and to operate the projector.

Permission to perform the experiment was obtained only on the assumption that the regular course in English could be given simultaneously. In order to prevent the contents of the regular lessons interfering with the experiment, the two teachers prepared a separate series of ("regular") lessons which was strictly adhered to by both. This course, which followed the ordinary textbook, was modified somewhat so as to avoid the syntactic problems treated in the experimental course. These were the some/any dichotomy, preposition + ing-form, possessive pronouns, the distinction between adjectives and adverbs, and the passive voice.

Although it would have been desirable to administer a test of retention the following term this proved impossible for administrative reasons, the subjects then being reorganized in a large number of groups.

The experimental schedule was very similar from project to project. The typical procedure was as follows:

(1) IQ testing

(2) Listening comprehension test (in one case, GUME A, a written diagnostic test in English)

(3) Pretest

(4) Introductory lesson explaining experimental aims, procedures, drill techniques, etc.

(5) The lesson series administered (the experiment proper)

(6) Posttest

(7) Pupil and teacher attitude tests

- -

(8) Re-test (GUME 1-3 only).

In some cases the IQ test or the listening comprehension test had to be administered somewhat later in the experimental sequence (the reader is referred to the part reports); these alterations cannot be supposed to have influenced the results.

The listening comprehension test mentioned in (2) above was in all part projects a variant of the so-called PACT test (Pictorial Auditory Comprehension Test). It is intended to measure foreigners' comprenhension of spoken English and was originally developed by John B. Carroll. In GUME 1–3 mimeographed copies of the original version were used by kind permission of Dr. Carroll. In the remaining studies, however, new versions were worked out, although with the original testing technique preserved. The pupils listen to a taped conversation or description of an object or event, etc., and then mark which of four alternatives (in the form of pictures) correspond to what was said on the tape. The test was included in the experiments as a potential covariate in the analyses of covariance. It should be mentioned in this connection that a secondary project objective has been development of foreign-language tests (see fig. 3, p. 92). Although auditory tests have been available in Swedish schools, none has been uncontaminated by reading ability (the options have mostly consisted of *written* alternatives). PACT has been further developed within the project and is included in the national test which will be administered to the student population in May this year. In Appendix 10 the testing technique is illustrated by an example.

The following table gives, in concentrated form, various characteristics of the different part projects. The numbers of subjects refer to the individuals for whom complete data are available, i.e. the observations that the treatment comparisons are based on.

Table 1. Survey of various features of the 10 experimental groups.

Part project	Grade	Appr. age level	N of classes (groups)	Total N of subjects	N of classes (groups) in each treatment			N of lessons per treatment
					Im	Ec	Es	
GUME 1 sk	7	14	12	227	4	4	4	6
GUME 1 ak	7	14	6	104	2	2	2	6
GUME 2 sk	7	14	12	247	4	4	4	6
GUME 2 ak	7	14	6	98	2	2	2	6
GUME 3 sk	7	14	12	170	4	4	4	6
GUME 3 ak	7	14	6	57	2	2	2	6
GUME 4	6	13	27	577	9	9	9	12
GUME 5 sk	8	15	12	235	4	4	4	6
GUME 5 ak	8	15	12	152	4	4	4	6
GUME A	7	adults	6	125	3	–	3	10

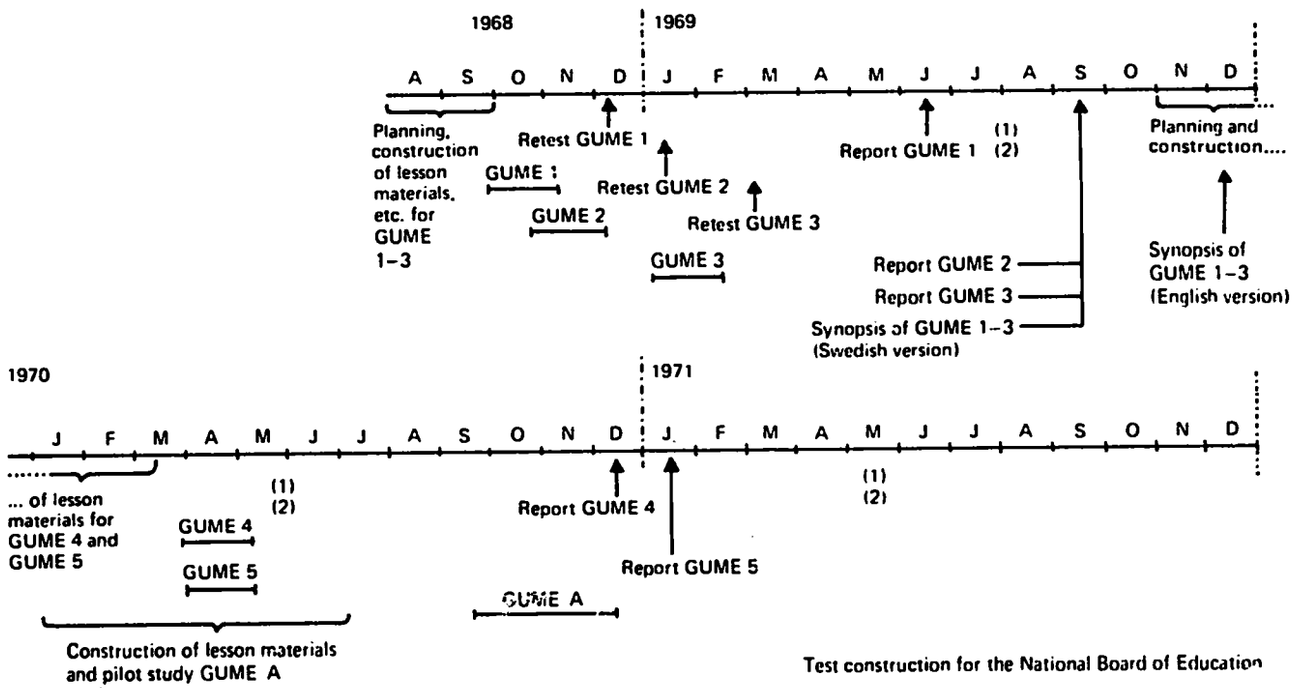


Fig. 3. Survey of the GUME research activities 1968–1971

Fig. 3 gives a survey of the research activities during the years 1968-1971. The duration and position (in time) of the various part projects is indicated by horizontal lines, the time of publication of the different part reports is indicated by vertical arrows. At one point a clarification is necessary: the figures (1) and (2), appearing in three positions, indicate that the criterion tests used in GUME 1 and GUME 2 respectively were administered in a number of control classes at three different times. The purpose was to find out to what extent the structures taught during the GUME experiments are actually learnt in one or two years' time without the teachers' paying special attention to those structures. The results will be presented in chapter 12.

CHAPTER 8

CHARACTERISTICS OF THE TEN EXPERIMENTAL GROUPS

Distribution of classes among schools

This description is relevant only with respect to GUME 1-5; GUME A utilized the entire population at one school only. In order to minimize between-school variance it would have been desirable to block on schools, i.e. ideally each treatment should be represented in each school. However, this procedure proved impossible for various reasons. First, many schools have English at exactly the same time; in the case of GUME 1-3 the mini-labs installed (see p. 87) led to overhearing in certain cases which made possible the use of only one or two classes simultaneously. Secondly, the requirement that a specific textbook should have been used previously necessitated the exclusion of some classes in the case of GUME 4. Thirdly, in GUME 5 it proved difficult to recruit classes for the experiment (see p. 88), and the total number of classes simply represent those where the teachers were willing to participate.

Table 2 illustrates the distributions of classes among schools in the case of

Table 2. Survey of distribution of classes among schools in GUME 1-5.

	GUME 1			GUME 2			GUME 3			GUME 4			GUME 5		
	Im	Ee	Es	Im	Ee	Es	Im	Ee	Es	Im	Ee	Es	Im	Ee	Es
School No. 1	x	x	x	x	x	x	x	x	x	o	o	o	x	x	x
School No. 2		x	x		x	x		x	x	o	o	o	x	x	
School No. 3	x	●		x		x	x	x		o	o	o		x	x●
School No. 4	●	x		x	●		x	x		o	o	o	●	●	x●
School No. 5	x		●		x	●	x		x	o	o	o	x●	●	
School No. 6	●		x			x			x	o	o		x		
School No. 7		●	●	x					●		o	o		x	
School No. 8		x			x		●			o	o				x
School No. 9			x	●			●				o	o	●		●
School No. 10	x			●				●		o				●	
School No. 11						●		●		o			●		
School No. 12					●				●			o		●	
School No. 13												o			●

x = sk classes ● = ak classes o = non-streamed classes

GUME 1-5. It should be noted that school No. 1, 2, 3, etc., are not identical for the different part projects.

The blocking procedure was thus possible in 5 schools in the case of GUME 4; in the other projects, in one school only. Each part-project thus utilized between 10 and 13 schools in all. We shall presently investigate if the variation due to differences between schools produced any significant differences between the treatments groups.

Drop-out rates

In educational experiments where the treatments are applied at successive time intervals, a certain drop-out rate is inevitable. The longer the duration of the experiment, the more severe the cumulative effect of the average rate of absence (it is not necessarily the same pupils who are absent from time to time). If no absence were to be allowed in order to include an individual in the treatment comparisons, a dramatic loss of data would usually occur in experiments of this kind. In the GUME projects the following subjective criteria for cancelling a pupil from the statistical computations were applied: In the experiments of relatively short duration (GUME 1-3 & 5, six lessons) the pupils absent from 2 or more lessons were excluded, in the remaining two experiments (GUME 4 & GUME A, 12 and 10 lessons respectively) the pupils absent from 3 lessons or more were cancelled. In addition to this, pupils for whom pretest as well as posttest scores were not available were excluded from the analyses. The unavoidable loss of individuals is a potential cause of bias in two respects, namely (a) the comparability of the treatment groups, and (b) the representativity of the experimental sample. We shall discuss our data in relation to both these problems in the present chapter. Table 3 on the following page illustrates the magnitude of missing data in the various samples.
In the case of GUME 3, both sk and ak, the drop-out rate is particularly high. In contrast to GUME 1 and GUME 2, which were performed under otherwise similar conditions, GUME 3 had a criterion test which took *two* hours, separated in time, to administer; since this implies four testing periods instead of two, the risk of absence increases.

On an average, ak groups tend to have a higher frequency of absence than the sk groups, which is according to expectations.

The pupils excluded because of too high absence (according to the criteria previously mentioned) were compared with their respective original populations in order to find out if the absence might have been selective. That is to say, did the absence-groups deviate from the original groups with respect to background variables? In two of the part projects, GUME 3, sk and ak, this

Table 3. Number of subjects included in the treatment comparisons.

	Original N of subjects (a)	N of subjects included in the analyses (b)	(b) in % of (a)
GUME 1 sk	299	227	75.9
GUME 1 ak	134	104	77.6
GUME 2 sk	309	247	79.9
GUME 2 ak	142	98	69.0
GUME 3 sk	283	170	60.1
GUME 3 ak	127	57	44.9
GUME 4	685	577	84.2
GUME 5 sk	297	235	79.1
GUME 5 ak	222	152	68.5
GUME A	141	125	88.7

was the case: the absence-groups proved to have significantly lower IQ scores, grades, and listening comprehension test scores. Detailed information about the drop-outs is presented in one of the earlier GUME reports (Levin 1969, p. 44). One can not completely rule out the hypothesis that the teaching procedures of GUME 3 have created lack of motivation in the less talented pupils and thus caused them to play truant in certain cases; the high frequency of absence for the GUME 3 groups (in table 3) lends support to this hypothesis.

During the experiments it was judged impossible to administer lessons to absent pupils on a later occasion; this would simply be asking too much of the teachers. Considering this and our criteria for including an individual in the data processing, the number of available observations is, with the exception of GUME 3, surprisingly high. Henceforth only the individuals included in the treatment comparisons (i.e. column b in table 3) will be dealt with.

Assignment to treatments

At the planning stage of each experiment a request for participation was sent to a large number of teachers. In cases where a surplus of positive answers was obtained, the final choice of classes was based on various criteria, some of which have been mentioned previously; the experience of the teacher, the boys/girls ratio, the textbook used, schedule and overhearing considerations. The final number of classes thus obtained was randomly distributed among

treatments, though with one restriction: in no school were two classes within the same course (sk or ak) allowed to receive the same treatment. In two cases (see table 2 p. 94), namely GUME 5, schools nos 3 and 4, one sk and one ak class were both assigned to Es. However, the risk that the pupils in the two courses would communicate in matters relevant to the experiment was judged negligible. (The randomization does not apply to GUME A; see p. 89)

Since the assignment to treatments was made at the school class level, there is a certain possibility that the actual number of subjects will vary from treatment to treatment depending on varying class sizes. This variation may be increased or levelled out because of drops-outs. The following table illustrates the final number of pupils per treatment included in the statistical analyses. In each case a X^2-test was applied to test if the observed values deviated from even distribution. The X^2-values are given in the table below.

Table 4. Number of pupils per treatment in each of the GUME 1-5 projects

	N					
	Im	Ee	Es	Tot.	X^2	
GUME 1 sk	69	77	81	227	0.97	
GUME 1 ak	23	42	39	104	6.25	$p < .05$
GUME 2 sk	84	92	71	247	2.71	
GUME 2 ak	38	38	22	98	4.65	
GUME 3 sk	50	63	57	170	1.51	
GUME 3 ak	16	20	21	57	0.73	
GUME 4	180	194	200	574	1.22	
GUME 5 sk	70	92	73	235	3.86	
GUME 5 ak	50	49	53	152	0.28	

The pupils are evenly distributed among treatments in all cases but one, GUME 1 ak (the critical X^2-value for df = 2 is 5.99).

Representativity of the GUME 1-5 samples

In order for the results to be generalizable, it must be shown that the experimental groups are random samples of their respective populations. When intact groups are sampled, it is usually difficult to provide a rational basis for assuming this. Furthermore, in eight of our ten experiments the groups con-

sist of samples of subpopulations, sk and ak, the characteristics of which are largely unknown. In these cases we shall lump together the sk and ak samples and discuss the representativity of the entire group in relation to the population at large.

An interesting aspect of the representativity problem would be to state that no given populations exist since the sk/ak proportions depend on the pupils' – and parents' – free choice. The actual sk and ak percentages may thus be viewed as partly a matter of chance. If so, there are no fixed populations to which our results might be generalized, and the problem of external validity would be of little importance. However, the sk/ak proportion, though theoretically flexible, appears to be relatively stable from year to year, and we have felt a need to let our samples reflect this proportion. Although it may imply overstressing the representativity aspect, we shall investigate the relation between our samples and their respective "populations".

At the end of this section we shall return to the representativity question in the case of GUME A.

THE SK/AK DISTRIBUTION.

The number of classes sampled for sk and ak were chosen so as to reflect the actual distribution of sk/ak pupils in the population. In table 5 below each part project will be compared with the population values for the year of 1969, which are based on more than 92.000 pupils all over Sweden. The official statistics (the population values) refer to grade 8; it matters little whether one uses grade 7 or grade 8 values or which year one chooses for the comparison, since the figures are relatively stable from year to year. In the

Table 5. The distribution of sk/ak pupils within GUME 1-5 and the population.

		GUME 1	GUME 2	GUME 3	GUME 4	GUME 5	THE POPULATION
sk	N	227	247	170	432	235	66,443
	%	68.6	71.6	74.9	75.3	60.7	71.8
ak	N	104	98	57	142	152	26,128
	%	31.4	28.4	25.1	24.7	39.3	28.2

case of GUME 4, where the pupils are not yet divided into the two courses, we have used the *choices* made by the pupils with respect to grade 7. These values may reflect a certain bias towards sk, assuming that this course has higher status value. Fig. 4 gives a visual impression of the values contained in table 5.

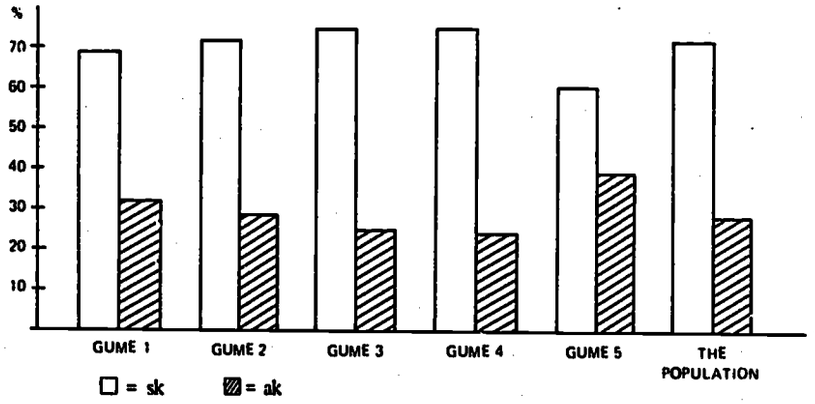


Fig. 4. The distribution of sk/ak pupils within GUME 1-5 and the population (percentages).

The correspondence to the population values was investigated in each case by a chi-square test. The X^2-values for the part projects are respectively 1.81, 0.01, 1.07, 3.44, 23.61 (df = 1); thus the sk/ak relations in GUME 1-4 are in accordance with the expected values at the 5 % level, whereas GUME 5 deviates strongly. In GUME 5 a relatively large number of ak classes were included in order to counteract the circumstance that ak classes ordinarily contain few pupils. It is obvious that this strategy disturbed the sk/ak relation.

SOCIAL CLASS

For GUME 1-5 information about the parents' occupation was collected at the headmasters' offices. In a number of cases either no information was obtained or the pupil's mother was given as the guardian without any mention of profession. In cases where information was available, the assignment of pupils to social class is based on a hierarchical description of professions and occupations from 1958 (1958 års Valstatistik). The criteria used in this publication are to some extent arbitrary and even inconsistent, but it was the

only source available at the time of our investigations. (A more consistent system has recently been developed for Swedish conditions; see Svensson, 1971.) According to offical statistics for the Gothenburg area (Andrakammarvalet i Göteborg 1968, U 1969:2, pp. 63-69) the overall figures for social group in Gothenburg are: 1: 8.2 %, 38.4 %, 3: 53.4 %. The distributions within each GUME project are given in table 6 below (group 1 corresponds roughly to "upper middle class", group 2 to "lower middle class", and group 3 to "working class").

Table 6. Distribution according to social class within each sample (GUME 1-5)

	No information	Social class No. 1	2	3	N
GUME 1 sk	19	80	76	52	227
GUME 1 ak	6	3	31	64	104
GUME 1 tot.	25	83	107	116	331
	%:	27.1	35.0	37.9	100.0
GUME 2 sk	45	29	77	96	247
GUME 2 ak	18	1	21	58	98
GUME 2 tot.	63	30	98	154	345
	%:	10.6	34.8	54.6	100.0
GUME 3 sk	9	27	70	64	170
GUME 3 ak	8	1	19	29	57
GUME 3 tot.	17	28	89	93	227
	%:	13.3	42.4	44.3	100.0
GUME 4	80	41	204	252	577
	%:	8.3	41.0	50.7	100.0
GUME 5 sk	56	41	66	72	235
GUME 5 ak	12	2	39	99	152
GUME 5 tot.	68	43	105	171	387
	%:	13.5	32.9	53.6	100.0
"The Norm" %:		8.2	38.4	53.4	

Fig. 5 below is a graphic representation of the distribution according to social class within the entire GUME 1–5 samples (sk + ak) and the Gothenburg population.

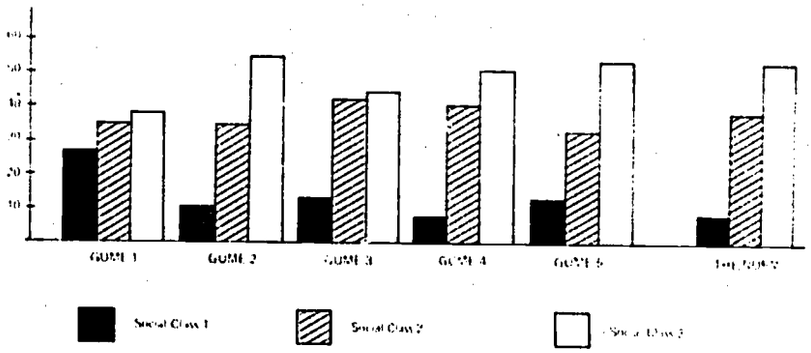


Fig. 5. Distribution according to social class within each GUME project (GUME 1-5) and the Gothenburg population.

In each case the sample distributions were compared with the population values by a chi-square test. The X^2-values are respectively 77.58, 3.12, 11.13, 1.52, 13.75 (df = 2). Thus only GUME 2 and GUME 4 are in accordance with the population values (within the limits of random errors); GUME 1 ($p < .001$). GUME 3 ($p < .01$), and GUME 5 ($p < .01$) deviate strongly. In the case of GUME 1 the deviation is largely due to the fact that three classes from a private school were included, in GUME 5 the deviation is somewhat surprising since that sample was biassed towards a surplus of ak classes (in table 6 the strong relationship between social class and course affiliation is apparent).

GENERAL APTITUDE (DBA) SCORES

Three parts of the so-called DBA-test (DBA = differentiell begåvningsanalys, i.e. differential intelligence analysis), namely the verbal, inductive, and spatial parts, were administered to each sample. In actual practice, the test is used mainly as an aid in vocational guidance. The three subtests, taken together, are considered to be a reliable measure of general ability or scholastic aptitude. The tests were administered in the following order: Verbal (10 min), Inductive (15 min), Spatial (12 min). The pupils' scores are expressed in stanine points; the total score is the unweighted sum of the three stanine scores. In table 7 the characteristics of the various experimental groups with respect to the part tests and the total are given. It should be observed that it was not possible to obtain test results for the complete samples; in the percentage column is indicated how large a proportion of each group did take the DBA tests.

Table 7. Means and standard deviations for the GUME 1-5 groups on the DBA test.

	N	% of entire group	DBA verbal $\bar{x}$	DBA verbal s	DBA inductive $\bar{x}$	DBA inductive s	DBA spatial $\bar{x}$	DBA spatial s	DBA total $\bar{x}$	DBA total s
GUME 1 sk	215	94.7	5.74	1.63	5.71	1.72	5.42	1.86	16.89	3.77
GUME 1 ak	96	92.3	3.52	1.37	3.70	1.78	4.12	1.82	11.23	3.37
GUME 1 tot	311		5.05	1.86	5.09	1.97	5.02	1.94	15.14	4.49
GUME 2 sk	230	93.1	5.47	1.52	5.56	1.94	5.23	1.91	16.31	4.06
GUME 2 ak	90	91.8	3.70	1.53	3.72	1.61	4.07	1.65	11.49	3.55
GUME 2 tot	320		4.97	1.72	5.04	2.03	4.90	1.91	14.95	4.48
GUME 3 sk	155	91.1	5.58	1.76	5.76	1.73	5.34	1.90	16.59	4.07
GUME 3 ak	54	94.7	3.56	1.60	3.00	1.30	3.93	1.95	10.48	2.93
GUME 3 tot	209		5.06	1.93	5.05	2.03	4.98	2.01	15.01	4.65
GUME 4	564	98.3	5.30	1.79	5.79	1.93	5.56	1.97	16.66	4.33
GUME 5 sk	214	91.1	5.87	1.63	5.89	1.73	5.59	2.08	17.37	3.94
GUME 5 ak	120	78.9	3.67	1.19	3.81	1.73	4.54	1.89	12.02	3.56
GUME 5 tot	334		5.08	1.82	5.14	2.00	5.21	2.08	15.45	4.59

DBA scores were obtained for more than 90 % of each sample with the exception of GUME 5 ak. With this exception the values may be regarded as representative of the entire groups. In GUME 5 the relatively many ak pupils in relation to sk pupils has the effect of decreasing the means somewhat.

In all cases except GUME 4 the samples do not deviate significantly from the population parameters. GUME 4 is above the norm as far as general scholastic aptitude is concerned. However, a new standardization of the DBA test in grade 6 has shown that the old norms are outdated (Härnqvist, 1969). For the part tests utilized in our investigations, a clear increase in test scores was noticeable in relation to the old norms. Thus the figures in table 7 overestimate the bias of the GUME 4 sample. In all cases the sk groups are significantly above the ak groups, which is a previouly well-attested fact.

GRADES

Grades in English, Swedish, and Mathematics were collected for each individual. As can be seen in table 8 below grades were obtained for practically the whole samples; the few observations missing cannot be supposed to influence the group means.

In the case of GUME 1-3 the grades were given at the end of the term preceding the experiment, i.e. at the end of the 6th form. At that time the pupils did not take separate courses, which means that they constituted one single reference group as far as grades are concerned.

The grades for the GUME 4 sample may reflect some subjectivity since they were given before the standardized achievement tests had been administered; the grades had thus not been adjusted according to the standardized tests.

Finally, the grades for the GUME 5 sample were given during the term preceding the experiment, i.e. the first term of grade 8. In Swedish the sk and ak groups take the same course and make up one reference group, whereas in English and Maths there are two courses. As it appeared, a high correlation exists between choice of course (advanced/easy) in English and Maths; thus the pupils in our sk group take the sk course in Maths in most cases. However, when this is not the case, the Maths grade is adjusted downwards by one point. Correspondingly, an ak pupil (in English) who is following the sk Maths course, gets his Maths grade adjusted upwards by one point. The intention behind this somewhat subjective procedure, which was applied in the limited number of cases where it was necessary, is to equate the grades in English and Maths.

Table 8. Grade scores (means and standard deviations) for GUME 1-5.

		% of total group	Grades English		Grades Swedish		Grades Maths		Grades Total	
			$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
GUME 1 sk	225	99.1	3.68	0.84	3.65	0.84	3.57	0.88	10.92	2.18
GUME 1 ak	103	99.0	2.12	0.74	2.20	0.68	2.13	0.74	6.45	1.79
GUME 1 tot	328		3.19	1.09	3.19	1.04	3.12	1.07	9.52	2.93
GUME 2 sk	239	96.8	3.53	0.87	3.50	0.81	3.48	0.94	10.51	2.26
GUME 2 ak	94	95.9	2.23	0.72	2.37	0.66	2.31	0.76	6.91	1.73
GUME 2 tot	333		3.16	1.02	3.18	0.92	3.15	1.04	9.49	2.67
GUME 3 sk	168	98.8	3.63	0.81	3.58	0.88	3.53	0.94	10.74	2.25
GUME 3 ak	56	98.2	2.14	0.75	2.30	0.74	2.38	0.80	6.82	1.82
GUME 3 tot	224		3.26	1.02	3.26	1.01	3.24	1.03	9.76	2.74
GUME 4	570	98.8	3.09	1.03	3.15	0.92	3.08	0.97	9.32	2.58
GUME 5 sk	233	99.1	3.33	1.00	3.47	0.84	3.21	1.06	10.02	2.45
GUME 5 ak	148	97.4	2.86	0.86	2.30	0.60	2.66	0.89	7.82	1.84
GUME 5 tot	381		3.15	0.98	3.02	0.95	3.00	1.03	9.17	2.48

In each case the unweighted grade scores have been added to a total. The relation between the various sample means and the population paramenters (theoretical means for the separate school subjects and the total are 3.00 and 9.00 respectively) will be investigated. However, in the case of GUME 5, where it is illogical to add the three scores to a total, only grades in Swedish will be considered.

In practically all cases the means exceed the expected average of 3.00. The samples thus appear to be positively biassed as far as grades are concerned. However, one can not rule out the hypothesis that the high grades are a sign of the teachers' generosity. This generosity effect has been demonstrated earlier by Marklund (1960, p. 172) and is also noticeable in Svensson's (1971, p. 53) large-sample data. Further support for the generosity hypothesis is provided by the fact that the groups, with the exception of GUME 4, were shown not to deviate from the norm in the case of general scholastic aptitude. Since no information is available about the grade means in the population, the representativity problem in this respect is impossible to solve. However, considering the facts mentioned above, we shall regard our results as indicative of existing grading practices. In GUME 5, where it is only permissible to consider grades in Swedish, i.e. the only school subject where the pupils constituted a single reference group, the sample has somewhat lower grades because of the surplus of ak classes.

The following table is an attempt to summarize the dicussion on the representativity of the GUME 1-5 samples. A (+) sign indicates that the sample is in accordance with a certain norm or population parameter, whereas a (–) sign indicates a deviation.

Table 9. Survey of the representativity of the GUME 1-5 samples in various background variables.

	sk/ak proportion	social class	DBA	grades
GUME 1	+	–	+	+
GUME 2	+	+	+	+
GUME 3	+	–	+	+
GUME 4	+	+	–	+
GUME 5		–	+	+

The values indicate that our samples, with the exception of GUME 2, are not strictly representative of their respective populations in the variables investigated. Caution must be observed in generalizing the results. It should

be noted, though, that the above description concerns the samples as a whole, i.e. including both sk and ak. In the teaching method comparisons the sk and ak samples will be treated separately.

Representativity of the GUME A sample.

As was stated earlier (p. 89) the GUME A group consists of the entire adult group taking English (the 7th grade course) at the Gothenburg Municipal School for Adults at the beginning of the school year of 1970. Although similar groups exist elsewhere, above all in large cities, it is difficult to hypothesize a population, of which our group might be considered a sample. Similar groups are very heterogeneous in practically all background variables (age, IQ, previous schooling, amount of time devoted to studies, particular courses taken, etc.) and furthermore, the groups vary in these respects from term to term. In the present investigation the background variable of major importance is the age factor. Thus, in order to give a conception of the kind of adult population investigated, we shall present some charateristics of our group.

The age of the group varies from 17 to 60 with a mean of 33 years. The actual distribution is as follows (N = 125):

-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60
4	25	29	25	22	7	5	4	4

In a survey of adult students taking similar courses, Johansson & Molander (1970) report a median age of 26.90 years. The authors also report that women are in a majority in adult courses of this kind. The GUME A sample contains 83 women and 42 men.

The educational background of the members is fairly homogeneous. With two exceptions ("realskola") none has any academic training beyond the compulsory school. The individuals' formal training in English varies from 0 to 3 years. However, those who have no formal training possess sufficient knowledge to be able to follow the grade 7 course.

Test scores on the verbal part of the so-called F-test are available for 111 individuals (88.8 % of the sample). The mean of the group proved to be 51.40, which roughly corresponds to the median score for the various adult norm groups presented in the test manual.

Adult students are free to decide what number of courses they should take during the same term. The distribution of school subjects read simultaneously is as follows:

	Subjects read					
	one	two	three	four	No inform.	Total
N	63	20	32	7	3	125
%	50.4	16.0	25.6	5.6	2.4	100.0

These figures are probably related to the amount of time that the individuals have at their disposal for studies. The main occupation by the individuals is reflected in the following survey:

	Working full-time	No employment full-time studies	Working part-time	House wives	No inform.	Total
N	57	8	16	37	7	125
%	45.6	6.4	12.8	29.6	5.6	100.0

Although the GUME A sample is heterogeneous in all variables investigated, the following generalizations are warranted:

(a) it consist of "adults"
(b) the members have no academic training beyond the compulsory school level
(c) their previous knowledge of English corresponds to a proficiency level normally reached at the beginning of grade 7
(d) the majority of the group have occupational duties and devote only a relatively limited time to studies

It is worth pointing out that most investigations at the adult level seem to have been concerned with college students. The discrepancy between select groups of this kind and the GUME A sample should be apparent from the above description.

Characteristics of the treatment groups (GUME 1-5)

As was stated earlier (p. 79) the pretest scores will continuously be used as covariates in our analyses of covariance. It is important to compare the pre-

test means of the various treatment groups, since in cases where the treatment groups differ significantly from each other, the comparison of adjusted means will have low precision. The pretest means were compared by analysis of variance; table 10 below presents the details. Each analysis was preceded by a comparison of the variances by the Bartlett test for homogeneity of variance. In no case was a significant difference between the variances found.

Table 10. Analysis of variance (one-way) of Pretest scores for GUME 1-5

	Im	Ee	Es	F	Sum of squares between	Sum of squares within	df	
GUME 1 sk	70.87	70.39	71.90	0.177	94	59436	2/224	
GUME 1 ak	48.17	46.71	50.21	0.815	247	15320	2/101	
GUME 2 sk	64.32	66.58	62.96	0.933	550	71912	2/244	
GUME 2 ak	48.05	47.18	43.82	0.652	260	18997	2/95	
GUME 3 sk	86.92	82.32	89.53	3.485	1604	38432	2/167	$p<.05$
GUME 3 ak	63.38	64.40	68.00	1.042	228	5895	2/54	
GUME 4	48.83	53.14	52.28	2.259	1927	243489	2/571	
GUME 5 sk	60.69	60.22	56.19	2.084	908	50546	2/232	
GUME 5 ak	33.12	30.80	30.68	1.930	191	7381	2/149	

There is one significant F-ratio in the table, namely for GUME 3 sk (Es > Im > Ee). In this case it can be supposed that only large differences between treatments will be detected. A safeguard in such cases is to compute confidence limits for some of the differences; if the F-test alone is made, this point can easily be overlooked (cf Snedecor & Cochran 1967, p. 430). It thus appears that the sampling procedure – or the loss of individuals – has disturbed the comparability between the treatment groups in the case of GUME 3 sk.

Similar comparisons between treatment means in respect of other background variables are not equally important for the interpretation of the analyses of covariance. However, as complementary information on the characteristics of the treatment groups we shall present the comparisons between them with respect to DBA and Grades means.

In each part project the analysis of variance was preceded by a test for homogeneity of variance. In the case of Grades the variances in all experiments are homogeneous according to the Bartlett test; with respect to DBA scores the variances in two projects, GUME 1 sk and GUME 2 ak, deviate from equality, the B characteristic being 6.38 and 6.69 respectively (as compared to the critical value 5.90). In the following two tables the comparisons of group means for DBA and Grades are given.

Table 11. Analysis of variance (one-way) of DBA scores for GUME 1-5.

	Im	Ee	Es	F	Sum of squares between	within	df
GUME 1 sk	16.58	16.39	17.65	2.411	68	2977	2/212
GUME 1 ak	10.85	11.50	11.14	0.265	6	1071	2/93
GUME 2 sk	16.29	16.56	15.98	0.367	12	3759	2/227
GUME 2 ak	11.29	11.71	11.43	0.122	3	1119	2/87
GUME 3 sk	16.82	16.38	16.64	0.153	5	2540	2/152
GUME 3 ak	11.00	9.65	10.95	1.293	22	434	2/51
GUME 4	16.44	16.92	16.59	0.613	23	10408	2/558
GUME 5 sk	18.28	16.82	17.11	2.732	83	3210	2/211
GUME 5 ak	11.79	11.28	13.08	2.683	66	1442	2/117

None of the F-ratios is significant. The results thus indicate that the treatment groups within each sample do not deviate from each other as far as general scholastic aptitude is concerned.

In table 12 below the results of the treatment group comparisons on Grades total are given. It should be observed that adding the three grade scores to a total is inadequate in the two GUME 5 samples; however, the procedure should not affect the *differences* between the treatment groups.

Table 12. Analysis of variance (one-way) of Grades total for GUME 1-5.

	Im	Ee	Es	F	Sum of squares between	within	df	
GUME 1 sk	10.97	10.56	11.20	1.703	16	1051	2/222	
GUME 1 ak	6.09	6.39	6.72	0.935	6	319	2/100	
GUME 2 sk	10.54	10.59	10.38	0.178	2	1214	2/236	
GUME 2 ak	6.77	6.97	7.05	0.199	1	278	2/91	
GUME 3 sk	10.90	10.44	10.91	0.827	8	834	2/165	
GUME 3 ak	7.44	5.85	7.30	5.125	30	153	2/53	$p<.01$
GUME 4	9.25	9.42	9.29	0.219	3	3781	2/567	
GUME 5 sk	10.06	10.17	9.78	0.535	6	1387	2/230	
GUME 5 ak	7.84	7.45	8.18	1.989	13	482	2/145	

In all cases but one (GUME 3 ak) there are no differences between the treatment groups as far as grades are concerned. In GUME 3 ak (Im = Es > Ee) the significant F-ratio probably indicates varying grading practices rather

than true differences between the groups; this hypothesis is supported by the fact that the groups did not differ in general aptitude and pretest scores.

Thus the overall picture is one of equality between the treatment groups in the variables investigated (pretest scores, DBA, and grades). Two deviations from this pattern were found: in GUME 3 sk (pretest scores) and GUME 3 ak (grades) the treatment groups are not strictly comparable.

Characteristics of the treatment groups (GUME A)

Since the background variables of GUME A are not identical with those of GUME 1-5 we shall, for convenience, present them separately. The following table gives the results of the comparisons between the Implicit and the Explicit group on some variables.

Table 13. t-values for differences between Im and Es in GUME A.

	Im			Es				
	N	$\bar{x}$	s	N	$\bar{x}$	s	t	
Pretest	57	56.56	18.32	68	53.18	13.57	1.15	
F-test verbal	48	51.27	10.19	63	51.49	8.49	0.12	
Diagn. Engl. test	57	31.00	10.13	68	30.54	8.94	0.27	
PACT	57	32.84	11.35	67	29.54	10.24	1.69	
Age	57	30.68	8.08	68	34.90	9.53	2.68	$p<.01$

There are no significant differences between the two groups in the various cognitive variables. The difference in age between the two groups cannot be supposed to influence the main treatment comparisons substantially since the correlation between the age variable on one hand and the posttest as well as the progress variable on the other is low (see Appendix 4, table X). However, we shall later investigate the relationship between the age factor and the dependent variable.

As stated previously the adult group contains 83 females and 42 males. The sexes are distributed between the two methods as follows:

Table 14. Distribution of sexes between treatments in GUME A.

	Im	Es	Total
Females	28	55	83
Males	29	13	42
Total	57	68	125

The frequencies deviate from even distribution, the X^2-value being 14.46 ($df = 1; p < .001$). Since the sexes are so unevenly distributed between the two methods, we found reason to compare them on the following background variables:

Table 15. t-values for differences between the females and males in GUME A.

	Females			Males				
	N	$\bar{x}$	s	N	$\bar{x}$	s	t	
Pretest	83	53.41	15.34	42	57.31	16.80	1.26	
F-test verbal	77	50.71	9.48	34	52.94	8.52	1.21	
Diagn. Engl. test	83	30.96	9.69	42	30.38	9.09	0.33	
PACT	82	29.78	10.58	42	33.55	11.06	1.82	
Age	83	35.04	9.29	42	28.90	7.04	4.12	$p < .01$

The females in the GUME A group appear to be around 6 years older than the males. In the cognitive variables no differences are found.

Thus, although no differences are detected in the cognitive background variables, either between the two methods or the sexes, the following observations can be made: (1) the Es group is somewhat older than the Im group (2) the females are somewhat older than the males (3) the Es groups contains a disproportionately large number of females. These observations, taken together, warrant an investigation of the sex *x* method interaction; this will be undertaken in chapter 11.

To summarize the findings in the present chapter it appears that the comparability between the various treatment groups, i.e. the internal validity of the experiments, is satisfactory. The following deviations from the general pattern of equality were found: GUME 3 sk (pretest: Es > Im > Ee), GUME 3 ak (grades: Im = Es > Ee), and GUME A, where the Es group contained comparatively older students and comparatively many females. Thus the group sampling procedure does not seem to have seriously disturbed the internal validity of the experiments. As regards the generalizability, i.e. the external validity, of our experiments, the following should be kept in mind: in the case of GUME 1-3 and 5 we will treat the different courses (sk/ak) separately in all calculations; however, we have tried to discuss the representativity of the whole groups (sk + ak) in relation to their respective populations, simply because population parameters are not available for the courses separately. In two background variables (DBA and grades) that correlate substantially with the dependent variable, the GUME 1-3 and 5 sample do not deviate from the norm. The GUME 4 sample, which scored relatively high on the DBA test, is such as to warrant generalizations to other large city groups. The GUME A poses a specific generalizability problem since it is difficult to

visualize a population of which our group may be considered a sample. We think the results in GUME A may be regarded as valid for adult groups possessing the general characteristics as described on p. 106. All in all we think that the internal validity of the investigations is satisfactory; however, caution must be observed in generalizing the results.

CHAPTER 9

THE LESSON SERIES

General considerations

The research presented in this report comprises a total of 128 lessons. It goes without saying that a detailed account of the teaching procedures of each lesson is out of the question. The problem which we are confronted with at the moment is to give the reader a description of the methods which enables him to form a judgment on the results and to estimate their educational relevance. Baker (1969), in reviewing a large bulk of evaluation research, deplores that many researchers do not specify the subject matter with which they are dealing:

> "Too often, the preoccupation with satisfying the requirements of design and statistical models violates the instructional treatment and reduces the utility of the research to zero" (p. 340).

This notorious deficiency of much research has been pointed out by several authors; Wittrock (1966) emphasized the necessity for specifying precisely the instructional variables, and Gagné (1967) stated that one cannot draw valid conclusions about differing *methods* of instruction unless there is an experimental way of controlling *content* (p. 36). The two last mentioned authors have further elaborated these views in a recent conference report (see Wittrock & Wiley, 1970).

Apart from the question of satisfactory specification of content there is still one of importance, namely the description of learning outcomes in relation to specified objectives. The more well-known techniques used for this purpose, such as Bloom's (1956) taxonomic approach, Gagné 's (1965) hierarchical descriptions of learning structures, and Stake's (1967) model in terms of antecedents, transactions and outcomes, seem to have achieved limited application within the field of second-language learning. The kind of taxonomy used in this area is normally one which identifies various linguistic elements (vocabulary, grammar, morphemes, pronunciation, realia) on the one hand and language skills (speaking, listening, writing, reading) on the other. Different varieties of this type of model exist side by side (see, for instance, Lado 1961, Valette 1968, Carroll 1968). By using an elements x skills matrix it is thus possible to identify learning outcomes.

In the GUME experiments we do not aspire to cover all outcomes inherent in such a matrix. The particular element chosen for study is grammar or, rather, a limited number of grammatical structures. On the other hand, all four skills have been included in the teaching procedures. It should thus be observed that our teaching methods include a limited number of possible outcomes and that they should not, therefore, be evaluated against a global foreign-language learning objective (cf p. 62).

All lessons were recorded on tape. One may ask whether this procedure, adopted in order to eliminate teacher variance, implies a greater handicap for any particular method. It might by hypothesized, for instance, that the Implicit method suffers most from tape-recording because there is no live teacher to reinforce the pupils during the drills, to increase or decrease the tempo as the situation demands, etc. However, it can similarly be argued that the Explicit methods have been curtailed most; the tape does not await the proper moment to explain or summarize, nor does it perceive whether or not an explanation has been grasped by the majority of the pupils, etc. We have discussed the problem concerned here with a number of experts and received conflicting answers. It thus seems to be a matter of subjective judgment which method is most hampered by tape-recording. It is our contention that tape-recording does not provide optimal conditions for either one of the methods. Although taped lessons and programmed materials may be of great value as complements to teacher-led instruction, it is difficult to conceive of a foreign-language teaching method completely bereft of the live teacher. As was stated previously taped lessons were adhered to as an experimental necessity.

In the following sections the teaching procedures will be described as fully as is feasible for space considerations. The reader is also referred to the separate reports (see Appendix 1) for close scrutiny; in some of the reports complete recording manuscripts for the explanations in the explicit groups are given.

GUME 1-3

As was mentioned earlier (p. 86) GUME 1-3 were planned and performed as a unity. The lessons consisted of three parts: an oral with structure drills, a written for written exercises, and a part for reading and listening practice, each taking roughly 10 minutes. The Im lessons were the starting point: the exercises were composed according to Im principles, i.e. there were no explanations at all. The explanations in the E groups were approximately 9 minutes per lesson which is close to 30 % of each teaching session. This is

more than would be considered optimal by most foreign-language teachers. However, it was judged necessary to give the explanations disproportionately long time in order to detect their effect, if any (cf p. 68). The explanations were divided into three 3-minute sections, one in each of the three parts of the lessons. The explanations were inserted in what was considered a suitable place in the exercise and a corresponding part of the exercise was excluded. A graph illustrating the organization of the GUME 1–3 lessons is given in fig. 6 below.

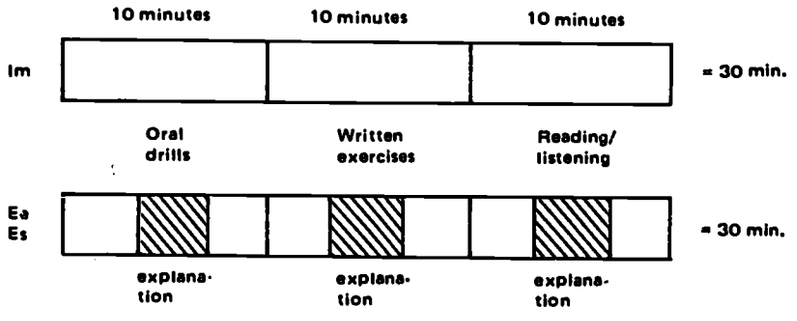


Fig. 6 General outline of the lesson sequence in GUME 1-3.

The teaching time of each lesson should optimally be the same between the teaching methods. Although there are minor variations between the methods in single lessons, the total teaching time is, for all practical purposes, the same; this holds for all GUME experiments.

ORAL DRILLS

Most drills are so-called four-phase drills (question – pupil's response – correct response – pupil's repetition of correct response). In some instances three-phase-drills are used, i.e. the pupils are not given time to repeat the right response. This is the case when, for instance, dialogues are converted into drills by simply letting the pupils act one of the parts. As far as possible contextualized drills have been used in order to avoid the use of isolated sentences. There is one difference between the part projects as far as stimuli presented during the drills are concerned. In GUME 1 the pupils saw a picture during the drill whereas, in GUME 2 and 3, they had some kind of written stimulus, usually the pattern practised, in front of them. In GUME 3 the

grammatical construction concerned, the passive voice, lends itself particularly well to transformation drills (active to passive and vice versa); the latter are accordingly frequent in GUME 3. In the case of GUME 2, where the some-any dichotomy is taught, it proved difficult to achieve contrastive drills where the pupil is called upon to use his built-in grammatical knowledge and to select the relevant item for his answer. This means that if a drill is used to illustrate the use of *some* in a particular context the pupils have to use *some* all through the exercise. This inevitably causes monotony to some extent; in relation to the other two part projects the drills in GUME 2 may be supposed to be somewhat less powerful. It should be noted that all the treatment groups did oral drills, except that the Im group did more and longer drills to make up for the time spent on explanations in the E groups.

WRITTEN EXERCISES

The purpose of the written exercises, or drills, was to consolidate what had already been taught during the oral drills. The written work was in most cases heavily structured so that the chances of mistakes were minimal. The procedure was as follows: The pupils were asked to look up a certain page in their workbooks (which had been specially made for the experiments), instructions as to how the drill should be done were given orally on the tape, one or two examples were done, and then the pupils were given a number of minutes to write. Sometimes they were allowed to go on and do as many pages as they had time for. After this the normal procedure was to read at least a number of the sentences in the correct form so that the pupils could correct their own attempts. Most of drills were of the fill-in type, simply in order to save time; if the pupils had been asked to write out whole sentences they would have spent an inordinately long time on things which, from the project point of view, would have been irrelevant. The written drills, being strongly structured, may be supposed to be of particular importance in the Im group. In the E groups, where the written drills are not so frequent, their role is taken over by the explanations.

THE READING TEXTS

Texts for the third part of the lesson, that is for the reading, were the same for all the three teaching strategies. Reading means here that the students had the texts in front of them and listened to a performance by native speakers. By this device the difficulty which the pronunciation would otherwise have presented, was avoided. Words which were presumed to be new to the stu-

dents were given in Swedish in the margin. This was felt not to interfere with the strict adherence to an implicit method, since this method is not a direct method in the sense that translations are forbidden; the term Im only refers to the teaching of grammatical structures and occasional translations of words and instructions are not part of the definition. The majority of the texts were written by two native Englishmen; a limited number of texts were written or adapted by members of the GUME project. The criteria for selecting the texts were (1) they should be fairly easy, interesting and deal with everyday situations, (2) they should introduce new grammatical content gradually, i.e. they should be carefully structured, (3) they should abound in representative examples demonstrating the grammatical structure concerned, thus providing continual repetition.

THE IMPLICIT METHOD

The Implicit variant is implicit in the extreme. There are no explanations of any kind. The stress is entirely on practice and the reasons for the various exercises are never overtly expressed. The Implicit method is thus without "grammar" unless we mean that the ordering and structuring of the various items constitute the grammar of the language.

THE EXPLANATIONS

In the explicit groups the pupils were given grammatical explanations, meant to direct their attention to the problem and to show them what they were doing in the exercises. It should be noted that the pupils were not given grammatical rules that had to be learnt or remembered, nor were they confronted with grammatical terminology. For instance, in GUME 3 the designations *subject*, *object* and *agent* were replaced by the figures 1, 2, and 3. This procedure made it possible to describe the transformation of an active sentence into the corresponding passive sentence by pointing out how part 3 in the active sentence moved to the beginning of the passive sentence, and how part 1, preceded by the "word" *by*, was placed at the end of the passive sentence. It was finally stressed how part 2, the verbal part, kept its place in the middle of the sentence.

The explanations used in GUME 1 deviate somewhat from those applied in the two other projects. In this study explanations of a somewhat unconventional kind, slightly influenced by transformational grammar (see Chomsky 1967, p. 420), were given. A question morpheme, represented by a question mark, was introduced, and a "free" s-morpheme was shown to move from

after the subject to a position before the subject; the term morheme was never used, however. Part of a teaching sequence is given in Appendix 12 as an illustration of the technique used. It should be observed that this pedagogical application was not intended by Chomsky (op.cit., p.407); however, since it is theoretically possible that learning would be facilitated by this procedure, it was thought worthwhile to investigate it.

In most Swedish school grammars the usage of *some/any* (the structures taught in GUME 2) is explained by reference to the sentence types in which they occur. Thus *any* is used in negative sentences and in questions while *some* is used in affirmative sentences. Then there are rules for why *some* is also used in questions and why *any* is used in statements when the basic meaning is negative, or with the meaning "vem/vilken/vad/som helst" in Swedish (cf Slettengren–Widén, 1966). To avoid this complexity *some/any* was treated as a semantic problem in GUME 2. *Any* means "any at all" (någon alls, någon överhuvudtaget, någon som helst), while *some* has a more specific and restricted meaning (någon viss, någon sorts, somliga). This distinction in hinted at in Löfgren (1950, p. 87) and treated more fully in Ellegård (1969, p. 42-45). With this type of analysis it is possible to treat the whole complex without the involvement of exceptions. On the other hand, it was not considered prudent to demolish the knowledge the pupils already might have. It was therefore repeatedly stated that *any* has the meaning "any at all" and that this meaning is particularly common in negative sentences and in questions.

Explanations are relatively easy to handle in the Es group where we have recourse to Swedish. In the Ee group we used the helpword "at all" consistently. A typical direction to the pupils in GUME 2 might thus run like this: "Use *any* in sentences where you can put in "at all" and where this gives a correct meaning".

In GUME 3 the explanations used are of a formal as well as a semantic character. Formal criteria are used when changes in the word-order in the transformation from active to passive are demonstrated. When it is pointed out that an active sentence has the same meaning as the corresponding passive sentence, semantic criteria are used. The explanations start with the active sentence as a kind of kernel sentence and describe how the passive is derived from it.

The pupils never saw the grammatical explanations in print. However, the particular structures were printed in the work sheets (Ee: green paper, Es: red paper), and during the explanations the pupils were, in a number of instances, asked to fill in missing words, to underline sentences, etc.

GUME 4

The following grammatical phenomena were practised during the lesson series: the s-form of the verb in the third person singular present (he gets up late); the present and past contiuous tenses in contrast to the simple present and past (he is playing the piano – he plays the violin, she was reading when he came in); preposition followed by an ing-form of the verb (he is good at dancing); the positior of adverbs of time (he is always late, he always comes home late); the some-any dichotomy, including something, somebody, anything, anybody; the do-construction in questions and negative sentences, both in the present and past tenses, and in all persons (does he like tea? – Yes, he likes tea very much, etc); and finally the regular past tense in -ed (he walked home).

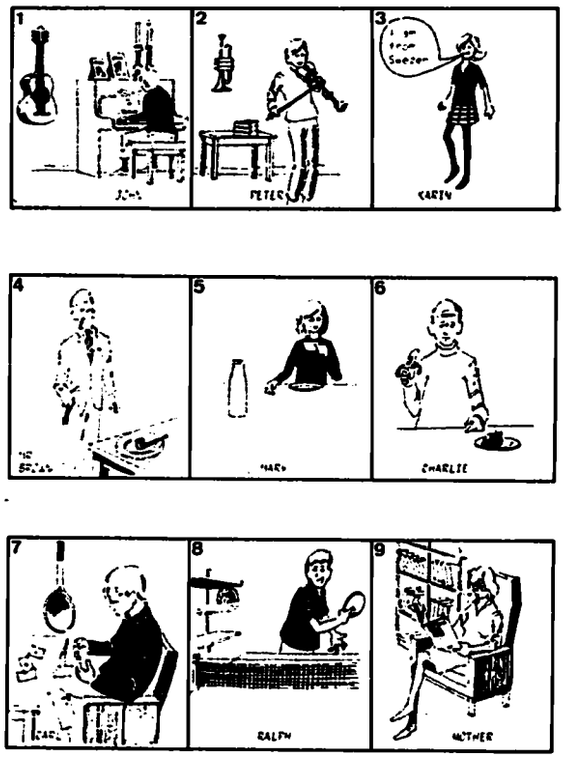
An attempt was made to vary the lessons as much as possible. Many different activities alternated: listening, oral drills with different stimuli, written exercises and reading. All four language skills were practised, but the main objective was the learning of the above-mentioned grammatical structures and the pupil's ability to use them; listening and reading, the passive skills, were thus of secondary importance and in speaking no kind of pronunciation control was introduced, and vocabulary learning did not occur except incidentally. Although the lessons outwardly resemble ordinary lessons in that they are varied and include practice in all four skills, they differ in that the goal is more limited (cf. p. 113).

In the case of GUME 4 we shall try to give an impression of the teaching procedures by presenting one lesson in some detail.

THE IMPLICIT METHOD (LESSON NO. 7):

First the pupils listened to chapter 3 of a story (which continued through five lessons) which contained a large number of examples of 'some' and 'any' and their compounds. The pupils had the text, one page, in front of them. A few questions were then asked on the text and the answers, most of which contained examples of 'some' or 'any' were given; the pupils were just listening. This first part, during which the pupils were silent (but hopefully not completely passive!) took just over 4 minutes.

Then the pupils were asked to turn to page 2 (see fig. 7 for a diminished copy of it). This is a mechanical drill of 'not anything' in the sense of 'nothing'. First the pupils listened to the whole dialogue and then they were asked to take over Bill's part. Normally drills of this kind were made as 4-phase: Tom's sentence is the stimulus, one pupil speaks Bill's part (the teacher points to a pupil who answers), the tape gives the right sentence, and



A Dull Day.

Tom and Bill come home. Life is dull. There's nothing they can do.

Tom: There's nothing on TV tonight.
Bill: There isn't anything on the radio either.
Tom: There's _____ on the table.
Bill: There can't _____ in the fridge either.
Tom: _____ to drink.
Bill: _____ to eat either.
Tom: _____ to smoke.
Bill: _____ to drink either.
Tom: _____ to read.
Bill: _____ to listen to either.
Tom: There's nothing to do.
Bill: But we must do something.

Figure 7. Two pages from the pupils' booklet (GUME 4)

then the whole class repeats this. Working with this page took about three minutes.

After this they were allowed to relax while they listened to a song, the text of which was given on page 3 of their booklets.

On page 4 the pupils practised 'any' in questions in a written drill. After a short introduction in Swedish they were given 4.5 minutes to write in. The teacher had an overhead copy of the page with the correct phrases in it. He put this on the overhead projector after 2 minutes, so that the pupils could correct what they had written as they got ready. The weakest pupils who might not have known what to write could copy the correct phrases, but experience showed that very few did that. When one minute remained soft piano music was played on the tape to warn the pupils that it was time to start correcting what they had written. Not all of them had time to write everything.

Next the pupils looked up the pictures on page 5 (see fig. 7). In all these pictures there is somebody doing something at the moment, but there is also something to indicate that at other times he or she does something else, e.g. in number 1 John is playing the piano, but on the wall is his guitar: "He plays the guitar very well." This is meant to practise the meaning of the simple present and the present continous.

First the pupils listened while the voices on the tape spoke about the pictures, next they were asked to repeat after the tape, and then they answered questions, like "Does John play the guitar?", "Is he playing the guitar now?", "What is he playing?"; for Swedish pupils, in whose language the difference between the simple and continuous tenses does not exist, the difference in meaning poses a greater problem than the forms. This exercise took a little over 12 minutes in all.

Finally they had pictures 4, 7 and 9 reproduced on page 6 in their booklets and were asked to write down answers to questions similar to those that they had answered orally before. They had 4 minutes to do this. They had an overhead key and music to warn them that time was up just as in the previous written exercise.

The total running time of this lesson was 31.5 minutes; this happens to be the shortest lesson of all.

THE EXPLICIT LESSONS

The comments given in the explicit groups were somtimes very short, like "When you write this, remember to have the 's' after 'he', but not after 'I' and not after 'they' ", sometimes very long, taking 4 or 5 minutes. In the latter case they were combined with written or oral practice, they were not

just long lectures on theoretical grammar but rather commented drills where the pupil was "taken by the hand". No pre-determined fixed time of explanations per lesson existed, as it did in our previous experiments. The explanations were meant to be "optimal", simply defined as the best we could produce for our purpose and taking as long as they had to. The explanations in Ee and Es were of almost equal length, even though this was not a fixed condition. There were between two and eight explanations in each lesson.

The most common procedure in GUME 4 was to have a short introduction either in the form of a few examples that the pupils just listened to, or in the form of a short drill, then came the explanation, and after that followed the main body of the drill. This seems to be slightly different from the common audio-lingual practice: "(the) generalization sets out in organized form what he *has been doing* in the drill' (Rivers, 1968, p. 43, italics ours). The Authorized Swedish Curriculum (Supplement English, p. 14) also recommends that generalizations – if they are to be given or formulated at all – should come in at the end as a confirmation. This might be a point worth investigating but it was not part of the present project, and we put in explanations at what was felt to be the best possible points.

The same structure was explained or commented on more than once, of course. Normally the first time was in the form of a short *eye-opener*, e.g. in lesson 10: "Now listeners, before you answer the questions I will tell you what we learn from these examples. After 'good at' we have the ing-form of the verb. So it's not enough to say 'sing' or 'swim' after 'good at'. We must say 'good at singing', 'good at swimming'." Then follows, sometimes after another short reminder, *the main explanation*, which often takes the form of a discussion, a dialogue between the voices on the tape, and with the pupils participating orally and by writing down certain phrases. Then, in a following lesson, there is a *reminder*, as in lesson 11: "So, listeners, here we are going to practise sentences where we say 'afraid of'. What form of the verb must we have after 'afraid of'? /// (Pause for the pupils to think and answer) – We must have the ing-form. – Yes, that's right. Listen, please. 'He is afraid of taking the medicine' And why do we have the ing-form? /// –Well, it's because of the little word 'of'. (etc)

LESSON "Ee 7"

In lesson 7, the implicit version of which was described in detail above, explanations in the explicit versions came in at the following places. The first very short comment came in just before the pupils listened to page 2; it took 25 seconds and it pointed out that "in this little exercise we practise 'anything' in sentences with 'not' ".

The next one came in just before they started writing on page 4 and it pointed out in the form of a dialogue between the voices on the tape that 'any, anything' are used in negative sentences and questions and 'some, something' in "other sentences". It took 39 seconds.

The third one, which took no less than three minutes, replaced the introduction to page 5. Instead of a mechanical but systematic discussion of all the pictures and the two things that they all expressed, a commented version, concentrating on the first two pictures and then going over the others very rapidly, was given.

The fourth and last theoretical comment in this lesson was in the form of a short reminder before the pupils started writing on page 6. It took 40 seconds. (Times given here refer to Ee; Es differs by twenty seconds only).

The total running time of the explicit lessons (lesson 7) was about the same as that for Im.

Gume 5

The lesson materials in this study is to a certain extent identical with the materials of GUME 3. However, as a higher form was chosen for the GUME 5 experiment, the grammatical content as well as the lesson materials was enlarged. The lessons consisted of speaking, writing, and reading modules, but it was not a matter of course, as in the GUME 3 experiment, that the order between these activities should be: (1) speaking, (2) writing, and (3) reading. So, for instance, writing drills can occur both att the beginning and at the end of a lesson. The exercises were the same for all three strategies with the exception that the Im-group had continued practice during the time taken up by explanations in the other two groups. The various kinds of drills as well as the explanations are of the same kind as those practised in GUME 3.

GUME A

In this part project, coming last in chronological order, only two strategies were compared, Im and Es. The main differences between these methods and those investigated at the comprehensive school level will be made clear below.

The following five grammatical structures were selected for study: (1) the use of some and any and their compounds; (2) adjectives and adverbs; (3)

preposition + gerund; (4) possessive pronouns; and (5) the passive voice. The proportion of time devoted to the various structures was as follows: structure (1) was covered in lessons 1–3, structure (2) was given in lessons 4 and 5, structures (3) and (4) shared the next three lessons, while structure (5) was dealt with in the last two lessons. Apart from a short revision of the previous lesson made at the beginning of each lesson, the structures were not dealt with on subsequent occasions. The time allotted to each structure, to revision and new material was very much the same in the two lesson series. There was a difference of ten minutes in the total duration of the two series in favor of the 1m version; however, this difference is explained by an instructional phase (at the beginning of the first lesson) aimed at explaining the drill technique.

During the lessons, which were all tape-recorded, the subject was supplied with a workbook containing the basic dialogue and some written exercises. A set of transparencies with series of pictures, illustrations, grammatical tables, and paradigms, facilitating structure drills, other oral activities, and grammatical explanations, accompanied each lesson. By these arrangements the role of the two teachers (who were identical with the investigators – see pp. 89–90) was limited to purely mechanical activities, such as handing out and collecting workbooks, and operating the tape-recorder and overhead projector.

THE IMPLICIT METHOD

The structure drills, which were carefully structured, were mostly of the three-phase type; occasionally, especially when they were not based on pictures, four-phase. Practice of audio-lingual skills was predominant. The basic text utilized each lesson had a dialogue containing the new grammatical pattern which was repeated several times. The lessons were entirely monolingual and thus contained no translation exercises.

THE EXPLICIT METHOD

In a typical Explicit lesson the structure presented in the basic text was carefully explained to the student by comparing or contrasting it with the corresponding Swedish structure. Exercises, both oral and written, were mostly of fill-in type or translations. Audio-lingual skills were not given priority, and owing to the explanations and translation exercises, a good deal of the lesson was given in the native tongue. Exercises that could be labeled as pattern drills were avoided.

The presence of such techniques as grammatical explanations in the native tongue and translation may suggest an identification of this type of teaching

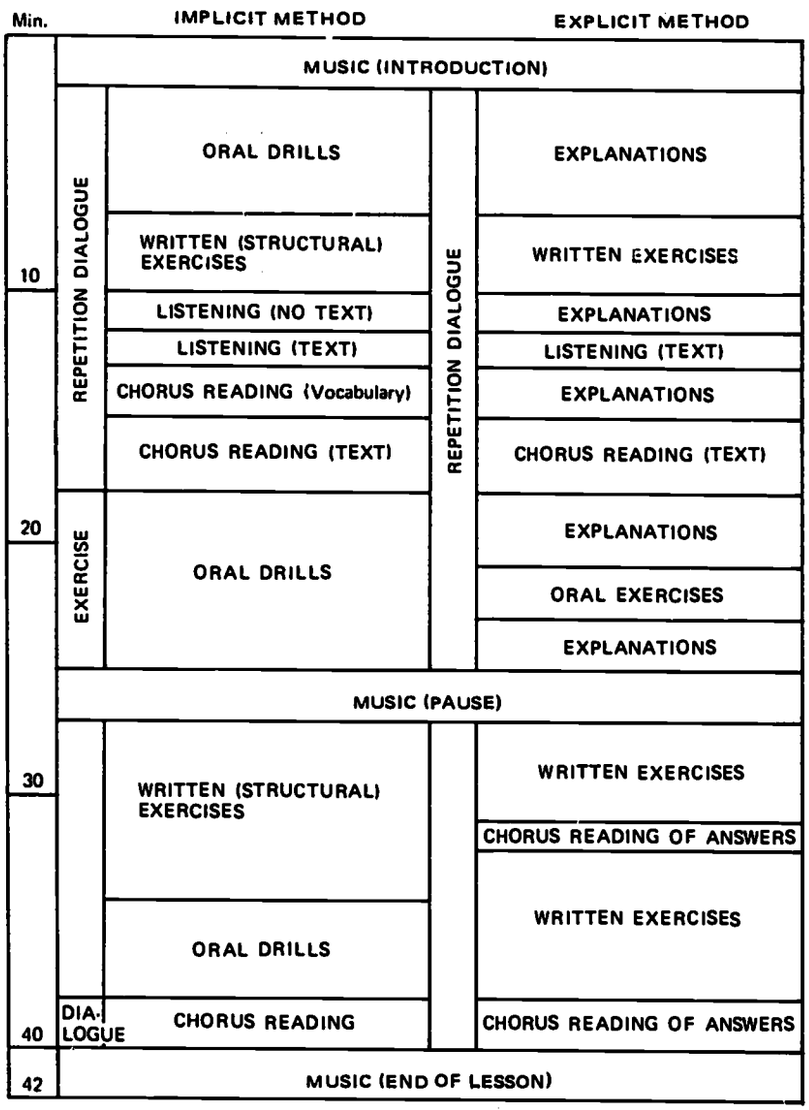


Fig. 8 Prototype lesson illustrating the distribution of activities (GUME A).

with the previously mentioned grammar-translation method (cf.pp 23-24). The Es method in GUME A, however, does not correspond to any definition of the grammar-translation method given by authorities on language teaching. In the Es lessons grammar was not taught as an end in itself, but was always followed by exercises containing every-day sentences, giving the learners the opportunity of immediate application of rules.

The main differences between the Es approach in this study and the previous one should be observed: in GUME A the Es method contains no systematized structure drills; on the other hand translation exercises (Swedish-English and English-Swedish) and rule-giving are utilized in order to make the subjects conscious of how the foreign language operates. In common with the previous Es methods it has the use of explanations and reference to the Swedish language. All in all, the method investigated in GUME A is of a more traditional character than those of the earlier part studies.

The lessons ordinarily followed one and the same pattern. Fig. 8 (see page facing) presents the sequence which the activities followed.

CHAPTER 10

EVALUATION INSTRUMENTS

General considerations

As was the case in the previous chapter it is also here impossible, for reasons of space, to present all details of concern. Again, the reader is referred to the part reports.

A criterion test intended to measure progress was constructed in each part project. Each test was to measure what had been specifically taught in the respective project; of necessity the test should have high content validity. The composition of the test varied from experiment to experiment (see table 16, p. 130).

With one exception (GUME A) only written tests were used. It may be argued that the spoken lanugage, which is an important aspect of language mastery, has been unduly neglected in our tests. A word of comment is in order. It should be stressed again (cf p. 113) that we never planned to cover the whole field of language learning; we are only interested in the pupils' active mastery of certain grammatical structures. It is very improbable that the pupils, in experiments of as short duration as the present ones, would increase their general speech production capacity, their pronunciation or intonation. Accordingly the training of these capacities was not included in the objectives of the experiments. On the other hand it may be argued that the pupils' ability to generate, in spoken form, the grammatical patterns concerned – with disregard of pronunciation and intonation errors – should have been investigated. We think that the marking procedure adopted in the written productive tests compensates for the lack of speaking tests. When the students' written answers were corrected no attention was paid to spelling errors (within reasonable limits – the marking was performed by assistants according to careful instructions, and all uncertainties were discussed with the project staff). Thus, if a wrongly spelt answer indicated that the student would be able to pronounce the word, or structure, correctly, he was given credit for his answer.

Similarly it may be argued that the criterion tests were biassed towards one method or the other. For instance, the Implicit method, in which the aural-oral skills are comparatively important, may be supposed to suffer most from our tests. We think that the arguments presented in the preceding para-

graph invalidate this criticism. On the other hand, the Implicit method may have been favoured by the testing technique which, in a number of cases, has a certain resemblance with the structural drills. This possibility can not be completely ruled out since the Implicit method contained more drills than did the Explicit methods; however, our impression during the testing periods was that the testing techniques caused no problems whatever. The hypothesis has been put forward that the testing time should operate in favor of the Explicit groups. In ordinary communication an individual has to deliver his answers rapidly whereas, in the tests, the subjects may be said to have had inordinately long time for their answers (which would favour the Explicit pupils when enough time is given to recall the explanation or rule and apply it). As is apparent from table 16, page 130, each test contains a fairly large number of items, and the time factor can hardly have had the kind of influence indicated. Finally, it has been stressed that test items of the fixed response variety may hamper the Explicit students since the generative aspect of language is neglected in "mechanical" tests of this kind. The generative aspects, or competence (as opposed to performance) is supposed to be particularly well developed by a cognitive approach. The counter-arguments may be put forward that (1) within each criterion test there is a balance between productive and fixed-response items, and (2) the correlations between the two types of subtests are of approximately the same magnitude as the intercorrelations within each variety of subtests.

Thus, to summarize this somewhat lengthy discussion we would say that it is difficult to gauge the bias, if any, of our tests. We are of the opinion that, in the light of the general objectives of our experiments, they do not favour any particular method.

Technical description of the tests

In this section a brief description will be given of the general nature of each test, particularly the relation productive/fixed-response items and their respective characteristics. Total number of items, testing occasions and testing time are given in table 16, p. 130.

GUME 1

The total test consists of 12 part tests, one of which is a listening test. The first two deal with the problem of how to answer questions, eight deal with the problem of how to ask questions, and the last two take up negative

sentences. Eight of the tests, including the listening test, are of the fixed-response type (2-, 3-, and 4-choice), four are productive in nature.

The following are sample items from the four subtests utilizing open answers; the pupils are to fill in one or more words, in some cases whole sentences: Subtest 1: *Do you go to school on Mondays?* Yes, ___. 2: *What colour did he paint his house?* He _____ it red.5: The pupils are to construct questions related to specific stimulus sentences.) *Ask me if I was in Scotland last summer.* _____ *in Scotland last summer?* 8: (The pupils are to change statements into questions.) *She sings very well.* _____ ? In the listening test (No. 4) the pupils are to mark, on a separate answer sheet, whether a spoken sentence is grammatically correct or incorrect.

GUME 2

The total test consists of three parts. The first two require the pupils to select the correct form *(some, any, somebody, anybody, something, anything)* in a given context. In the first part the pupils mark their answers on a separate sheet (6-choice), in the second part they fill in the right form into lacunas in a running text (the six forms were given at the beginning of the text). In the third subtest the pupils indicate whether written sentences are grammatically right or wrong; besides the six forms mentioned, also *somewhere* and *anywhere* are included.

GUME 3

The criterion test consists of 7 part tests, one of which is a listening test. In the latter twenty passive sentences are read from a tape-recorder. The pupils mark the correct answer to questions following this pattern: Stimulus sentence: *The flowers have been run over by the cars.* (pause) *What has been run over?* The options on the answer sheet are: (a) *the cars* (b) *the flowers.* Of the remaining 6 part tests, three are 2-choice tests and three require the students to write whole sentences. In one of the latter the students are to transform passive sentences into active, in another to transform active sentences into passive; the third test of a productive kind is simply a translation test (Swedish into English); this test consists exclusively of passive constructions to be translated.

GUME 4

The test contains seven parts, some of which utilize testing techniques and, in some cases, items identical with those in GUME 1 and GUME 2. Four of the tests require the students to produce their own answers, three are of the

fixed-response variety. The following sample items illustrate the types of answers asked for in the four productive tests: Subtest 1: *Does your father live in Oslo? No, he _____ in Gothenburg.* (Thus, only the proper form of the word underlined in the stimulus sentence is being tested.) 2: (Cf GUME 1, subtest 5): *Ask me if Susan watches TV every evening. _____TV every evening?* 7: (The student is required to ask a question which might follow logically upon the stimulus sentence): *He speaks many languages. _____ German, too?* 7: (The student is to answer a question; in doing so, he is supposed to agree with the first part of the question but to disagree with the last part): *I suppose Mr Austin has a car and washes it every week? Well, he has a car, but he _____ it every week.*

GUME 5

The test consists of six parts, one of which (No. 4) is identical with No. 4 in the GUME 3 criterion test. Generally speaking, the testing procedures in GUME 5 and GUME 3 have much in common. This is only natural since both tests stress, as did the teaching, the interrelationship between active and passive sentences and the formation of the verbal part of the passive sentence. One of the subtests is a listening test. In this a running text (a story about Dr. Doolittle and his friends) is read from a tape-recorder. The tape is occasionally stopped and the students asked to answer 4-choice items on the contents. Of the remaining five parts, one is a 2-choice test and four are of a productive kind. No. 1 consists of a completion test where 11 different forms of the auxiliary *be* have been removed in a running text. The student is to fill in the blanks. No. 2 requires the students to transform passive sentences into active; only the crucial words are to be filled in: *The film has already been forgotten by the children. The children _____ already _____ the film.* No. 5 is to measure the students' ability to form the verbal part of the passive sentence, and the verb to be employed is given in the infinitive. *Does anybody visit that old museum? Yes, it _____ (visit) by many people on Sundays.* Finally, in No. 6 the students are to write out the passive sentence corresponding to an active sentence of the following kind: *They sell beautiful clothes in Paris.*

GUME A

The criterion test consists of three parts, the first of which is a listening test. The students listen to a short conversation, 2–4 exchanges, between a male and a female voice. The last part of the last exchange, containing the crucial words, is left out on the tape. The students mark, on a separate sheet, which of three options constitutes the right completion of the taped dialogue. Sample item: *HE: Peter and his girl-friend have their lunch at a restaurant*

every day. SHE: He needs a lot of money then. HE: Not really. He pays for his *lunch and the girl-friend pays for* ___. The options are: (1) *hers lunch* (2) *hers* (3) *her.* Part 2 of the total test is a 3-choice test, whereas part 3 is a production test in which the students are to fill in the crucial element in incomplete English sentences. The meaning of each sentence is clarified either by the complete Swedish equivalent or by a Swedish cue word. The two following sample items illustrate the two testing techniques used: (a) *JIM: Are all those* ____ *(dina)? SUE: Yes* ____ *(naturligtvis) they are* ____ *(mina). You never write to* ____*(någon).* (b) *Mrs Williams tackade mig för att jag kom. Mrs Williams thanked me for* ____ .

GUME A is the only part project where an oral productive test has been administered. The test was only given as a posttest; a speaking test would probably have caused the adult subjects unnecessary irritation if given before the experiment started. In this test, which consists of 30 items, the student hears questions and incomplete answers. He is to repeat the latter, thereby also filling in the missing word or structure. The oral test was administered simultaneously with a listening test where the students had to identify grammatical and non-grammatical sentences, combine sentences with pictures, or state whether sentences were applicable to a certain picture or not. Instead of marking their answers on a sheet, the students gave an oral answer ("right" or "wrong", "number 1", "number 2", "number 3", etc) which was recorded on

Table 16. Survey of various characteristics of the criterion tests.

	Total number of						
	subtests	productive subtests	fixed-response subtests	listening subtests x)	items	testing occasions	testing time (minutes)
GUME 1	2	4	8	1	12x10=120	1	40
GUME 2	3	1	2	–	40,21,70=131	1	40
GUME 3	7	3	4	1	9,20,38,38, 38,10,10,8=133	2	30+30
GUME 4	7	4	3	–	10,15,45,20, 15,40,15=160	2	40+40
GUME 5	6	4	2	1	11,10,9,40 14,10=94	2	24+30
GUME A	3	1	2	1	60,50,20=130	2	25+39

x) The listening tests are always of the fixed-response variety and are included in the number given in the preceding column.

tape. The listening and oral productive parts have been added together in all computations; the combined test is called Oral Test. All recordings were made in an audio-active language laboratory.

Table 16 gives a survey of the features of the criterion tests. In the table only tests administered both as pre- and posttests have been included.

To reach as great uniformity as possible in the different classrooms the testing procedure was regulated from a tape; the tape ran through the whole testing period and was thus responsible for timing the test. All instructions were in Swedish. Each test was developed by a series of try-outs. Details about the revision work are given in the part reports.

In the present report only the students' total score on the criterion tests will be considered in the teaching method comparisons. Part test scores, intercorrelations between subtests, etc., are discussed in greater detail in the part reports.

Reliability

Test reliabilities are presented in Appendix 3. The magnitude of the reliability coefficients, which have been calculated on the pretest scores, is more than satisfactory for the purpose of the investigations (group comparisons). A word of comment is in order. The conventional types of reliability coefficients are related to the variation in scores in a particular group (see, for instance, Levin & Marton 1971, p. 41). However, it would have been advantageous, from an experimental point of view, if the students had been completely ignorant of the grammatical structures dealt with in the experiments; in such a (theoretical) case, the "zero-point" discussed earlier (pp. 67–68) would have prevailed, the variation in pretest scores would have been none, and the reliability would have been 0. When dealing with a foreign language which the subjects have come in contact with earlier it is probably out of the question to obtain test items (within reasonable limits) which do not reflect different achievement levels in a particular group of students. The high coefficients in our pretests are thus not desirable *per se*. However, since the great variation in ability among the pupils is a fact, the values indicate that this variation is measured with precision.

Validity

As stated above (p. 126) the content validity aspect is important in tests used to evaluate different teaching methods. However, this type of validity can be

checked only by a careful comparison between the tests and the instructional content in the lesson series. We are confident that the content validity of all the criterion tests is satisfactory.

Since the pretest scores reveal great variation we have calculated product-moment correlations between this test and an independent measure, namely Grades in English (in GUME A, where no grades were available, the diagnostic English test similarly offered at the beginning of the experiment, was used as criterion). For purposes of comparison we have calculated the corresponding correlations between the posttest and the two types of criteria mentioned. The values are given in the following table.

Table 17. Validity coefficients (uncorrected for attenuation).

	Correlations between Grades English* and		
	Pretest	Posttest	N
GUME 1 sk	.679	.688	227
GUME 1 ak	.572	.562	104
GUME 2 sk	.518	.582	247
GUME 2 ak	.455	.604	98
GUME 3 sk	.628	.645	170
GUME 3 ak	.659	.697	57
GUME 4	.697	.735	574
GUME 5 sk	.721	.683	235
GUME 5 ak	.425	.371	152
GUME A	.785	.729	125

*) In GUME A the diagnostic English test.

There are no systematic differences between the pre- and posttest correlations. In the two part projects where no division into courses was made (GUME 4 and GUME A), around 50 % of the variance in the criterion scores is explained by our tests, However, also in some of the more restricted groups the correlations approach this value. The low correlations in GUME 5 ak are partly explained by low reliability (see Appendix 3). Thus, whatever aspects are included in the criterion seem to be covered in fairly substantial degree by the pre- and posttest.

The student attitude test

In all experiments roughly similar questionnaires have been given. They contain items of two kinds, open answers and items with fixed-response alterna-

tives. The former will receive limited attention in the results chapter, whereas the latter will be accounted for in greater detail. We have grouped a number of the fixed-response alternatives together, assuming that these items reflect the pupils' general attitude towards the teaching method they received.

GUME 1–3

In the first three experiments the same attitude test was given. Nine items were selected for measuring the general attitude towards the experiments. Four of these (1–4) are about the series as a whole, five (5–9) about the technical quality and about the three parts of the lessons, oral, written, and reading-listening. The first four have five response alternatives, the last five questions have four. The following aspects are covered by the nine items:

1: learnt less more (than during ordinary lessons)

2: less fun – more fun

3: time went slower – faster

4: more tired – less tired

5: earphones bad – good

6: sound bad – good

7: oral drills bad – good

8: written drills bad – good

9: reading texts bad – good

Items number 1, 2, 5 and 6 are given in Appendix 11 to illustrate the two types of scales (4-point and 5-point). In each item the first alternative was given the highest (most positive) value, except in number 4, where "less tired" was considered – at least from the pupils' point of view – to be the most positive. The theoretical mean of the composite scale is 24.5, indicating a neutral attitude towards the experiment.

GUME 4

In this project the following seven items are supposed to gauge the students' general attitude towards the respective teaching method:

1: During project lessons I learnt *very little – very much.*

2: Project lessons were *very boring – great fun.*

3: In doing oral and written exercises I understood what to do *never – always.*

4: From the four-phase drills I learnt to speak English *very poorly – very well.*

5: From the four-phase drills I learnt English grammar *very poorly - very well.*

6: The four-phase drills were *very boring – great fun.*

7: The four-phase drills were *very difficult – very easy.*

All items are 5-point scales; the theoretical mean of the total attitude test is thus 21.0.

GUME 5

This test is identical with the test used in GUME 1–3, except that item number 5 is excluded since no earphones were used in the present project. The attitude test thus consists of four 5-point- and four 4-point scales, the total mean being 22.0.

The attitude tests accounted for hitherto will thus be used for comparing the three teaching methods (Im, Ee, Es) with respect to general attractiveness. We assume that the scales represent at least ordinal measurement and will apply a non-parametric test for k-sample cases, the Kruskal-Wallis one-way analysis of variance by ranks, (Siegel, 1956).

GUME A

Seven items were chosen from the total test to measure the general attitude; of these, four were identically the same in the Im and Es groups, whereas three were different. The latter were different in so far as they covered the main aspect of each method (Im: the oral drills; Es: the explanations), but they were identical in wording.

The following aspects were covered by the test:

1: The course as a whole was *very valuable – completely worthless.*

2: The lessons: *great fun – very boring.*

3: Would you recommend this course to be incorporated into other English courses? *Yes, absolutely – No, absolutely not.*

4: During the course my attitude was changed in the following manner: *I became more and more positive - I became more and more negative.*

In items 5–7 the items in the Im group refer to the oral drills, in Es to the explanations.

5: *very difficult – very easy*

6: *very effective – not effective at all.*

7: *ought to be substantially increased – ought to be completely replaced by explanations* (in the Es group: *oral drills).*

All items except number 4 had 5-point scales. This item had three response alternatives only: in the analyses they were assigned the numerical values 1, 3, and 5 respectively. The theoretical mean of the scale is 21.0.

Since only two teaching methods are compared we will apply the Mann-Whitney U test, which is relevant for 2-sample cases (Siegel, 1956).

The teacher attitude test

A questionnaire was administered in each of the GUME 1–5 projects. In GUME A, where the two experimenters administered the lesson series (cf. p. 90) opinions on the teaching procedures have been obtained from a number of observers visiting different lessons.

In the questionnaires the teachers participating in the experiments were asked questions on how they usually teach English themselves, which method they use, how they treat grammatical points, how much they speak English, etc. One part of the questionnaire required the teachers to comment on various aspects of the lesson series: the grammatical explanations (in the Explicit groups), the oral exercises, the written exercises, the reading passages, the tempo of the lessons, the sound quality of the tapes, the reactions on the part of the pupils, etc. In chapter 11 we shall briefly comment on the teachers' opinions in some of these respects.

CHAPTER 11

MAIN RESULTS

Overall progress.

In this chapter we shall investigate the differential effect, if any, of our teaching procedures on the students' acquirement of English. A necessary prerequisite for studying *differences* in progress between teaching methods is that the treatments have had measurable or, preferably, substantial effects on the pupils. In other words, did the pupils, irrespective of teaching method, learn anything from the respective lesson series? Before presenting the treatment comparisons proper we shall therefore give a picture of the overall progress during each experiment. Progress will be expressed in raw gain scores as well as in (A/P P)-scores, the latter relating actual progress to possible progress (cf pp 74–75).

Table 18. Raw gain scores and (A/P P) scores for the ten GUME groups.

		Raw gain scores		A/P P	
	N	$\bar{x}$	s	$\bar{x}$	s
GUME 1 sk	227	10.69	8.70	23.57	21.52
GUME 1 ak	104	3.31	8.08	4.15	12.07
GUME 2 sk	247	16.54	10.68	26.12	17.39
GUME 2 ak	98	12.85	13.57	14.95	16.06
GUME 3 sk	170	11.18	10.05	25.55	24.80
GUME 3 ak	57	4.14	8.44	5.67	13.44
GUME 4	574	17.26	12.32	17.63	14.42
GUME 5 sk	235	7.88	8.04	19.92	20.60
GUME 5 ak	152	3.72	6.80	5.39	10.42
GUME A	125	19.38	13.33	29.96	19.20

In all cases progress is made although in the ak groups (with the exception of GUME 2 ak) it is strikingly low. It is, however, theoretically possible for differences between treatments to exist. The sk groups and the two relatively heterogeneous samples, GUME 4 and GUME A, progress approximately 20–25 %, the majority of the ak samples around 5 %. A noteworthy fact is the magnitude of the standard deviations in relation to their respective means; in

all ak groups the SD's exceed the means. This indicates that the within-group variances are great and that a number of pupils make negative progress, i.e. they regress. Fig. 9 below is an illustration of this. The figure represents GUME 2 sk; the black field signifies regress scores. (In Appendix 9 the corresponding distributions are given for all groups.)

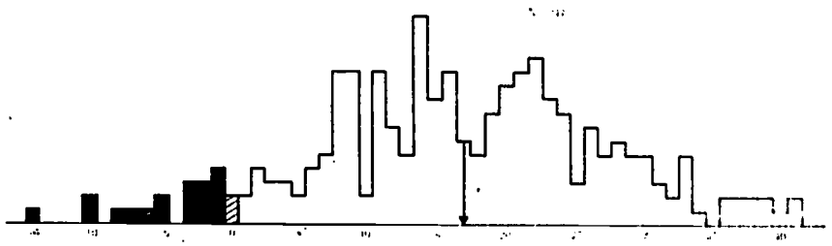
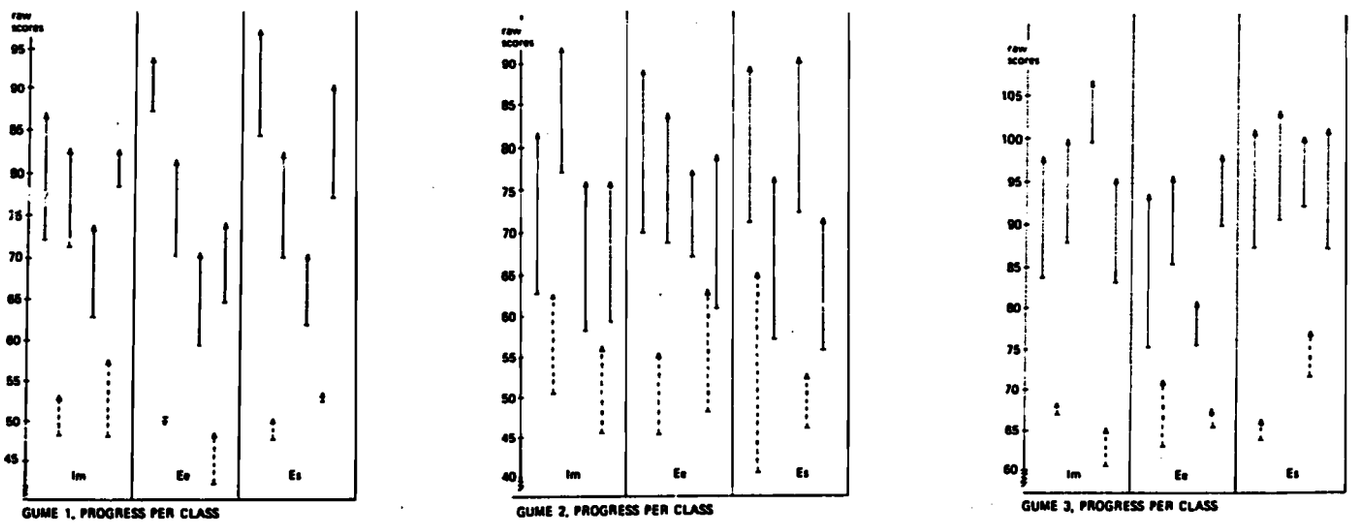


Figure 9. Distribution of raw gain scores (and regress scores) for GUME 2 sk; N = 247.

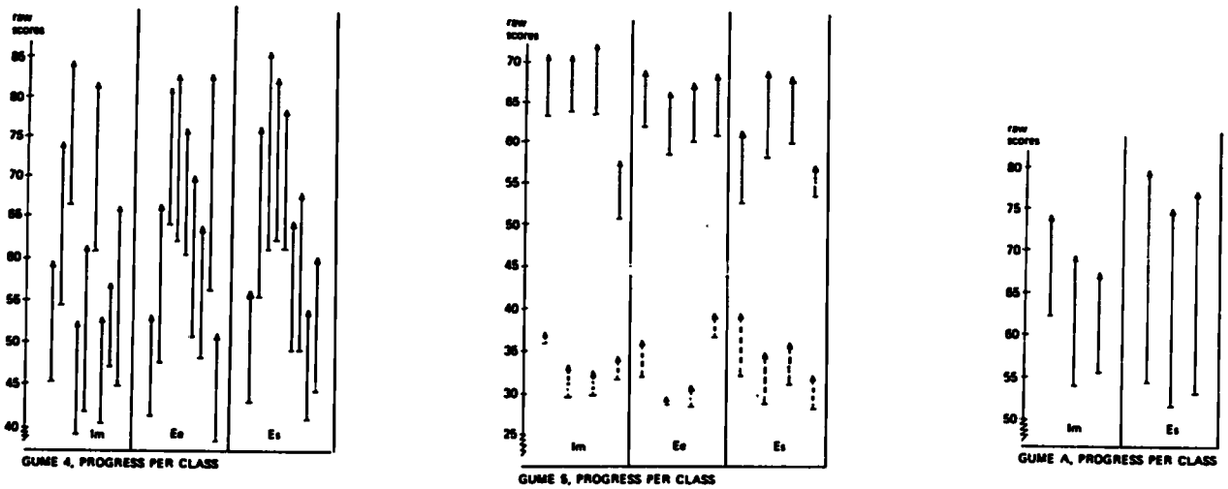
As the figure demonstrates a fairly large number of pupils have regressed. It is hardly probable that a regress score of, say, 5 points or more, represents a true score. More likely, it is a test effect, caused by lack of motivation on the posttest occasion. Of course, very high progress scores might similarly be explained as test effects because of low motivation on the pretest occasion. However, all scores, whatever their nature in this respect, have been included in the analyses. In all likehood, our decision not to exclude extreme regress scores implies an underestimation of the general effect of our teaching procedures.

General outline of the treatment comparisons

Before presenting the various analyses performed we shall give, in graphic form, a representation of the general outcome of the investigations. In the figures below the different school classes are indicated by arrows. The bottom end of each arrow signifies the pretest score, the top end gives the posttest score and the length is an indication of the magnitude of the progress made. The arrows are arranged in groups, one for each teaching method. In experimental groups containing both sk and ak classes, the latter are represented by broken arrows. In each case the scale (raw scores) is given on the left-hand side. It should thus be observed that the arrows represent classroom means and not individual scores.



Figures 10-15. Progress scores of participating school classes in the GUME 1-5 and GUME A samples.



Figures 10-15. Progress scores of participating school classes in the GUME 1-5 and GUME A samples.

We shall first comment on the figures for GUME 1-5, i.e. the groups at the comprehensive school level. In cases where both sk and ak classes occur, the most striking feature is the marked division into two groups of arrows (solid and broken). With the exception of three instances in GUME 2, the sk and ak arrows do not overlap. As far as length of arrows is concerned there is also a great difference between sk and ak; the sk classes generally make significantly greater progress than the ak classes. This fact, pointed out in table 18, is very prominent in the figures.

The two groups of arrows should be considered separately. Even so, the main impression is one of great variation within rather than between teaching strategies. It is an interesting finding *per se* that school classes vary so widely; as a matter of fact, the mean pretest score of many classes surpass the posttest score of others. The school class mean variation for pretest as well as posttest scores (the starting point and the end of the arrows) is greater for sk classes than for ak classes; in the more heterogeneous GUME 4 classes it is still larger. There is no doubt that this variation represents a reality of great educational impact. Against this background it is somewhat surprising that, in the recent curriculum (Lgr 69, p. 145), it is stated that the teaching objectives should be the same for sk and ak.

In the figures representing our samples at the comprehensive school level it is difficult to detect any systematic differences between teaching strategies. The general impression is one of great variation within strategies and between classes, not so between strategies.

The figure representing GUME A, on the other hand, shows a more consistent pattern. The Es arrows, starting at a somewhat lower point than the Im arrows, reach higher, which indicates greater progress. We shall now proceed to investigate what significance, in statistical terms, the graphic representation of the various outcomes may have.

Investigation of main effects

THE INDIVIDUAL SCORE AS THE UNIT OF ANALYSIS

We shall first present the various analyses of covariance, all performed with the posttest as the dependent variable and the pretest as the covariate.

Table 19. Analyses of covariance

Dependent variable: Posttest
Covariate: Pretest

	Adjusted means				ss'y			
	Im	Ee	Es	F-ratio	be-tween	with-in	df	b_w
GUME 1 sk	81.91	80.57	82.77	1.311	193	16404	2/223	.904
GUME 1 ak	55.75	50.55	50.43	4.246	497	5848	2/100	.845
GUME 2 sk	81.58	80.28	82.32	0.815	175	26067	2/243	.846
GUME 2 ak	58.73	60.43	59.73	0.163	55	15800	2/94	.679
GUME 3 sk	98.24	94.97	98.96	2.798	524	15533	2/166	.817
GUME 3 ak	68.07	70.06	70.27	0.374	50	3537	2/53	.732
GUME 4	68.49	68.83	68.89	0.065	18	78739	2/570	1.182
GUME 5 sk	67.17	66.96	66.84	0.038	4	13215	2/231	.806
GUME 5 ak	34.48	33.74	37.36	4.543	377	6133	2/148	.765
GUME A	68.05		79.18	25.399	3797	18238	1/122	1.021

——— = significant at the 5 % level ═══ = significant at the 1 % level

Before commenting on the results we shall present the same analyses corrected for unreliability of the covariate.

Table 20. Analyses of covariance, corrected for unreliability of the covariate.

Dependent variable: Posttest
Covariate. Pretest

	Adjusted means				ss'y			
	Im	Ee	Es	F-ratio	be-tween	with-in	df	b_w
GUME 1 sk	81.93	80.64	82.69	1.701	168	11006	2/223	1.005
GUME 1 ak	55.79	50.85	50.08	7.477	516	3447	2/100	1.030
GUME 2 sk	81.62	80.11	82.49	1.418	238	20346	2/243	.940
GUME 2 ak	58.59	60.38	60.05	0.216	66	14375	2/94	.789
GUME 3 sk	98.15	95.39	98.58	2.376	345	12035	2/166	.929
GUME 3 ak	68.66	70.36	69.54	0.293	26	2309	2/53	1.017
GUME 4	68.73	68.68	68.82	0.015	3	53146	2/570	1.271
GUME 5 sk	67.04	66.87	67.07	0.028	2	9970	2/231	.885
GUME 5 ak	33.63	34.12	37.81	12.541	531	3135	2/148	1.296
GUME A	67.79		79.39	36.380	4119	13813	1/122	1.160

Correction for unreliability obviously does not change the main results from table 19. The adjusted means are nearly the same in both analyses. However, in three cases, GUME 1 ak, GUME 5 ak, and GUME A, there is a substantial increase in F-ratios when the correction is made. This increase is mainly due to the fact that the treatment means changed position from pretest to posttest; in fact, the group ranking first on the pretest ranked last on the posttest.

In these two tables we thus have a first indication of the general outcome or our treatment comparisons. In the sk courses at the comprehensive school level it matters little which method is used; in none of the sk courses is a significant F-ratio obtained. Similarly, in the heterogeneous GUME 4 group the F-ratio is far from significant.

In two of the four ak courses, GUME 1 ak and GUME 5 ak, significant F-ratios were obtained. Before commenting on them we shall make tests of homogeneity of regression (cf Snedecor & Cochran, 1967, p. 432 ff).

Table 21. Test of homogeneity of regression for GUME 1 ak

Within	df	$\Sigma(x-\bar{x})^2$	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(y-\bar{y})^2$	Regr. coeff.	Deviations from regression		
						df	ss	ms
Im	22	3075	3029	4278	.985	21	1294	61.62
Ee	41	7073	5602	6402	.792	40	1965	49.13
Es	38	5618	4584	6612	.816	37	2871	77.59
						98	6130	62.55
Pooled	101	15766	13215	17292	.838	100	6218	62.18

F = 88/62.55 = 1.41 N.S (df = 2/98)

The regression slopes do not deviate significantly from each other in the GUME 1 ak sample. Thus interpretation of the treatment effects (Im > Ee = Es) is permitted.

Table 22. Test of homogeneity of regression for GUME 5 ak.

Within	df	$\Sigma(x-\bar{x})^2$	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(y-\bar{y})^2$	Regr. coeff.	Deviations from regression		
						df	ss	ms
Im	49	2257	1435	2899	.636	48	1986	41.38
Ee	48	2407	2368	4060	.984	47	1730	36.81
Es	53	2866	1957	3698	.683	52	2361	45.40
						147	6077	41.34
Pooled	150	7530	5760	10657	.765	149	6251	41.95

F = 174/41.34 = 4.21 $p < .05$ (df = 2/147)

In the GUME 5 ak sample the regression lines for the three groups differ significantly. Although this is an interesting finding *per se*, indicating the existence of interaction between initial and final scores, it ends further search for main effects.

Thus, in our ak samples there are three no-difference results and one showing a statistical superiority for the Implicit method (GUME 1 ak: Im > Ee = Es). As was mentioned in chapter 9 (pp 116-117) the explanations used (in GUME 1) were of a transformational kind and in some cases rather elaborate in nature. It is probable that, in a group of limited scholastic aptitude such as our ak sample, explanations of this kind are out of the question.

The general impression of the analyses at the comprehensive school level is thus one of non-significant differences. We shall now turn to the analysis at the adult level. In the table below a test of homogeneity of regression is made.

Table 23. Test of homogeneity of regression for GUME A

Within	df	$\Sigma(x-\bar{x})^2$	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(y-\bar{y})^2$	Regr. coeff.	Deviations fr. regression		
						df	ss	ms
Im	56	19131	18060	22863	.944	55	5814	105.71
Es	67	12518	14233	28650	1.137	66	12467	188.89
						121	18281	151.08
Pooled	123	31649	32293	51513	1.021	123	18542	150.75

F = 261/151.08 = 1.73 N.S. (df = 1/121)

The regression slopes for the two treatment groups do not differ significantly, which permits us to interpret the main treatment effects.

The results for GUME A in tables 19 and 20 point to a significant superiority for the Es method. This finding, in strong contrast to the results previously discussed, indicates that adult students profit from a method utilizing explanations. When the results at the comprehensive school level and the adult level are compared the differences between the various teaching procedures should be kept in mind. In GUME A the difference between Im and Es is more marked than the differences between Im and Ee/Es in the GUME 1-5 experiments. Part of the Es-Im difference in GUME A may thus be explained by the characteristics of the teaching materials used. However, considering also the fundamental similarities between the GUME A materials and those used in GUME 1-5, the overall results in fact indicate that adults profit relatively more than younger students from a method including explanation and analysis of grammatical structures.

For purposes of comparison we shall now present the results of the analyses of difference scores. The general characteristics of difference scores

have been treated previously (see p.73 ff); their main advantage is that they provide a rough picture of the progress made during the lesson series. In table 24 and 25 below analyses of variance for raw gain scores and A/P P scores will be presented. In the case of GUME A, where only two groups are being compared, t-values are given instead of F-ratios.

Table 24. Analyses of variance (one-way) of raw gain (progress) scores

	Im	Ee	Es	F	SS between	SS within	df
GUME 1 sk	10.86	9.56	11.62	1.124	170	16951	2/224
GUME 1 ak	7.44	2.45	1.80	4.157	512	6216	2/101
GUME 2 sk	16.88	15.23	17.83	1.258	286	27769	2/244
GUME 2 ak	11.55	13.53	13.91	0.284	106	17759	2/95
GUME 3 sk	12.00	9.57	12.25	1.297	261	16820	2/167
GUME 3 ak	3.19	4.90	4.14	0.178	26	3961	2/54
GUME 4	16.52	17.64	17.55	0.470	143	86779	2/571
GUME 5 sk	7.76	7.64	8.30	0.148	19	15123	2/232
GUME 5 ak	2.58	2.39	6.04	4.973	437	6542	2/149
GUME A	13.37		24.43	5.19=t			123

Before commenting on the results we shall give the corresponding analyses for the A/P P scores.

Table 25. Analyses of variance (one-way) of A/P P scores

	Im	Ee	Es	F	SS between	SS within	df
GUME 1 sk	23.42	19.73	27.36	2.517	2301	102393	2/224
GUME 1 ak	10.65	2.74	1.85	4.642	1263	13742	2/101
GUME 2 sk	26.69	24.09	28.09	1.128	682	73705	2/244
GUME 2 ak	13.87	16.13	14.77	0.187	98	24919	2/95
GUME 3 sk	24.18	21.49	31.25	2.466	2980	100924	2/167
GUME 3 ak	3.13	7.10	6.24	0.410	151	9971	2/54
GUME 4	16.29	18.37	18.12	1.141	474	118638	2/571
GUME 5 sk	20.24	19.72	19.88	0.013	11	99321	2/232
GUME 5 ak	3.66	3.59	8.68	4.223	881	15553	2/149
GUME A	19.39		33.32	4.41=t			123

The two types of analyses yield almost identical results. Interestingly enough, they also coincide with the previous analyses of covariance. It thus appears that analyses of difference scores produce approximately the same results as the analyses of covariance when the correlation between the covariate and the dependent variable is relatively high. Since the results correspond more or less exactly to those previously reported, we shall not comment further on them here.

In order to investigate interactions (see below p. 154) we have performed a number of analyses of variance (two-way classification). Simultaneously these analyses represent treatment x levels designs and provide a second check on the results of the main treatment effects accounted for above. It should be noted that they do not imply control in a stricter sense since, in each sample, the sk and ak courses have been lumped together. This procedure might be questioned because the two courses represent different populations. However, since it is stated in the curriculum that the goal should be the same for the two courses and since, in our experiments, both courses received exactly the same treatment, we shall tentatively investigate treatment effects in the pooled groups. The main reason for not performing the analyses on each course separately is the fact that the number of observations is too low for a 3×3 cells analysis, especially in the ak courses.

The actual procedure followed is one suggested by Searle (1971) which takes so-called unbalanced data, i.e. unequal number of observations in the cells, into consideration. As was stated previously (p. 76) it is to some extent arbitrary how many subjects are included in the different cells. Searle's procedure implies fitting various effects to the following model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

In unbalanced data, the F-statistic for α (row effects) after fitting μ to the model is not identical to the same statistic after fitting μ and β to the model (which is the case for balanced data). Similarly, $F(\beta/\mu)$ and $F(\beta/\mu,\alpha)$ are not identical; the tests are not the same, and neither of them should be described, albeit loosely, as "testing β-effects" (op cit p 76). The general problem considered by Searle is: what conclusions can de drawn from the various combinations of results that can arise *vis á vis* the significance and non-significance of $F(\alpha/\mu)$, $F(\beta/\mu,\alpha)$, $F(\beta/\mu)$ and $F(\alpha/\mu,\beta)$?
The survey below indicates what conclusions may be drawn depending on the magnitude of the various F-ratios. The reader is referred to the details of the analyses of variance (two-way) in Appendix 7, tables I – VI, for a check of the conclusions to be presented below.

Survey of suggested conclusions according to significance and non-significance of F-statistics in fitting a model with two main effects α and β (from Searle 1971, p. 278).

Fitting α and then β after α	Fitting β and then α after β				
	$F(\beta/\mu)$:	Sig	NS	Sig	NS
	$F(\alpha/\mu,\beta)$:	Sig	Sig	NS	NS
	Effects to be included in model				
$F(\alpha\|\mu)$: Sig $F(\beta\|\mu,\alpha)$: Sig		α and β	α and β	β	α and β
$F(\alpha\|\mu)$: NS $F(\beta\|\mu,\alpha)$: Sig		α and β	α and β	β	α and β
$F(\alpha\|\mu)$: Sig $F(\beta\|\mu,\alpha)$: NS		α	α	α and β	α
$F(\alpha\|\mu)$: NS $F(\beta\|\mu,\alpha)$: NS		α and β	α and β	β	neither α nor β

The results of the various analyses have been interpreted according to this survey; they are presented in table 26 below.

Table 26. Interpretable effects in the analyses of variance (two-way classification). Dependent variable: Posttest. Each sample divided into three approximately equal parts according to pretest scores.

(See Appendix 6 for the critical scores used in dividing each sample in three approximately equal levels (Upper, Middle, Lower).

	Effects (x) to be included		
	Row	Column	Interaction
GUME 1	X	–	–
GUME 2	X	–	–
GUME 3	X	–	X
GUME 4	X	–	–
GUME 5	X	–	–
GUME A	X	X[1)]	–

1) Es > 1m $p < .01$

The column effects, i.e. the values indicating overall differences between the teaching methods, are in accordance with the results presented previously. This means that the Es method is clearly superior at the adult level (F-ratio 11.225; df = 1/119) whereas, at the comprehensive school level, the picture is one of no differences. The row effects are all strongly significant, which is of

little interest in this connection, however. Interaction will be commented on in a subsequent section (p 154 ff.). To summarize the findings presented thus far it may be stated that, in general, the results in the three types of analysis – analysis of covariance, analysis of variance, and treatment *x* levels designs – coincide.

We have consistently presented results pertaining to the posttest scores of the various criterion tests. In one case (GUME A, cf p 130) a criterion test consisting of listening and oral production items was administered. The test was only given as posttest after the experiment. A little more than two thirds of the experimental sample took the oral test. The two teaching strategies were compared on the oral test by analysis of covariance. As covariates were used the three tests selected first in the step-wise regression procedure described earlier (p 78 and Appendix 5); the pretest, the diagnostic English test, and PACT. The result is given in the table below.

Table 27. Analysis of covariance, GUME A
Dependent variable: Oral test
Covariate: Pretest + Diagnostic English test + PACT (weighted sum)

Adjusted means			Sum of sqs		
Im	Es	F-ratio	between	within	df
32.65	36.74	11.267	356	2718	1/86

$F = 6.96$ ($p < .01$)

In all previous analyses the results in this part project have been consistent: the Es group is clearly superior. This is also the case in respect of the oral test which might have been supposed to favour the audio-lingually oriented Im method. This finding lends extra support to the Es method at the adult level.

As stated earlier (p. 87) the criterion test was administered as a test of retention five weeks after the experiments in GUME 1, GUME 2, and GUME

Table 28. Analyses of covariance of retest scores for GUME 1-3

	Adjusted means				ss'y			
	Im	Ee	Es	F-ratio	be-tween	with-in	df	b_w
GUME 1 sk	85.35	82.12	85.04	2.358	432	18480	2/202	.890
GUME 1 ak	56.24	54.33	54.53	0.430	54	5413	2/87	.802
GUME 2 sk	82.52	83.36	83.17	0.119	31	28977	2/225	.811
GUME 2 ak	58.93	58.42	56.95	0.168	48	11999	2/84	.706
GUME 3 sk	99.65	99.19	99.03	0.037	9	17262	2/144	.798
GUME 3 ak	72.18	71.41	74.96	0.661	118	4116	2/46	1.012

3. (The reasons for not administering retests in GUME 4, GUME 5 and GUME A were given on pp. 89–90). In table 28 on the preceding page the analyses of covariance of retest scores are presented. Before commenting on the results we shall give the corresponding analyses corrected for unreliability of the covariate.

Table 29. Analyses of covariance of retest scores corrected for unreliability of the covariate; GUME 1-3

	Adjusted means				ss'_y			
	Im	Ee	Es	F-ratio	be-tween	with-in	df	b_w
GUME 1 sk	85.40	82.26	84.85	2.798	383	13808	2/202	.989
GUME 1 ak	56.30	54.59	54.22	0.743	60	3508	2/87	.978
GUME 2 sk	82.53	83.20	83.36	0.135	29	24363	2/225	.901
GUME 2 ak	58.84	58.43	57.08	0.151	38	10649	2/84	.821
GUME 3 sk	99.58	99.60	98.66	0.133	27	14376	2/144	.906
GUME 3 ak	73.23	71.55	73.95	0.597	52	1991	2/46	1.406

Again, the two types of analyses yield practically identical results. As both tables demonstrate, there are no significant differences between the treatments in retest scores. When the adjusted means in table 29 are compared with those in table 20 a slight progress from posttest to retest is noticeable (with the exception of GUME 2 ak, Ee and Es); however, the results are not based on identical subjects because of drop-outs from posttest to retest and should therefore be interpreted cautiously. In sum, whatever tendencies towards differences appeared in the posttest scores have vanished at the time of the retest.

THE CLASSROOM MEANS AS THE UNIT OF ANALYSIS

It is to be regretted that the following analyses, performed with the classroom mean as the statistical unit of analysis, cannot provide complete comparisons with the preceding analyses. In four of our samples, GUME 1 ak, GUME 2 ak, GUME 3 ak, and GUME A, the degrees of freedom are too limited for meaningful analyses to be made. However, the remaining analyses will make possible tentative comparisons.

Table 30. Analyses of covariance of school class means
Dependent variable: Posttest
Covariate: Pretest

	Adjusted means				ss'_y			
	Im	Ee	Es	F-ratio	be-tween	with-in	df	b_W
GUME 1 sk	82.11	81.27	83.75	0.585	12.4	84.7	2/8	.944
GUME 2 sk	81.85	80.51	83.03	0.685	12.4	72.2	2/8	.955
GUME 3 sl.	98.66	95.46	99.61	1.068	26.8	100.5	2/8	.664
GUME 4	68.47	68.41	68.57	0.007	0.2	259.9	2/23	1.250
GUME 5 sk	66.46	66.31	67.28	0.241	1.7	27.8	2/8	1.126
GUME 5 ak	34.20	34.34	37.43	6.189	23.2	15.0	2/8	1.062

A comparison with tables 19-20 and 24-25 makes clear that the analyses performed on the classroom means do not change the general picture, which is still one of insignificant differences between teaching methods. The significant F-ratio in the case of GUME 5 ak should not be taken to indicate interpretable differences between treatments. As fig. 16 below illustrates, two of the regression lines intersect.

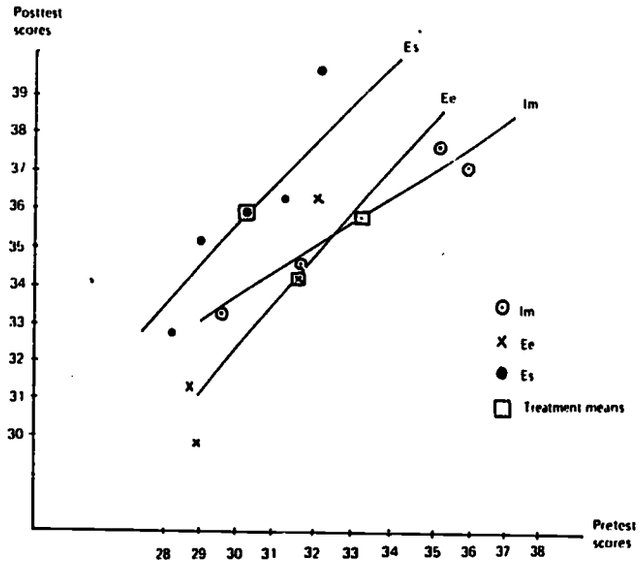


Fig. 16 Regression lines for the treatment groups (Im, Ee, Es) in GUME 5 sk. Classroom means used as observations.

A test of homogeneity of regression was made for the regression lines based on classroom means.

Table 31. Test of homogeneity of regression for GUME 5 ak (classroom means).

Within	df	$\Sigma(x-\bar{x})^2$	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(y-\bar{y})^2$	Regr. coeff.	Deviations from regression df	ss	ms
Im	3	26.75	17.79	13.00	.665	2	1.17	.59
Ee	3	41.50	49.05	63.75	1.182	2	5.77	2.89
Es	3	10.25	15.49	25.25	1.511	2	1.84	.92
						6	8.78	1.46
Pooled	9	78.50	82.33	102.00	1.049	8	15.64	7.82

F = 6.86/1.46 = 4.70 N.S. (df = 2/6).

Although the F-ratio in not significant, it is close to the critical value (5.14). Taken together, the statistical test and the previous figure indicate that the treatment differences in GUME 5 ak are hazardous to interpret. The general impression of equality between the teaching procedures thus prevails when the analyses are performed on classroom means.

The following table is intended to provide the reader with an outline of the findings presented thus far. The table also illustrates to what extent the different types of statistical treatment bring about similar results.

Table 32. Survey of main results in the treatment comparisons:
Dependent variable: Posttest scores

	ANCOVA (ind. scores)	ANCOVA (group means)	ANOVA (raw scores)	ANOVA (A/P P scores)	ANOVA 2-way (treatments x levels design
GUME 1 sk	0	0	0	0	0
GUME 1 ak	Im > Ee=Es	–	Im > Ee=Es	Im > Ee=Es	
GUME 2 sk	0	0	0	0	0
GUME 2 ak	0	–	0	0	
GUME 3 sk	0	0	0	0	0
GUME 3 ak	0	–	0	0	
GUME 4	0	0	0	0	0
GUME 5 sk	0	0	0	0	0
GUME 5 ak	0	(Es > Ee=Im)	Es > Im=Ee	Es > Im=Ee	
GUME A	Es > Im	–	Es > Im	Es > Im	Es > Im

0 = no differences found between teaching strategies – = no calculations made

Concerning the main results obtained in the various statistical analyses, the following conclusions seem warranted: The one-way analyses of variance, no matter whether they are calculated on the raw difference scores or on the A/P P scores, give results similar to those obtained by the analyses of covariance in nine cases out of the ten investigated. The only deviation from the general pattern of equality is found in GUME 5 ak where the analyses of variance indicate a main treatment effect whereas the analyses of covariance do not. In this particular case non-parallel regression lines obviated interpretation of main effects (which were similar to those found in the analyses of variance). Thus, our data do not substantially corroborate the findings of Feldt, Cronbach and others (cf. chapter 6) concerning the danger of utilizing difference scores. However, it should be noted that our findings hold for cases where there is a fairly large correlation between the x and y scores and when the scores are highly reliable; they should of course not be generalized to more unrelated or more unreliable variables. GUME 5 ak may be regarded as an illustration of this: the pretest reliability is particularly low (.59) and the pretest-posttest correlation is moderate, .63. Since the different scores are still more unreliable, and since the analysis of variance does not correct for initial scores, the result will be particularly influenced by chance.

The treatment x levels designs, although they are not strictly comparable with the analyses of covariance, coincide with the latter. That is to say, they underscore the insignificant differences found in GUME 1-5 and the superiority of the method utlizing explanations at the adult level.

It makes no difference, in our data, if the analyses are performed at the individual or school class level. In GUME 5 ak the main effect obtained at the school class level (within brackets in the table) might have been replaced by a zero sign; it is difficult to interpret the differences between the methods because two of the regression lines intersect.

Thus, if the analyses of covariance and analyses of variance (two-way) are chosen as the most valid analyses, the general outcome of the treatment comparisons is clear: At the comprehensive school level, there is no evidence of differences between the teaching strategies compared. The only exception, GUME 1 ak, is probably best explained by the fact that the transformational-generative explanations used proved too complicated for pupils of relatively low ability. At the adult level the results are unequivocal: in all comparisons the Es method proves superior, in statistical terms strongly significantly.

The interpretation of the results at the comprehensive school level are valid in so far as the various experiments are analysed separately. However, since the different investigations may be looked upon as a series of independent observations, it is of interest to regard the results in a more global perspective. Within each experiment the ranking of the three methods will follow one of six possible permutations. Thus, the probability of any particu-

lar ranking is 1/6. If the ranking of methods then follows the same pattern consistently in n experiments the probability of obtaining this result would be $(1/6)^n$. In the sk samples (n = 3) this is precisely what happens. In the analyses of covariance corrected for unreliability (table 20), in the analyses of variance of raw gain scores (table 24), and in the analyses of covariance of school class means (table 30) the ranking of methods within sk is identically the same, namely (1) Es (2) Im (3) Ee. The values are given in table 32 below.

Table 33. Treatment means (the sk samples) in various analyses.

	Analysis of cov., corr. for unreliability (table 20)			Analysis of variance of raw gain scores (table 24)			Analysis of covariance of school class means (table 30)		
	Im	Ee	Es	Im	Ee	Es	Im	Ee	Es
GUME 1 sk	81.93	80.64	82.69	10.86	9.56	11.62	82.11	81.27	83.75
GUME 2 sk	81.62	80.11	82.49	16.88	15.23	17.83	81.85	80.51	83.03
GUME 3 sk	98.15	95.39	98.58	12.00	9.57	12.25	98.66	95.46	99.61
GUME 5 sk	67.04	66.87	67.07	7.76	7.64	8.30	66.46	66.31	67.28

Within each analysis the probability of repeating exactly the Es> Im > Ee order is $(1/6)^3$, i.e. 0.005. Some of the differences between means are admittedly small, but in view of the great within-group variance in our data even marginal differences are of interest. It should be noted that the analyses of variance of A/P P scores showed the same ranking of methods as the analyses just presented in three cases out of four; in GUME 5 sk the order of means is (1). Im, (2) Es, (3) Ee. However, we think it is safe to conclude that, at the comprehensive school level, the pupils belonging to the advanced course tend to profit most from the Es method and least from the Ee method, the Im method coming somewhere between the two in efficacy. On the other hand, no such tendency is discernible in the easier course; in this case the conclusions drawn previously hold even when the different experiments are considered as a whole.

Investigation of pre- and posttest variances

In view of the fact that the different teaching strategies produced small or no differences in terms of means (at the comprehensive school level), one might still ask if they influenced the group variances differently. For instance, a

teaching method tending to heterogenize the group strongly would complicate individualization. It is a common finding (see, for instance, Anastasi, 1958, p. 195) that group variances increase as a result of instruction. In our case there is reason to investigate two things: (a) do the group variances increase?, (b) do they increase differently? In the GUME 1–5 samples we have calculated, for each of the teaching methods, a posttest/pretest or pretest/posttest variance ratio; that is to say, we have consistently put the highest value in the numerator. In table 34 below a (+) sign indicates an increase in variance from pre- to posttest and a (–) sign indicates a decline.

Table 34. Posttest/pretest (+) and pretest/posttest (–) variance ratios in GUME 1-5

	Im		Ee		Es	
	N	F-ratio	N	F-ratio	N	F-ratio
GUME 1 sk	69	1.076 (+)	77	1.059 (+)	81	1.151 (+)
GUME 1 ak	23	1.392 (+)	42	1.105 (–)	39	1.177 (+)
GUME 2 sk	84	1.177 (+)	92	1.140 (+)	71	1.079 ( )
GUME 2 ak	38	1.274 (+)	38	1.334 (+)	22	1.266 (+)
GUME 3 sk	50	1.491 (–)	63	1.271 (+)	57	1.183 (+)
GUME 3 ak	16	1.520 (–)	20	1.504 (+)	21	1.451 (+)
GUME 4	180	1.434 (+)	194	1.754 (+)	200	1.880 (+)
GUME 5 sk	70	1.225 (–)	92	1.006 (–)	73	1.104 (–)
GUME 5 ak	50	1.282 (+)	49	1.685 (+)	53	1.291 (+)

The (–) signs make it clear that there is not always an increase in variance from pre- to posttest. However, in the cases of decrease no significant F-values are obtained; there is thus no decline either. In all three GUME 4 methods and in GUME 5 ak, the Ee group,the variances increase significantly from pre- to posttest. In all, there is a general tendency for the variances to increase (20 cases out of 27). In order to investigate whether the changes in variance differ from treatment to treatment we have used the following procedure: Within each part project and each teaching method the pre- and posttest variances for school classes have been compared. For each school class, the difference between the post- and pretest variance has been calculated, and then these differences have been ranked according to size. The ranks have been compared by the Kruskal-Wallis one-way analysis of variance by ranks; the calculations yielded the results presented in table 35 below. As was the case previously when the analyses were performed with the school class means as the unit of analysis, the ak groups (in GUME 1–3) will have to be left out because of the small number of observations.

Table 35. Kruskal-Wallis analysis of variance by ranks of posttest-pretest variance differences.

	Sum of ranks			
	Im	Ee	Es	H
GUME 1 sk	25	20	33	1.65
GUME 2 sk	19	26	33	1.88
GUME 3 sk	33	19	26	1.88
GUME 4	173	114	91	6.31
GUME 5 sk	27	19	32	1.64
GUME 5 ak	26	20	32	1.37

In GUME 4, where a significant difference is obtained between the three treatments ($p < .05$), pairs of treatments were compared by the Mann-Whitney U test. One significant difference was obtained: Es > Im ($p. < .02$). Thus, the Explicit-Swedish method tends to increase the variation among the pupils more than the Implicit method does. This result indicates that, in our youngest group (grade 6), the method utilizing explanations in the mother tongue tends to favour the more able students and put a handicap on the less able. This is tantamount to stating that the GUME 4 result indicates the existence of interaction between aptitude level and treatment. (Interaction problems will be dealt with in the following section). In the remaining experiments no tendency towards differential treatment effects on the variances were found.

Investigation of interaction effects

In each experiment two analyses of variance (two-way classification) have been calculated, in both cases with the posttest as the dependent variable. In the first analysis the samples are divided into three thirds according to aptitude scores, in the second analysis the division is made according to pretest scores; the latter type was presented as treatment *x* levels designs on p 145 where we were concerned with overall treatment effects. However, in this section we will investigate whether the teaching methods produced different learning effects at different levels of ability.

In the case of GUME 1–5 general aptitude ("intelligence") is measured by the DBA test; the composite test includes the verbal, inductive, and spatial parts. In GUME A it proved impossible, for practical reasons, to administer more than the verbal part of the F-test. The critical scores for dividing the experimental samples into three thirds according to scores on the mentioned tests as well as the pretests are given in Appendix 6.

The two analyses are not based on an identical number of observations since, in each sample, some pupils were not present on the aptitude test occasion. Table 36 below illustrates the differences in this respect.

Table 36. Number of observations in each of two analyses of variance (two-way classification).

	Dependent variable: Posttest		Loss of observations from (1) to (2)	
	Indep var: Pretest	Indep var: Aptitude scores		
	(1)	(2)	N	%
GUME 1	330	311	19	5.8
GUME 2	344	320	24	7.0
GUME 3	227	209	18	7.9
GUME 4	574	561	13	2.3
GUME 5	386	334	52	13.5
GUME A	125	111	14	11.2

One way to check whether the loss of observations has been systematic with respect to teaching method is to compare the column means for the two types of analyses. Inspection of these means (see Appendix 7, table I–XII) makes it clear that the differences are of such small magnitude as to be negligible. The two types of analysis may thus be considered equal as fas as underlying observations are concerned.

In all samples the correlation between the pretest and posttest is higher than between the aptitude test and the posttest. This is also reflected in the residual errors which are consistently less in the analyses utilizing the pretest as independent variable (see Appendix 7).

The result of the two series of analyses are summarized in table 37 below. We have indicated those cases where an interaction effect is interpretable.

Table 37. Interpretable interaction effects (x) in the analyses of variance (two-way classification). Dependent variable: Posttest.

	Independent variable: DBA scores*)	Independent variable: Pretest scores
GUME 1	–	–
GUME 2	–	–
GUME 3	X	X
GUME 4	–	–
GUME 5	–	–
GUME A	–	–

*) in GUME A F-test (verbal).

With the exception of GUME 3 there is no evidence of interaction between treatment and ability, no matter whether the latter is defined as general scholastic aptitude (GUME 1–5), as a specific verbal ability (GUME A), or as ability to perform on our English tests. The interaction effects found in GUME 3 are somewhat difficult to interpret. The analysis utilizing the DBA test scores as independent variable gives, besides the interaction effect (F-ratio 3.020;df= 4/200), a significant column effect indicating that the Es method may be preferred at all levels. The interaction effect is thus explained by the different effects of Im and Ee at particularly two of the ability levels; at the upper level Ee is ahead of Im whereas the opposite is the case at the middle and, to a lesser extent, the lowest level (see Appendix 7, table IX). In the parallel analysis utilizing pretest scores as independent variable the tendency is for the Explicit methods to excel at the upper level of ability and for the Implicit method to be superior at the lowest level (F-ratio 2.733;df = 4/218; see Appendix 7, table III, for cell means). However, in GUME 5 where the interaction term is only slightly below significance (F-ratio 2.314;df = 4/378; F_{crit} = 2.39), the tendency towards interaction is in the opposite direction (Appendix 7, table V). Considering the fact that the same grammatical structure was taught in GUME 3 and GUME 5, it is difficult to interpret these contrasting interactions. Considering further that the two parallel analyses in GUME 3 gave partly different results, it seems impossible to draw meaningful conclusions about the interactions appearing in our data.

At the end of the preceding section (p. 154) we presentedsome evidence of interaction in GUME 4, the youngest sample. Although none of the methods produced any differential effects on the treatment means the Es method appeared to increase the group heterogeneity from pre- to posttest more than the Im method. This finding suggests that a method utilizing explanations in the native language favours the more able students and puts a handicap on pupils of low ability. However, GUME 4 is the only experiment where the Es method proves to increase the group variance most (cf table 35, p. 154) and the result should be interpreted cautiously. One cannot rule out the possibility, though, that the finding applies to younger age levels, GUME 4 being the youngest group.

As was pointed out earlier (p 66) significant aptitude treatment interactions are fairly exceptional when the personological variable is of a general character. In the majority of our analyses no significant interactions were obtained which, at least in part, may be attributed to the fact that our personological variables are generel in nature (DBA and pretest). Our experiments were not planned with the ATI concept in mind, but we thought it might be worth the effort to investigate the interactions, having the proper kind of data available.

Two further comments will be made with respect to interactions in the

adult sample. On page 109 we stated that we should investigate the interaction between the age factor and the dependent variable, i.e. the posttest. Although the Es sample contained significantly older subjects we judged it improbable that the age factor would influence the treatment comparisons, the correlation between age and the dependent variable being only −.197. The details of the analysis are found in Appendix 7, table XIII. There is no evidence of interaction between age and the posttest, the F-ratio being 0.019 (df = 2/119). A significant row effect is obatined (F-ratio 3.338; df 2/119; $p < .05$); the younger the subject, the better the posttest result. The superiority of Es over Im is demonstrated at each age level (within the limits of the samle).

Similarly, on page 110 we stated that investigation of interaction between sex and the dependent variable should be undertaken in order to find out if the two methods did have any differential effect on females and males. The details of this analysis are given in Appendix 7, table XIV. The F-ratio for the interaction term is 0.120, indicating no interaction. (As a further check on the sex variable two analyses of covariance, one for each sex, were performed with the posttest as dependent variable and the pretest as covariate. Both analyses gave results similar to those presented previously for the whole sampe, i.e. the superiority of the Es method was demonstrated irrespective of sex; see Appendix 8).

Pupil reactions to the teaching strategies

In this section the pupils' attitudes towards the various teaching methods will be presented. In each questionnaire a number of fixed-response items supposed to measure the pupils' general attitude towards the treatments are utilized for statistical comparisons. For reasons of space the separate items will not be discussed except when the item result is contrary to the test in general. The pupils' spontaneous reactions as they appeared in the open answers will only receive brief mention.

In GUME 1–5 differences between the teaching methods as far as pupil attitudes are concerned will be tested by the Kruskal-Wallis one-way analysis of variance by ranks. The H statistic used in the test is distributed as chi square with df = k - 1, i.e. in all our cases df = 2. Each of the observations are replaced by ranks; in table 38 below the middle rank (MR) for each teaching method is given.

Significant differences are found in six of the nine analyses. In GUME 1 the Ee group ranks last in both sk and ak; in the latter course the pupils are also less positive towards Es than towards Im. A reasonable interpretation of

Table 38. Kruskal-Wallis one-way analysis of variance by ranks of pupils' attitude scores.

	Im		Ee		Es		H-value (after corr. for ties)	sign.	Atti-tude mean	Theor. mean
	N	MR	N	MR	N	MR				
GUME 1 sk	59	113.13	71	86.81	74	109.08	7.89	.02	25.47	24.5
GUME 1 ak	17	59.44	36	33.33	30	42.52	13.64	.01	26.02	24.5
GUME 2 sk	76	99.04	86	126.27	61	108.03	7.54	.05	28.92	24.5
GUME 2 ak	27	38.48	33	40.89	21	44.40	0.75	–	29.05	24.5
GUME 3 sk	47	72.19	56	78.91	51	80.84	1.01	–	25.06	24.5
GUME 3 ak	15	21.33	20	27.80	20	33.20	4.73	–	25.76	24:5
GUME 4	166	260.47	171	286.11	187	242.71	7.42	.05	22.98	21.0
GUME 5 sk	64	118.81	73	93.68	63	89.79	9.63	.01	22.55	22.0
GUME 5 ak	48	74.75	42	56.32	50	78.33	7.52	.05	21.34	22.0

the negative attitudes to Ee seems to be that the specific type of explanations used (cf pp. 116–117) caused consternation when presented in the foreign language. If the absolute attitude values for the Ee groups are considered (see Appendix 2, table I and II), it may be stated that the Ee method was tolerated, but only just. The general result is reflected in the separate items with one exception, item number 9 (the reading texts). In both the sk and ak course the Ee group did not rank last (Im = Ee = Es). This result probably indicates that the Ee students looked upon the reading activities as relatively relaxing.

In GUME 2, where the pupils' attitudes towards the treatments are fairly sympathetic on the whole, no differences are found in the ak group. This holds for the composite test as well as for each individual item. In the sk group the Ee method ranks first and Im last. This result might be taken to indicate that among the more advanced students the oral drills were either considered too monotonous or too frequent. This is not the case, however. The particular item (number 7) asking the pupils' opinions on the oral drills ranks Im > Ee > Es (H:7.25). Thus, the generally negative attitudes of the Im pupils are counterbalanced by a positive acceptance of the oral drills.

In GUME 3 the teaching strategies do not differ with respect to attractiveness as it is measured by the composite test. However, the total score conceals certain differences at the item level. For instance, in the sk group the Es pupils think they learnt relatively more (than did the Im and Ee groups) during the project lessons (item no. 1). On the other hand the Ee pupils consider themselves less tired after the lessons (item no. 4). Although these responses may reflect true opinions, it is also probable that they reveal some inconsistency in the pupils' answers to the questionnaire. In GUME 3 ak there is a clear tendency for the two E groups to be more positive than the Im group towards the technical arrangements and the drills (items nos 5, 6, 7, 8).

One fixed-response item in the questionnaire asked whether the explanations made it easier or not to understand the lesson content (5-point scale: *much easier, somewhat easier, no difference, somewhat more difficult, much more difficult).* Only the E-classes were instructed to respond to this item. In spite of this two Im classes in GUME I answered the question; they rather liked the (non-existent) explanations. No statistical comparisons were made between the Ee and Es groups; the following table gives the attitude means to make possible a rough comparison.

Table 39 Attitude means in GUME 1–3, the E groups, on one item measuring attitude towards the explanations (5-point scale)

	Ee		Es	
	N	$\bar{X}$	N	$\bar{X}$
GUME 1 sk	105	3.36	82	3.44
GUME 1 ak	48	3.44	43	3.51
GUME 2 sk	95	3.96	70	3.84
GUME 2 ak	46	4.02	28	3.94
GUME 3 sk	75	3.76	46	3.22
GUME 3 ak	71	3.85	39	3.80

No systematic differences between the methods appear in the figures. The only discernible tendency is that the ak groups in all cases are slightly more positive than the corresponding sk groups. The means may be taken to indicate that all groups inclined towards a positive attitude to the explanations, all being above the mean, 3.0. Whether the Im pupils felt frustrated because nothing was explained is still an open question. This information is not obtained from the attitude test.

The part of each of the GUME 1–3 questionnaires that consists of open answers covers a wide range of opinions from extremely positive to extremely negative. As might be expected, diverging opinions are expressed on one and the same feature. For instance, some pupils complained that the earphones hurt them and that the sound was bad, others declared that the earphones were the most exciting thing about the whole project. Although it is difficult to see any trend in the open answers, the following may be tentatively stated: The music, the inserted songs, the fact that no homework was given, and the greater variety of the instruction during the experiment were appreciated. Those who were negative complained that this type of instruction gave them no opportunity to ask questions and that the tempo during the lessons was too high. The lenght and the general character of the open answers in some of the classes indicate that the pupils have been influenced by each other and

possibly also by the teacher. This being so, we shall not utilize them for method comparisons.

In GUME 4, where a significant H-value is obtained, the Ee method ranks first and Es last. At the item level there is one deviation from this general pattern: the Im pupils *feel* they learnt less grammar than the others (item number 5). One possible explanation of this may be that the Im pupils, being given no explanations, simply were not conscious of the fact that they were taught grammar. The comparatively more positive attitude on the part of the Ee pupils is difficult to explain. Their teachers (see below) declared that they mostly gave explanations in the mother tongue during ordinary lessons; perhaps the Es pupils, although appreciating explanations, felt it less relevant to get them in Swedish.

Four items in the questionnaire concerned the explanations given (intended for the Ee and Es groups) and not given (Im). Almost 25% of the Im pupils think they have got explanations and 15% and 6% respectively in Ee and Es think they have not got any; incidentally, the Im pupils have a very positive attitude towards the explanations (which they never got). The results indicate that the pupils' responses to these questions are not wholly dependable. No differences are found between the Ee and Es groups as fas as their views on the explanations are concerned.

In GUME 5 the sk pupils have a neutral attitude towards the lessons whereas the ak pupils incline towards the negative. Six lessons devoted to one single grammatical structure is apparently very demanding on the less apt pupils at this age level. In the sk group the Im pupils are positive towards the lessons as a whole; the Ee and Es pupils seem to have tolerated them, but not more. In the ak group the Ee pupils are decidedly negative towards the lessons; the Es method ranks first, which is in accordance with the tendency found earlier for learning effects.

One item asked the pupils' opinions of the explanations. The means for the Ee and Es groups are 3.60 and 3.71 respectively (5-point scale). Both groups thus think that the explanations facilitated understanding.

As was the case in GUME 1–3, the open answers do not lend themselves to comparisons between the teaching methods. The following things were mentioned as positive features: The songs and music; no home work; more fun, more change; possible to check oneself; the funny stories. The negative answers fall in these categories: dull, too slow, too long pauses; too much repetition, harping; just listening to a tape-recorder; no teacher.

In the adult sample differences between the two methods with respect to student attitudes were compared by the Mann-Whitney-U-test. A statistical difference in favor of the Es group was obtained; the z-value is 5.405 ($p < .01$). All items pointed in the same direction as the total test. Thus, in GUME A the method that produced better learning also induced more favor-

able attitudes. Or perhaps, the more favorable attitudes – whatever caused them – produced better learning. However, the cause-effect relation problem, which of course can not be solved here, is of no interest in this case; the favorable attitudes may be looked upon as part of the method.

A few open answers were included in the test. However, the students' comments were ordinarily very sparse and can not be utilized for method comparisons.

In sum, the results of the attitude tests at the comprehensive school level display a moderate correspondence with the learning outcomes. The two most noteworthy exceptions are GUME 2 sk and GUME 4 where the attitudes of the Ee groups were most favorable despite the fact that they did not learn more than the other groups. The popularity of the Ee method in these cases is somewhat difficult to interpret. The pupils generally appreciate being given explanations, but they seem to feel that explanations in the mother tongue are less relevant. Thus, in GUME 1–5 there is no consistent pattern in the data referring to pupil attitudes. Conclusions concerning the value of the three teaching methods should not be based on their respective attractiveness as expressed by the pupils.

At the adult level the superiority of Es in learning effects is supported by favorable attitudes. It is perhaps not astonishing that students as mature as those in GUME A feel more condfident with a method that bears a certain resemblance with the one they presumably met earlier at school.

Responses to the teacher attitude test

The questionnaires administered in GUME 1–5 will be briefly commented on in this section. Each of them dealt with both teaching methods in general and with specific problems pertaining to the respective part project. Altogether 101 teachers were engaged in GUME 1–5; of these, 87 answered the questionnaire.

The first question to be presented asked the teachers to state their views on which teaching strategy would succeed best with pupils of good, average, or poor ability. (This particular item was included only in GUME 1–3 and 5). Not all teachers answered this question, and among those who did, some answered it inconsistently, not indicating any suitable teaching method for the medium and poor levels. However, the following table gives a rough indication of the teachers' opinions.

Table 40. Teacher responses to item: Which method suits which type of pupil?

	Im	Ee	Es	Tot
Good	8	24	32	64
Medium	5	7	43	55
Poor	20	1	32	53
Total	33	32	107	172

There is a majority for the Es method at all levels. The responses also reflect some belief in a structural drill method (Im) at the lowest level and in Ee at the uppermost. The latter method is considered rather useless for the less gifted pupils.

This question may be compared with two others. The teachers were to indicate, on a 4-point scale, whether explanations should be given. The answers among the teachers (N = 58) were: *every lesson* (10%), *fairly often and regularly* (64%), *once in a while* (26%), *never* (0%). It is perhaps mildly surprising that so many explain so often. One would have thought that the alternative *once in a while* more exactly fits the philosophy behind the directions in the curriculum (cf pp. 41-42). Incidentally, the opinions of the 27 teachers involved in GUME 4 on this question were distributed as follows: 4%, 37%, 59%, 0% (the percentages refer to the same options as presented above). As might be expected, the teachers at the medium stage (grade 6) of the comprehensive school explain grammatical features less frequently than the teachers at the upper stage (in so far as this item reflects actual classroom practice).

The second question with which the figures in table 40 might be compared is the following: *Should explanations be given in* (a) *Swedish* (b) *English?* On this item there was a strong majority in favor of explanations in Swedish (The teachers in GUME 4 vere not given this question).

Despite the fact that the participating teachers agreed not to discuss project matters with the pupils, teacher attitudes are a potential source of bias in the pupils' attitudes. However, the teachers' preference for Es is not reflected in the pupils' responses, as was shown above. When asked to state which of the three methods the teachers' own practice corresponded to, most of them pointed out that any method ought to be modified according to circumstances. Thus, the practices are not necessarily so Es-oriented as the answers to the fixed-response items indicate.

The teachers' comments on various technical aspects will be of great help in further research and development work. Considerations of space make it necessary to leave them out here, however.

CHAPTER 12

SOME ADDITIONAL FINDINGS

In this chapter we shall render an account of some complementary results. Although they do not always have a direct bearing on the main results reported hitherto, they may shed some light on various problems related to second-language learning.

Four follow-up studies

Throughout the teaching method comparisons we have used no control groups in a strict sense, i.e. groups being given pre- and posttest but no treatment *or* groups being given the posttest but neither pretest nor treatment. The first type of control, when applied, is intended to check whether the subjects were "sensitized" by the pretest and made inordinately large progress because of that; the latter type of control is a check on whether progress might be merely a function of maturation. There are three main reasons why we did not include any control group of this kind. First, the lesson content was very specific (in eight of the ten part projects consisting of one single grammatical structure) and it was taught intensively during a relatively limited amount of time. Secondly, in a class whose teacher does not concentrate on the same structure(s) during a corresponding period progress would in all likelihood be close to zero. Thirdly, we are not interested in the amount of raw progress made as such but only in the difference in progress brought about by different methods of teaching irrespective of how great or small this progress is. However, we have felt a need to compare the results obtained in our experiments with what is normally achieved at the same age level during a relatively long period of time. In Sweden von Mentzer (1970) has investigated how much pupils learn of English grammar during the 7th form. However, the results are very uncertain because two different tests were used and different classes were tested. In our study it has been possible to retest pupils one and two years after the initial test. As may be expected it proved impossible to locate a number of pupils after two years' time; however, despite the relatively large drop-out rate we think that some tentative

conclusions may be arrived at.

The GUME 1 and GUME 2 tests were chosen for the follow-up study. Before the start of the autumn term, 1969, tests, tapes and instructions were sent to the headmasters of the same schools that had participated in the original studies. They were asked to distribute them to teachers who were to teach a group of 7th grade English during the coming year. Twelve sk and six ak control classes were used; this corresponds to the proportions used in the original GUME 1 and GUME 2 studies. The tests were given on the very first lesson in the autumn term (cf the graph on p. 92). They were then collected and marked but the teachers were neither informed of the results nor of the fact that the tests were to be administered again at the end of the school year. At the end of May, 1970, the tests were given a second time. On this occasion tests from four sk and two ak of the GUME 2 follow-up classes could not be obtained; this means a 30% drop-out rate but the proportions sk/ak were maintained. In May, 1971, the third testing took place. Thus, the results are indications of how much the pupils learn of two specific grammatical structures (the do-construction and some/any) in one and two years' time respectively, when the teachers do not pay special attention to these particular structures. In the table below only pupils present on all three testing occasions have been included.

Table 41. GUME 1 and GUME 2 FOLLOW-UPS and original results.

	Original samples			FOLLOW-UP					
	Oct/Nov, 1968			August, 1969		May, 1970		May, 1971	
	N	$\bar{x}$	s	N	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
GUME 1 sk	227	71.07	16.23	223	71.25	15.42	81.41	17.85	88.08	18.01
GUME 1 ak	104	48.35	12.29	70	49.31	10.03	56.09	11.73	58.07	14.66
GUME 2 sk	247	64.77	17.16	107	58.19	18.78	77.23	18.67	85.92	19.15
GUME 2 ak	98	46.77	14.09	19	45.37	9.12	52.47	11.09	59.74	11.00

On the May -70 testing occasion the follow-up classes contained the following number of pupils: GUME 1 sk: 270, GUME 1 ak: 93, GUME 2 sk: 154, GUME 2 ak: 66. From table 41 above it is thus obvious that the loss of observations is great in GUME 2 whereas it is surprisingly low in GUME 1. In GUME 2 the greatest loss of observations occurred at the Maj -71 administration of the test (sk: 47, ak: 46).

In the follow-up groups the August -69 means should be of approximately the same magnitude as those of the original samples. The differences were tested and the following t-values were obtained:

Table 42. Tests of significance between original and August -69 means.

	Original mean	August -69 mean	t-value	
GUME 1 sk	71.07	71.25	− 0.12	
GUME 1 ak	48.35	49.31	− 0.56	
GUME 2 sk	64.77	58.19	+3.10	$p<.01$
GUME 2 ak	46.77	45.37	+0.55	

We notice that for GUME 1 the pretest means in the original project, which started about four weeks after the beginning of the term, are equal to those in the follow-up study. The do-construction has been dealt with in grades 5 and 6 and obviously very little happens in the first few weeks. In GUME 2 sk the pretest means in the project, which did not start until November, are higher than in the corresponding follow-up group. The difference may be explained by the fact that the some/any problem has not been dealt with systematically before grade 7, and here the sk pupils make some progress in the first two months. As may be expected, an equally large progress is not found in the ak sample. Thus the two different tests administered to different groups one year apart give approximately the same results, a fact which lends support to the progress comparisons performed.

Table 43. Progress (raw gain scores) in the original samples and in the follow-up groups.

	Original samples			FOLLOW-UP				
	1968: six lessons			Aug. -69 to May -70			May -70 to May -71	
	N	$\bar{x}$	s	N	$\bar{x}$	s	$\bar{x}$	s
GUME 1 sk	227	10.69	8.70	223	10.16	10.35	6.67	9.71
GUME 1 ak	104	3.31	8.08	70	6.77	8.48	1.99	9.80
GUME 2 sk	247	16.54	10.68	107	19.05	12.57	8.68	11.69
GUME 2 ak	98	12.85	13.57	19	7.11	14.78	7.26	13.50

In all ak groups the standard deviations exceed the means. That it to say, whatever progress is made in the less advanced course, the variation among subjects in this respect is great. A number of pupils regress, a fact which has been commented on previously (cf p. 137). The progress comparisons of main interest are those between the original project groups and the Aug. -69 to May -70 progress scores. In the following table these differences are tested for significance.

Table 44. Tests of significance between original and Aug. -69 to May -70 progress scores.

	Original progress	Aug.–69 to May–70 progress	t-value	
GUME 1 sk	10.69	10.16	+0.59	
GUME 1 ak	3.31	6.77	- 2.69	$p < .01$
GUME 2 sk	16.54	19.05	– 1.79	
GUME 2 ak	12.85	7.11	+1.57	

In GUME 1 the ak group made less progress than the control group. It should be noted that all progress scores for the project groups are means calculated over the three treatments. As has been pointed out earlier, the E pupils in GUME 1 ak made particularly low progress and were relatively more negative (than the Im pupils) towards their respective methods. The low mean of the GUME 1 ak sample is thus largely explained by the insignificant progress made in the two E groups. The Im mean (7.43), on the other hand, corresponds well to the Aug. -69 to May -70 mean. In all other cases the two progress scores compared are similar. It thus seems as if the pupils learnt as much in the six project lessons as they do otherwise in one year with respect to two grammatical structures, the do-construction and the some-any dichotomy. During the following year the progress in the control classes, with the exception of GUME 1 ak, tend to decrease somewhat. However, in the case of these progress measures the standard deviations highly exceed the means, indicating that the general trend is hazardous to interpret.

In figures 17 and 18 below the distributions of scores on the three testing occasions are presented for the GUME 1 and GUME 2 follow-up groups. In each distribution the arrow (solid or broken) indicates the group median. It should be observed that the y-axis is not the same for the GUME 1 and GUME 2 distributions; despite the apparent similarity in size between the two series of frequency distributions, those of GUME 1 refer to approximately double the number of observations.

The distributions demonstrate that, on all three testing occasions, very few pupils in the easier course exceed the median of the advanced course, whereas a larger number of pupils belonging to the advanced course have scores lower than the median of the easier course. In the following section we shall discuss similar findings in the project groups. The figures also illustrate the progress made in the follow-up groups over one and two years' time. The tendency among the follow-up groups is the same as within each of our experiments (cf. p. 136), i.e. the sk pupils make relatively greater progress. In the figures this is illustrated by the two arrows gliding further apart from the top to the bottom figure (in GUME 2, where the number of observations is very low in ak, no such tendency is discernible from May -70 to May -71).

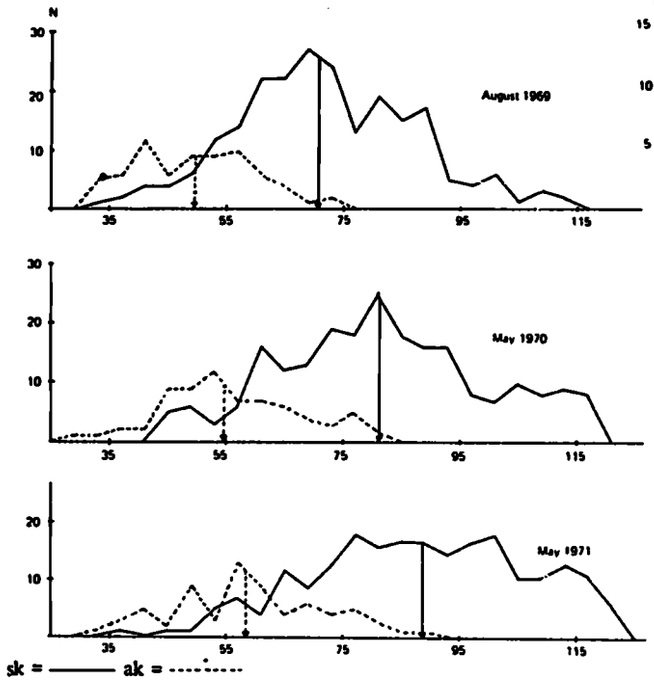


Figure 17. Distribution of scores on the criterion test, GUME 1 FOLLOW-UP. (N_{sk} = 223; N_{ak} = 70).

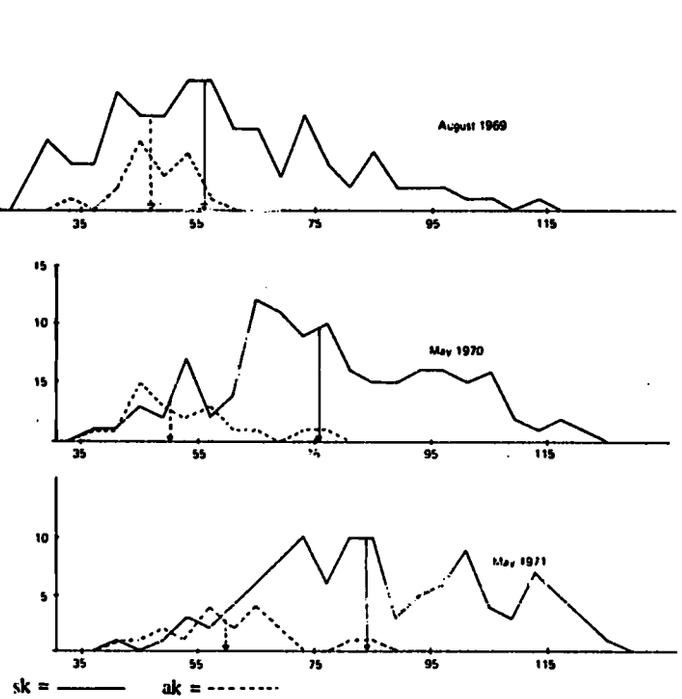


Figure 18. Distribution of scores on the criterion test, GUME 2 FOLLOW-UP. (N_{sk} = 107; N_{ak} = 19).

This "gliding-apart" effect probably reflects different learning capacities in the two groups. However, the hypothesis cannot be ruled out that part of the low progress in the ak group might be explained as an identification phenomenon; belonging to the easier course *causes* low motivation.

Findings related to choice of course

In the previous section some results related to the two courses in English were presented. We shall continue this discussion somewhat, although with respect to data from the project samples proper, i.e. GUME 1–5. It should be noted that the partition into two courses of the GUME 4 sample reflects the *choices* made by the pupils before the start of the upper stage (cf. p. 99), whereas, in the other samples, the actual courses will be dealt with. Of the various compulsory school subjects taught at the upper stage (grades 7 through 9), only English and Maths offer alternative courses. The pupil (and/or his parents) chooses course on his own. The curriculum is very explicit on this point: "Choice of course may be made even if it should conflict with the pupils' intellectual capacities, such as these are perceived by the school authorities. This means that there are no formal hindrances for admission to the different classes or courses. Nor can a pupil, even if his academic achievements are insignificant, be prevented from following a more theoretical course through school" (Lgr 69 p. 34). Despite the fact that the individual's choice of course is free in principle, there is obviously a substantial correlation between choice and social class. Table 6, p. 100, is a demonstration of this fact. In Sweden there has been some controversy over keeping two separate courses. The following findings may shed some light on the question.

In the figures on pp. 169–170 two series of frequency distributions (Pretest and DBA scores) for GUME 1–5 are presented. In each figure the sk and ak distributions are represented, the arrows indicating the medians for the two courses. The shadowed area to the left illustrates the proportion of sk pupils below the ak median, the corresponding area to the right shows the proportion of ak pupils above the sk median.

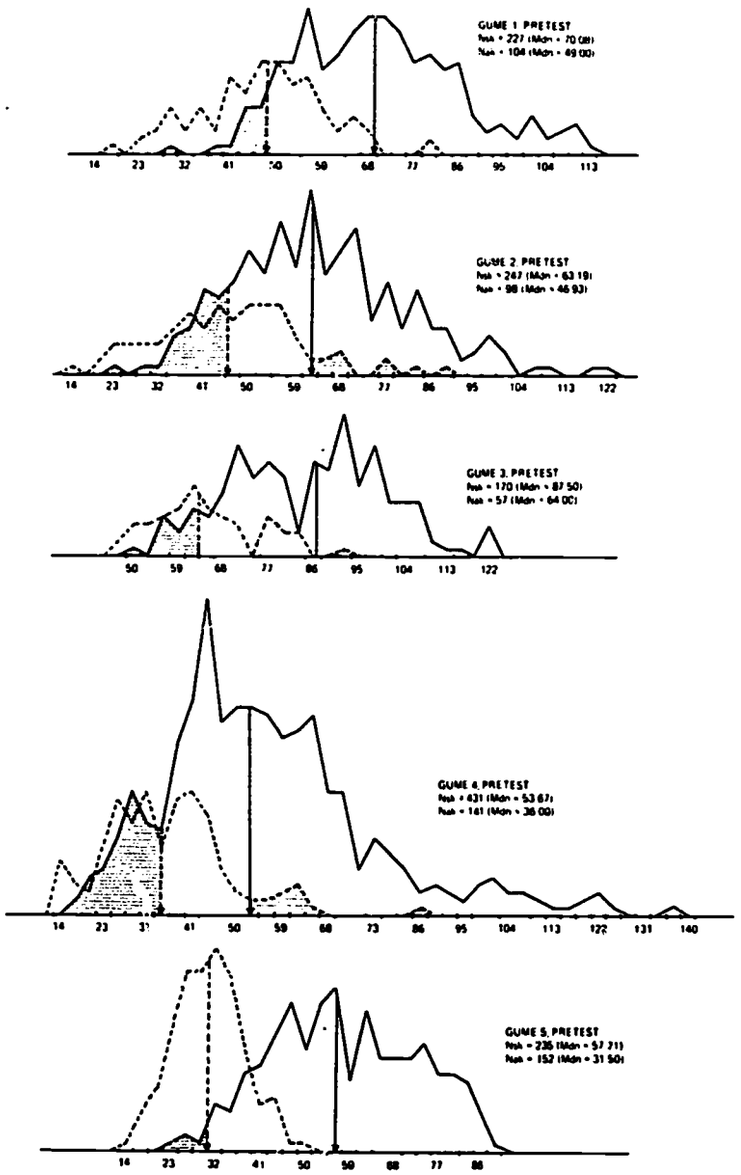


Fig. 19. Distribution of pretest scores in the GUME 1–5 samples.
(——————— = sk; - - - - - - - - - - - - - - = ak)

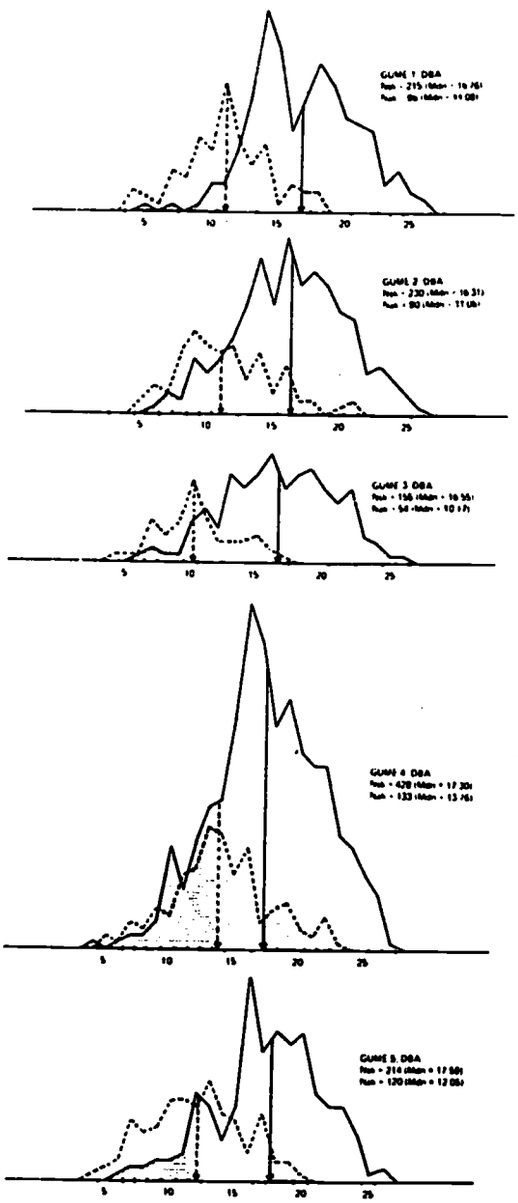


Fig. 20. Distribution of DBA scores in the GUME 1–5 samples.

(———————— = ak ---------- = ak)

The following observations can be made with respect to both pretest and DBA scores:

1. The medians of the two courses are clearly separated (= significantly different).
2. The range of scores in the advanced course is approximately equal to the total range (sk + ak).
3. There are relatively more sk pupils below the ak median than there are ak pupils above the sk median. The average percentages, calculated over all samples, are: Pretest: 9.1% and 4.2% and DBA: 11.9% and 9.3% respectively.
4. The most clear "gliding-apart effect" is found in GUME 5 pretest, represented by a bimodal distribution of scores.

Considering the obvious differences in ability and achievement between the sk and ak groups, it is apparent that teaching in the two courses ordinarily proceeds at different levels and different rates. It is of course hazardous to predict what consequences it would have to add the two courses together into one. However, considering the observations made above, we would venture the following hypotheses:

Lumping the courses together would not substantially change the range of ability in the total group as compared to the range prevalent in the advanced course. In the various distributions of pretest scores there are only 2% ak pupils on an average who have lower scores than the lowest sk score. In order to estimate the composition of a hypothetical composite (sk + ak) class, the following procedure has been adopted: Assuming that a composite class would consist of 25 pupils, we have located the score below which 4% of the weakest sk pupils would fall. This score indicates that in each composite class there would be one low-achieving sk pupil on the average. Similarly, we have calculated the percentage of ak pupils who fall below this critical score; this

Table 45. Number and percentage of ak pupils below the 4th percentile in the sk group.

	N_{ak} below 4th sk percentile (1)	$N_{sk + ak}$ (2)	(1) in % of (2) (3)
GUME 1	39	331	11.8
GUME 2	28	345	8.1
GUME 3	18	227	7.9
GUME 4	23	572	4.0
GUME 5	91	387	23.5

percentage is an illustration of how large a proportion of the ak pupils would be equal to or less successful than the weakest sk pupil in our hypothetical composite class.

The table presents an increase in percentages (column 3) from grade 6 to grade 8; grade 6 (GUME 4): 4%, grade 7 (GUME 1-3): around 10%, grade 8 (GUME 5): 23.5%. The grade 8 result is, again, an illustration of the fact that the two courses have glided apart during the relatively long period they have been taught differently. If the two courses were to be added together, this would take place at the beginning of grade 7; therefore the grade 7 results are of main interest in our case. The implication is thus that, in case the courses were lumped together, approximately 10% of the lowest-achieving ak pupils would be included in the same class as the weakest sk pupil. In sum, our hypothetical 7th grade sk + ak class would contain approximately 3 more pupils of equal or somewhat less ability then the weakest pupil in the original sk class. It is difficult to see that the problems of individualization would drastically increase because of this.

However, if the courses were added together it is reasonable to assume that the teaching would have to be adapted so as to fall somewhere between the previous ak and sk levels as far as difficulty and speed are concerned. One practical consequence of adding the courses together would be to save teacher hours. Existing ak classes are ordinarily small in size and thus consume a relatively large number of teacher hours per pupil. If, by this hypothetical step, teacher hours were saved, they might be used for giving both superior pupils and slow learners extra training, partly in the form of pre-produced programs, partly by the live teacher. It is difficult to foresee what effect adding the courses together would have on discipline; the negative effects, if any, ought to be evaluated against the fact that the risk of wrong choices, causing low motivation, is eliminated. By "wrong" is implied that able students choose the easier course and weaker students choose the advanced. The correlation between choice and social class alluded to previously indicates that the actual choices partly reflect social handicaps.

One may ask whether the difference between the sk and ak courses in pretest scores corresponds to an equally large difference in DBA, i.e. general scholastic aptitude. In so far as the former substantially exceeds the latter, this may be looked upon as support for the "identification hypothesis" put forward previously (p. 168); that is to say, the ak pupils would, in such a case, perform less well than might be expected on the basis of their general ability. In order to get a conception of these differences, we have adopted the following procedure: th sk-ak differences in pretest and DBA have been divided by their respective standard deviations, the latter being calculated for the composite (sk + ak) group. The ratio thus obtained indicates, in terms of standard deviations, how much the sk group exceeds the ak group. Finally, the ratios for pretest and DBA are compared. Table 46. Illustrates the computational steps.

Table 46. Differences between the sk and ak groups with respect to Pretest and DBA.

	(1) Standard deviation in composite (sk + ak) group		(2) Difference sk -- ak mdn (raw scores)		(3) Difference sk -- ak mdn (SD's)		(4) Difference between (2) and (3)
	Pretest	DBA	Pretest	DBA	Pretest	DBA	
GUME 1	18.04	4.49	21.08	5.68	1.17	1.27	-0.10
GUME 2	18.24	4.48	16.26	5.25	0.89	1.17	-0.28
GUME 3	16.88	4.65	23.50	6.38	1.39	1.37	0.02
GUME 4	20.89	4.33	17.67	3.54	0.85	0.82	0.03
GUME 5	18.30	4.59	26.21	5.54	1.43	1.21	0.22

There is no systematic trend from grade 6 to grade 8; in GUME 4 (grade 6) the superiority of sk is equally large in pretest and DBA; in grade 7 (GUME 1--3) the figures vary, one group being similar to GUME 4 and the two other demonstrating a smaller sk-ak difference in the pretest than in DBA; in grade 8 (GUME 5) the tendency is for the sk-ak difference to be larger in the pretest than in DBA. Thus, in the oldest samle (GUME 5), where the pupils have been taught in separate courses for more than a year, the difference in pretest scores between the two courses is larger than might be accounted for by their respective general ability. The results should be interpreted cautiously, there being no systematic trend from grade to grade and the ak groups being of limited size. However, we think that the finding in GUME 5 suggests that the ak pupils, having identified themselves as low-achievers, do not work up to their capacity.

Basically, the problem touched on here is a covariance problem. It would have been theoretically possible to calculate, within each total sample, the sk as well as the ak pupils' pretest scores, utilizing their respective DBA scores as covariate. This kind of calculation would in all likelihood produce results similar to those just presented. However, considering the fact that the sk and ak groups are both selected, the covariance procedure would have been dubious, and we have refrained from it.

Some correlations

The ten correlation matrices, one for each experimental sample, are presented in Appendix 4, tables I – X. Here we will limit ourselves to commenting on some correlations, thereby comparing them over all experimental groups. The comments will mainly refer to the GUME 1–5 samles; a different set of concomitant variables was used in GUME A.

PROGRESS CORRELATIONS

As was pointed out previously (p 73), progress (defined as raw gain score) is practically always negatively correlated with initial, i.e. pretest, scores. In table 47 below the correlations between progress and some other variables are given. The N's in the table indicate the number of pupils for whom progress scores are available. In the Progress-Grades and Progress-DBA correlations these N's are, in a few instances, slightly lower. However, the actual N has been taken into consideration in each case when testing whether the correlations deviate from zero. The same procedure has been adopted in the subsequent tables of correlations in this section.

Table 47. Correlations between Progress (=raw gain score) and some other variables.

		Progress (= raw gain score) –				
	N	Pretest	Posttest	Grades English	Grades total	DBA total
GUME 1 sk	227	–.175	.345	.079	.222	.079
GUME 1 ak	104	–.239	.396	.037	.004	.026
GUME 2 sk	247	–.254	.356	.130	.134	.153
GUME 2 ak	98	–.335	.555	.222	.259	.231
GUME 3 sk	170	–.245	.388	.074	.107	.133
GUME 3 ak	57	–.317	.459	.091	–.028	–.144
GUME 4	574	.307	.688	.488	.434	.283
GUME 5 sk	235	–.356	.195	–.130	–.067	.087
GUME 5 ak	152	–.259	.587	.022	.073	.317
GUME A	125	–.019	.633	–	–	–

_____ = significant at the 5 % level ═══ = significant at the 1 % level

There is a clear tendency for the progress scores to be negatively correlated with pretest scores. In all cases except GUME 4 the relation $r_{12}s_2 > s_1$ holds (1 and 2 stand for pre- and posttest respectively; cf p. 73). The calculations are left out here but can easily be checked by the reader. Thus, within each group the pupils with low pretest scores tend to progress relatively more than pupils with high pretest scores. This finding cannot be explained as a ceiling effect (in the posttest) in the ordinary sense, since practically no pupils reached the maximum score on the posttest occasion. A more natural explanation is a general regression effect, caused by less than perfect reliability in the pretests. The pupils whose true scores were underestimated in the pretest and, similarly, the pupils whose true scores were overestimated, have regressed towards

their respective means on the posttest. A second and somewhat tentative explanation is that the less able students (within each course) gained insight into grammatical structures presented earlier during the course of ordinary teaching and therefore progressed more than the more able students who possessed some knowledge at the start of the experiment. The exception from the general pattern of negative correlations between progress and pretest scores is GUME 4. This project was performed in the 6th form, i.e. at a stage where the pupils, both according to the curriculum and the commonly used textbooks, have not yet met several of the structures taught. At this comparatively early age the more able students (as defined by the pretest scores) progress more rapidly than the less able when faced with new learning materials.

In all experiments except GUME 4 the correlations between progress on one hand, and Grades English, Grades total, and DBA on the other tend to be between zero and slightly positive. That *increase* in learning correlates only moderately with other variables, follows from the fact that the individuals' *relative standing* in their group does not alter with practice, which in turn follows from the high correlations between pre- and posttests (cf, for instance, Anastasi 1958. p 195).

CORRELATIONS BETWEEN THE CRITERION TEST AND OTHER VARIABLES

In table 48 below the correlations between the criterion test and some variables are given. Since the correlation between the pre- and posttests are ordinarily high, their respective correlations with other variables present a similar pattern; we will therefore only give the pretest correlations.

Table 48. Correlations between the Pretest and some other variables.

		Correlations between Pretest and:							
	N	Grades English	Grades Swedish	Grades Maths	Grades Total	DBA Verbal	DBA Ind	DBA Spat	DBA Total
GUME 1 sk	227	.679	.502	.328	.588	.462	.301	.162	.415
GUME 1 ak	104	.572	.267	.242	.439	.280	.301	(.114)	.349
GUME 2 sk	247	.518	.469	.254	.475	.452	.325	.182	.419
GUME 2 ak	98	.455	.329	.284	.440	.308	.236	(.098)	.285
GUME 3 sk	170	.628	.616	.457	.658	.542	.331	.220	.483
GUME 3 ak	57	.659	.409	.267	.554	(.239)	.430	.265	.506
GUME 4	574	.697	.633	.492	.682	.586	.408	.220	.522
GUME 5 sk	235	.721	.571	.481	.675	.551	.276	.179	.452
GUME 5 ak	152	.425	.324	.420	.510	.366	(.127)	(.066)	.219

Correlations within brackets do not deviate from zero. The remaining correlations do, in the majority of cases at the 1 % level.

The different variables included in each study were intended to provide background information about the subjects and to be potential covariates in the analyses of covariance. There were no factor-analytic considerations behind our choice of variables, and the studies are not designed so as to elucidate what factors constitute foreign language learning ability. However, the general pattern of correlations in table 48 seems to warrant the following observations:

The pretest-grades correlations are, with one exception (GUME 5 ak) of the order Grades English > Grades Swedish > Grades Maths, which lends support to the validity of the criterion test. However, it is apparent that the differences between the Grades English and Grades Swedish correlations are not substantial, nor are the differences between the Grades English and Grades total correlations; in the latter case the Grades total correlations are higher in two instances. Taken together, these observations seem to indicate that it is difficult to devise foreign-language tests without measuring a more general, supposedly verbal, scholastic ability factor at the same time. A similar picture is obtained in the pretest-DBA correlations, where the pretest-DBA verbal correlations are of about the same magnitude as the pretest-DBA total correlations.

ATTITUDE CORRELATIONS

As was illustrated in the preceding chapter, no clearcut relationships were found between attitudes and teaching methods at the comprehensive school

Table 49. Correlations between the Attitude test and some other variables.

		Correlation between Attitude test and –							
	N	Pre-test	Post-test	Pro-gress	A/P progr.	Grades Engl.	Grades total	Verbal DBA	DBA total
GUME 1 sk	204	–.108	–.121	–.036	–.069	–.115	–.131	–.100	–.023
GUME 1 ak	83	.201	–.039	.251	.202	–.147	–.163	–.256	–.179
GUME 2 sk	223	–.030	–.027	.005	.005	–.004	–.100	–.026	–.095
GUME 2 ak	81	.083	.071	–.005	.025	.052	.020	–.206	–.153
GUME 3 sk	154	–.106	–.044	.090	.082	–.049	–.019	–.198	–.148
GUME 3 ak	55	.090	.112	.032	.039	.048	–.063	–.116	–.235
GUME 4	529	.181	.264	.258	.267	.206	.171	.105	.097
GUME 5 sk	200	–.171	–.094	.158	.126	–.167	–.094	.024	.085
GUME 5 ak	140	–.064	.054	.134	.142	–.078	–.012	.091	.123
GUME A	119	.045	.163	.201	.199				

_____ = significant at the 5 % level ═══ = significant at the 1 % level

level. At the adult level, on the other hand, the method producing better learning was associated with more positive attitudes towards it. In table 49 on the preceding page the correlations between attitude scores and some variables, irrespective of teaching method, are presented.

In GUME 4 the correlations between attitudes and all variables are statistically significant. It is, again, impossible to conclude whether awareness of making progress produced favorable attitudes or whether more positive attitudes caused better learning. The significant, although low, correlations between attitudes and the cognitive variables indicate that, in the youngest of our samples (grade 6), the more able pupils tend to be comparatively positive towards the experiment. In the other, an older, samples at the comprehensive school level no such tendency is discernible. The general tendency is for the attitudes to correlate zero or, in a limited number of instances, slightly negatively, with progress as well as the remaining variables. This finding is in accordance with, for instance, the results obtained in the Pennsylvania study (cf p. 56).

In the adult sample the positive correlations between attitudes and progress corroborate the general finding that the superior method is associated with more favorable attitudes.

SOCIAL CLASS CORRELATIONS

In Appendix 4, tables I - X, the correlations between social class and all other variables are given for each sample. In all cases they are product-moment correlations, which is somewhat dubious considering the underlying social class scale. For each total sample (sk + ak) at the compulsory school level we have therefore computed point-biserial correlations between social class and some variables. In these cases the dichotomous variable has been obtained by adding social class 1 and 2 into one category whereas social class 3 represents the other category. The correlations are presented in the table below.

Table 50. Point-biseral correlations between Social class and some other variables.

	N	Pre-test	Post-test	Pro-gress	Atti-tudes	Grades Total	DBA Total
GUME 1	306	.365	.399	.209	−.102	.400	.352
GUME 2	282	.185	.224	.078	−.051	.217	.200
GUME 3	210	.216	.247	.082	−.062	.257	.258
GUME 4	494	.168	.156	.066	.087	.183	.129
GUME 5	319	.282	.314	.127	.102	.263	.308

The correlations between social class and the various cognitive variables are all statistically significant at the 1 % level; the magnitude of them corresponds to what is ordinarily found in similar studies. The progress scores, and particularly the attitude scores, appear to be unrelated to social class.

CHAPTER 13

DICUSSION AND CONCLUSIONS

Earlier in the present report it was stated that we regard the GUME project as a conclusion-oriented undertaking. Our series of investigations are intended to shed some light on the tenability of two opposing foreign-language learning theories, the audio-lingual habit theory and the cognitive code-learning theory. The studies are thus not intended to compare distinct foreign-language teaching programs and materials and to lead to recommendations, in view of the research results, about which one to use. Our review of the research literature demonstrated that the two theories have undergone revision and modification and even that the theoretical conflict is considered apparent rather than real in some quarters. However this may be, it is obvious from current discourse in methodogical matters that both theories, even in their most uncompromising form, still have staunch defenders. There is thus no reason to believe that our results have become outdated because the Swedish debate, at least in its more violent form, has abated.

The most clearcut finding in our experiments concerns the adult sample where the Explicit group excelled in all treatment comparisons. The members of the adult group varied in age from 17 to 60 and had, with two exceptions, no academic training beyond the compulsory school level. Although the results should not be generalized to other adult groups of a different composition, they support the majority of hypotheses forwarded in respect of linguistically mature groups. That is to say, they suggest that explanations in the mother tongue clarifying linguistic patterns are efficient in internalizing the English grammar even when supplied at the expense of practice. They also suggest that pattern drills are of limited value as long as the adult has not been provided insight into the structure of language. Since our adult group is very heterogeneous in age, we investigated if age tended to interact with teaching method. No such interaction was found; the Es method is superior at all three age levels investigated. The younger subjects (–25) achieved significantly higher than the older subjects (41–). No hypothesis was formulated in this respect; however, the result seems to make sense in the foreign-language learning area.

Thus, the contention that the mediational role of the native language should be utilized in teaching adults seems to have much to recommend it (cf

Ausubel 1964). Ellegård (1971) has pointed out that the adult who is learning the syntactic and phonetic structures of a foreign language has to do this consciously – in so far as they differ from those of his native language. A common result of indadequate learning in natural situations is so-called pidgin language, i.e. the adult uses the vocabulary of the foreign language adapted to the phonetic patterns of the native language, while the syntax is mainly reduced to what is common to the native and the foreign language (op. cit., p. 122). In our experiments, where the adults were exposed to formal training, it is apparent that the group which did not have recourse to the native language as a mediating link learnt less.

At the comprehensive shool levels investigated, i.e. grades 6, 7, and 8, the pupils being approximately 13,14, and 15 years of age, the results are not as conclusive as at the adult level. The pupils belonging to the easier course generally made very little progress during the experiments, irrespective of teaching method assigned to them. In fact, the amount of learning was so small as to minimize the probability of obtaining differences between treatments, if such exist. The small progress in the easier course may be due to the fact that the teaching material presented was a compromise between what might be considered optimal in each course; the difficulty level apparently gravitated more towards the sk than the ak standards. Somewhat surprisingly, this was not reflected in the attitude tests; the ak students, despite their relatively insignificant progress, were slightly more positive than the sk students in the majority of cases. Although the teaching procedures may be accepted for the main purpose of our investigations, i.e. testing whether explanations facilitate learning, the ak pupils were probably relatively more handicapped by the speed and the somewhat unrealistic situation with taped lessons. All in all, if the teaching materials in the ak groups had been more adequately adapted to the ability of the pupils, our experiments would probably have been more promising in so far as detecting treatment differences is concerned.

In the more advanced courses, where the progress is more substantial, the overall pattern of results suggests a certain rank ordering of the teaching strategies, namely Es > Im > Ee.We should prefer to discuss this finding in terms of convincingness rather than in terms of statistical proof. The large amount of error variance in our data implies that true differences between treatments tend to escape detection. In no single sk group do we find a significant difference between treatments but, considered as a whole, the various experiments display a consistent pattern. The results in the sk group thus show a certain correspondence with the results at the adult level. The concept of linguistic maturity has been commented on previously; it makes sense to believe that the pupils at the comprehensive school level, being 13 to 15 years of age, are relatively mature linguistically and thus fairly similar to

the adult group in this respect. If so, the results in the sk group suggest that, at the upper stage of the comprehensive school, explanations in the native language tend to facilitate understanding. In view of this the following paragraph in Lgr 69: II Eng, seems somewhat overstressed: "Every grammatical rule *must* (italics ours) be formulated with English as the starting-point. This means that, when grammatical items are discussed in the native language, which may be judged necessary in exceptional cases, Swedish usage should not be compared with English, but the discussion should solely take the English structure as its starting-point" (p. 14).

Considering the fact that the recommendations of the curriculum are intended to specify, in explicit terms, its general objectives (which are, admittedly, liberal in nature; cf. p. 42), categorical statements of the kind quoted become more questionable the more unsupported they are by research findings. We contend that teachers of English, rightly or wrongly, would feel the Im and Ee methods to be in accordance with the intentions of the curriculum while the Es method would be considered alien to them. There is obviously a serious undertone in the humorous phrase "Not a word of Swedish in my lessons", a phrase appearing now and then in the Swedish debate and intended, we believe, to reflect the presumed intentions of the curriculum.

The tendency towards superiority of the Es method should not be generalized to explanations in a general sense. Too abstract or otherwise inadequate explanations would simply be a waste of the students' time. In our samples at the comprehensive school level, where the lessons had to be a compromise between what might be considered optimal for the two courses, the hypothesis cannot be ruled out that the explanations were beyond the comprehension of some of the low-achieving students.

One important difference between the Implicit and Explicit procedures at the comprehensive school and adult levels should be kept in mind. At the former level the Im and E strategies were similar to a fairly large extent, the only difference being that, in the E groups, a number of drills were replaced by explanations each lesson. At the adult level, on the other hand, the two methods were rather extreme in that certain techniques were excluded from each method: habit-forming procedures from the Explicit method and formal grammar from the Implicit method. It is thus impossible to conclude whether the somewhat different results obtained at the ordinary school and adult levels depend on differences between materials or whether different learning strategies are used by adult and teenager. A more conclusive cross-validation for isolating these effects would be to offer the adult materials at the comprehensive school level and vice versa. Incidentally, this type of further investigation is being planned at the moment. However, the similarities between the procedures at the different levels probably outweigh the differences, and our hypothesis is that essentially similar results will be obtained.

A word of caution is in order about the risk of drawing too far-reaching conclusions from our findings. The results should not be used for rejecting certain foreign-language teaching *methods*. We simply did not compare complete, or global, methods, but rather, specific procedures or techniques related to two theories or learning models. Much of the debate in the last few years has been beset with the inadequacy of treating methods in a global and diffuse way, often without any attempt at defining aspects such as the age of the learner, the amount of language already mastered, possible differences between different languages from a learning point of view, etc. Similarly, it was not always stated whether the matter of dispute concerned grammar, vocabulary, reading, listening, etc. Our experiments have consistently concentrated on the learning of grammatical structures, and the results should not be generalized to other aspects of language learning. It is intersting to note, in this connection, that Rivers (1968), in discussing the "two levels of language", the level of manipulation of language elements and the level of expression of personal meaning, states that "it is clear that one type of teaching will not be sufficient for the task" (p. 72). Her hypothesis is that a habit-forming mechanistic model is adequate for the more elementary level, the handling of rule-governed aspects of, for instance, accidence, i.e. forms of nouns and verbs in certain positions. For the higher, more intellectually demanding level, requiring the subject to choose structures and vocabulary in expressing his thoughts, Rivers proposes a more cognitive, insightful approach. The particular aspects taught in our lessons mainly belong to the level of expression. This being so, the tendency towards superiority for Es seems logical. However, if our studies had been performed at younger age levels, which in turn had necessitated lesson contents and testing procedures more in line with the level of manipulation of language elements, it is very probable that the method comparisons would have yielded results different from the present ones. Again, our results should be considered in relation to the various conditions under which they were obtained.

A most regrettable consequence of misinterpreting our results would be to suggest a return to more traditional foreign-language teaching methods. None of our Explicit methods were of this kind. Although rules were presented to the adult Explicit group, grammar was never taught as an end in itself, but was always followed by exercises containing common every-day sentences, giving the learners the opportunity of immediate application of rules. Oral activities were also part of the method. However, the Explicit method used in the adult sample was admittedly more traditional in character than the Explicit methods at the comprehensive school level where, besides the explanations, numerous drills and oral exercises were included. Thus, our Explicit methods bear no resemblance to an old grammar-translation method with little or no conversation and a lot of rule-cramming.

Although our research design was steered towards searching main treatment effects, we have investigated whether interaction occurred between the various treatments on the hand and various levels of "intelligence" and achievement on the other. Thus, our research may be said to represent a compromise between the traditional comparative study and the interactionist approach which leaves no place for the traditional questions of educational research, such as "What is *the* best foreign-language teaching method?" (cf Cronbach & Snow, 1969, pp. 10-11). Our project was not planned with any subtle aptitude-interaction hypothesis in mind, but possessing suitable data, we have calculated the various interactions mentioned above. A few significant interaction terms appeared; however, they proved somewhat inconsistent and difficult to interpret. Earlier research has shown that aptitude-treatment interactions are generally rare, and exceptionally rare when the personological variable is complex in factor structure. Since our personological variables are of exactly this kind, it is not surprising that no clear interactions were found. It is still an inspiring task for researchers to develop foreign-language teaching methods hypothetically related to specific variables, and search for interactions.

We shall briefly comment on some considerations made in connection with the data treatment. Throughout our studies we have compared a number of statistical techniques, the respective values of which have been frequently discussed in the research literature. More particularly, we have compared three varieties of computations: analysis of covariance, the treatment x levels design, and analysis of variance of raw gain scores. The latter technique is ordinarily warned against because of the unsatisfactory properties of difference scores. In our data, which we considered suitable for an empirical comparison between the three techniques, the analyses of covariance and variance of difference scores brought about the same results in nine comparisons out of ten. Tre treatment x levels designs, being based on the total samples and not each of the sk and ak courses, generally coincide with the previous analyses. The great symmetry in results is probably best explained by the high correlation between pre- and posttests and the high reliabilities of the tests.

In experiments where the intact school is the sampling unit, sampling errors will occur in so far as the classes are more homogeneous than the population from which they are drawn. We have not calculated any intra-class correlations and thus have no precise measure of school class homogeneity. Calculations of main effects were made with the individual as the unit of analysis and, in cases where the number of observations was judged sufficiently large to permit this, with the school class means as the unit of analysis. In all cases the two types of calculation gave similar results. We think there is reason to believe that our results, being the same irrespective of method of computation and unit of analysis, are dependable.

Comparisons between the treatment groups with respect to a number of background variables revealed no systematic differences between the groups, which may be taken as an indication that the internal validity of the experiments is satisfactory.

As is always the case in broad field studies, full control over the experimental situation is lacking; our investigations admittedly do not meet the requirements of laboratory research. A hypothetical list of sources of invalidity might include such things as vague instructions to participating teachers and pupils, malfunctioning of technical equipment, changes in experimental shedule which might have been foreseen, variations in listening conditions between classrooms; indeed there are numerous potential causes of irrelevant influence. However, since our investigations may be looked upon as a series of mainly independent replications, it is very probable that extraneous factors of the kind mentioned have cancelled out.

In comparative research uncontrolled variance attributable to differences in teacher behavior has often obscured the findings. In all our experiments we have used tape-recorded lessons and preproduced material in order to achieve strict control over the stimulus situation and in order to eliminate the teacher as a source of error.

A few comments on our findings besides the main treatment effects is in order. At the comprehensive school level, where the treatment differences in terms of means are small or non-existent, we investigated whether the various teaching methods brought about differences in group variances from pre- to posttest. With one exception (GUME 4, Es > Im) the general picture is one of no differential effect on group heterogeneity. The duration of our experiments is probably too short for such differences, assuming they exist, to occur. The result in GUME 4 indicates that, in our youngest sample (grade 6), the method utilizing explanations in the mother tongue tends fo favour the more able students and put a handicap on the less able (as defined by pretest scores).

The responses to the pupil attitude test bear no clear relation to the learning outcomes at the comprehensive scool level. A few items pertaining to the explanations revealed that a number of pupils have no clear conception of whether they received explanations or not. This finding is perhaps not so surprising as it may seem; from the students' point of view it is simply " an English lesson going on". In the majority of cases the pupils appeared to have neutral or slightly positive attitudes to the lesson series. In only one case, GUME 5 ak, is the general attitude slightly negative; the pupils belonging to the easier course in the oldest group (grade 8) apparently find difficulty in enjoying grammatical structures. At the adult level the subjects belonging to the Explicit method had the most positive attitudes. Regarding attitudes as part of a method, it becomes of little interest to speculate over whether the

favorable attitudes produced the better learning or whether the students' awareness of making progress produced the sympathetic attitudes. As we see it, the positive reactions lend further support to the superiority of the Explicit method at the adult level.

At the comprehensive school level the teachers' attitudes, as expressed in the questionnaire, are decidedly in favor of the Es method. When asked to predict the success of our three methods at three levels of pupil ability, the teachers rank Es first at all levels. Some belief in the Implicit method at the lowest level was evidenced, though. A large majority of the teachers hold that explanations should be given fairly often and regularly, and an equally strong majority favour explanations in Swedish. These attitudes seem to coincide with teacher opinions appearing in the foreign-language teaching debate (cf p. 43); they obviously do not coincide with the curriculum.

The criterion test utilized in GUME 1 and GUME 2 were administered to a number of control classes which were not otherwise involved in our experiments. The classes took the tests on three occasions extending over a period of two years. Although a substantial loss of observations occurred, it proving difficult to locate a number of pupils on the two later occasions, some tentative conclusions may be put forward. The control classes learnt as much during one year of ordinary teaching as did the experimental classes in six lessons. It should be remembered, though, that the control teachers did not concentrate on the specific structures investigated (the do-construction and some-any) but covered a lot more during the year. In the control classes it was also found that the means of the two courses, sk and ak, tended to glide further apart from the first testing occasion to the third. This observation, which is supported by a similar finding in our experimental samples proper, may be looked upon as a kind of identification phenomenon; we venture the hypothesis that belonging to the easier course contributes to low motivation and partly causes the pupils not to work up to their ability.

In the comprehensive school the pupils are free to choose the more advanced or the easier course in English. However, the correlation between social class and choice of course indicates that social factors are at work in the actual choice situation. The distributions of pretest as well as DBA ("intelligence") scores in each sample demonstrate that the variation in scores in the sk group is practically as large as the variation in the composite, i.e. sk + ak, group. Considering this fact and the hypothetical identification phenomenon mentioned in the preceding paragraph, we have ventured to speculate on the consequences of lumping the two courses together. We shall not repeat our arguments in favor of such a step (cf pp. 171-172); suffice it to say that we believe that experiments in this direction will yield valuable insight into materializing the concept of individualization within foreign-language teaching. We are well aware that the problems of individualization

might be rather unique in Sweden where 100% of the pupils take English for seven years. As a comparison it may be mentioned that, in the US, only about 2% take a foreign-language course of four years (cf Strasheim 1970). However, keeping two different courses in the subject of English, which is perhaps the most prominent school subject for promoting international understanding at an early age, and simultaneously knowing that the division into courses tends to perpetuate social handicaps, is not in accordance with the general philosophy of a democratic school.

Our recent digressions have removed us somewhat from the main theme of the present book, the comparative studies. We shall conclude by shortly returning to them. The often quoted foreign-language teaching debate in Sweden has displayed a diversity of opinion in theoretical and methodological matters. The following quotation from Campbell & Stanley (1967) may help to bring the dispute into proportion:

> "When one finds, for example, that competent observers advocate strongly divergent points of view, it seems likely on a priori grounds that both have observed something valid about the natural situation, and that both represent a part of the truth. The stronger the controversy, the more likely this is" (p. 173).

I should be noted that the Swedish debate was particularly concerned with the teaching at the compulsory and 'gymnasium' levels, whereas the teaching of adults was mainly exempted from the debate. As far as we can judge, our results at this level, favouring the cognitive code-learning theory, find acceptance among the teaching profession and support in the research literature. At the compulsory school level our results do not substantiate the orientation of the curriculum towards the audio-lingual habit theory. However, the treatment differences at this level were generally very small and the slight superiority of the Explicit-Swedish method should not be taken as conclusive evidence but await confirmation by further research. What direction this research should take can only be speculated on at the moment. In our experiments the teacher variable was held constant, as an experimental necessity, by use of taped lessons. Some critics might argue that this is equal to hampering the teaching process inordinately, and probably the same critics would suggest a more process-oriented approach involving observation of the behaviors of the teacher and his interaction with the students. Both this type of research and the one adopted in the GUME project have their advantages and limitations, and a well-designed combination of the two will probably prove rewarding.

Finally, we have not compared "methods" of teaching in any other sense than the one attributed to them in this book. Apart from this, we hope to have contributed to fostering a more balanced view on the alleged superiority of whichever foreign-language teaching method the reader may have in mind.

CHAPTER 14

SUMMARY

The research presented in this volume has been carried out within the so-called GUME project (the Swedish equivalent of Gothenburg/Teaching/Methods/English) and is an interdisciplinary undertaking, the project members representing English and education as academic disciplines. The main purpose of the research, extending over a four-year period of time, has been to investigate the tenability of two foreign-language learning theories, the audio-lingual habit theory and the cognitive code-learning theory (the two terms have been coined by Carroll).

The audio-lingual habit theory has its roots in the twenties and thirties when structural linguists began to view language as a means of communication; it is closely related to Skinnerian behaviorism. Two of the major assumptions underlying the theory are (1) foreign-language learning is basically a mechanical process of habit formation, (2) analogy provides a better foundation for foreign-language learning than analysis. According to the audio-lingual theory conscious attention to the critical features of a grammatical structure will interfere with the fluent use of it. Audio-lingual techniques aim at giving the student automatic control of the language by means of pattern practice and structure drills, and so-called mimicry-memorization of dialogue material is intended to serve the purpose of rendering the linguistic behavior habitual and automatic. It is often stressed that language patterns should be learned to the point of overlearning. Among the main proponents of the audio-lingual method are Brooks and Politzer.

Criticism of the audio-lingual theory has been levelled by several authors, of which Chomsky, Saporta, Jakobovits, and in Sweden Ellegård, all representatives of the cognitive code-learning theory, may be mentioned. According to this theory imitation and reinforcement, two concepts closely connected with the behaviorist view, are inadequate for describing the learning of the native as well as a foreign language. Chomsky, for instance, has stressed that normal linguistic behavior is stimulus-free and innovative, and referred to this property as "the creative aspect of linguistic use". The child learning his native language, as well as the foreign-language student, has to learn not only sets of responses but also some form of internal strategies of plans; having

learned these plans seems to be synonymous to having achieved competence, knowing a language, or having gained insight. And, most importantly, this insight is supposed to be facilitated by explanation of the rules of the language.

The two conflicting theories have been reflected in the foreign-language teaching debate in Sweden during the last few years. A large proportion of this debate was concerned with the merits and deficiencies of the language teaching method recommended in the authorized curriculum, which may be said to be generally oriented towards the audio-lingual habit formation theory. Although the debate contained arguments pro and con, the majority of participants in the debate obviously favoured a method fostering "insight", i.e. a method in line with the cognitive code-learning theory. However, the methods discussed were seldom strictly defined by the debaters, methods were treated vaguely and globally and due consideration was not always paid to such aspects as the age and ability of the learner, the particular aspect of language to be taught (vocabulary, syntax), etc. Unfortunately, the deficiencies inherent in the debate seem to be shared by some current research.

Considering the strongly opposing opinions in foreign-language theory and practice, it is perhaps natural that a tendency towards eclecticism should occur. Various authors have suggested a synthesis of the two theories and stressed that there must be a constant interplay in learning by analogy and analysis, of inductive and deductive processes. Notable among theorists who have suggested this kind of theoretical compromise are Rivers and Carroll.

Within the GUME project various teaching strategies, designed so as to correspond to the two theories mentioned, have been compared at different age levels. We have concentrated on syntax learning; the lessons produced thus cover only one aspect of the foreign language. Within eight of our part projects, one specific area of English syntax, known to cause Swedish pupils great trouble, was chosen for investigation; in the remaining two projects five and seven different problems of syntax were included. Although the lessons outwardly resemble ordinary lessons in that they are varied and include practice in various skills (listening, speaking, reading, writing), they differ in that the goal is more limited: only learning of grammatical structures is concerned. The teaching procedures were modified somewhat from experiment to experiment depending on experiences and findings made in the course of our research. This being so, we prefer to regard the series of experiments as fairly independent replications.

The three methods compared are:

Im – The implicit method
Ee – The explicit-English method
Es – The explicit-Swedish method

The implicit method, based on the habit formation theory, is strictly systematized but contains no explicit formulations of either what the drills are about or how the grammatical problems should be solved. The pupil's attention is directed to the crucial features of the sentence by way of analogy or contrast, and the systematized drills are supposed to result in a subconscious assimilation of the rules. The Swedish language is not used on any occasion.

Both our explicit methods fall under the cognitive code-learning theory. The pupil is made consciously aware of the functioning of the language by verbalized generalizations and explanations about what he has just heard, spoken, read, or written. It is worth pointing out that no grammar rules in the traditional sense are given, no rules for the pupils to learn, but there are just explanations of and comments on what the pupils are doing in the drills. This description of the explicit methods holds for the experiments performed at the comprehensive school level. In one case the experimental sample is an adult group (see below); here a somewhat different approach is used: rules proper are given, translation exercises are practised, and a good deal of the lessons are given in the native language (in the adult sample Es is the only explicit method offered). In the experiments at the compulsory school level the explicit methods are not to be compared with a grammar-translation method; in fact, a large part of the time is taken up by structure drills, the same as in the implicit method.

At the comprehensive school level both varieties of the explicit methods are used. Ee gives the explanations in English, whereas Es uses the Swedish language. The explanations in English and Swedish are not merely translations of each other, as the Swedish version also includes comparisons with the corresponding Swedish structures.

The experiments were performed in grade 6, 7 (three parallel experiments), and 8 of the Swedish comprehensive school, i.e. when the pupils are approximately 13, 14, and 15 years of age. One investigation was undertaken at the Gothenburg Municipal School for Adults, the students varying in age from 17 to 60 with a mean of 33 years and having no academic training beyond the compulsory school level. The majority of the adult group have occupational duties and devote a relatively limited time to studies.

The experimental schedule was very similar from project to project. The essential features of the procedure in each case were, in chronological order: (1) IQ testing (2) Pretest (3) Introductory lesson explaining experimental aims, procedures, drill techniques, etc. (4) The lesson series administered (the experiment proper) (5) Posttest (6) Pupil and teacher attitude tests (7) Retest (only in the three experiments in grade 7).

As in often the case in school research, it was not feasible to sample experimental subjects on an individual level, but intact school classes had to

be used. At the planning stage of each experiment a request for participation was sent to a large number of schools and teachers. In cases where a surplus of positive answers was obtained, the final choice of classes was based on various criteria, such as the experience of the teacher, the boys/girls ratio, the textbook used, and schedule considerations. The final number of classes thus obtained was randomly distributed among treatments, though with one restriction: in no school were two classes allowed to receive the same treatment. The classes represent a large geographic variation within the Gothenburg area. The sampling procedure described does not apply to the adult sample; in this case the total group taking the grade 7 course at the Gothenburg Municipal School for Adults was engaged in the experiment.

The comparability between the various treatment groups was investigated in a number of background variables. The general picture is one of equality between the treatment groups. Thus, the internal validity of the experiments may be judged to be satisfactory. The different samples at the compulsory school level were compared with their respective populations in several respects. The experimental samples are not, in all cases, representative of the corresponding populations, and caution must be observed in generalizing the results. The adult sample poses a specific generalizability problem since it is difficult to visualize a population of which our group may be considered a sample. We think the results pertaining to the adult sample may be regarded as valid for adult groups possessing the general characteristics mentioned previously.

The samples at the compulsory school level thus represent grades 6, 7, and 8. In the former all pupils are taught one and the same course in English whereas, from grade 7 and onwards, the pupils are divided into two courses, sk (= "särskild kurs" = advanced course) and ak (= "allmän kurs" = easier course). In grades 7 and 8 we have treated the two courses separately in all computations. In total our investigations include ten more or less similarly designed experiments; the survey on the following page illustrates various characteristics of the groups and the chronological order in which the experiments were undertaken.

In order for a subject to be included in the statistical analyses, he was not allowed to be absent more than one lesson (in the case of the 6-lesson series) or two lessons (in the case of the 10- and 12-lesson series).

A criterion test intended to measure progress made during the lesson series was constructed in each part project. Each test was to measure what had been specifically taught in the respective experiment; of necessity the test should have high content validity. We have discussed, at some length, the probability that our tests might be biassed towards one teaching method or another. Although it is difficult to gauge this bias, if any, we have ventured to argue that, in the light of the general objectives of our experiments, the criterion

Part Project	Grade	Appr. age level	N of classes (groups)	Total N of subjects	N of classes (groups) in each treatment Im	Ie	Es	N of lessons per treatment
GUME 1 sk	7	14	12	227	4	4	4	6
GUME 1 ak	7	14	6	104	2	2	2	6
GUME 2 sk	7	14	12	247	4	4	4	6
GUME 2 ak	7	14	6	98	2	2	2	6
GUME 3 sk	7	14	12	170	4	4	4	6
GUME 3 ak	7	14	6	57	2	2	2	6
GUME 4	6	13	27	577	9	9	9	12
GUME 5 sk	8	15	12	235	4	4	4	6
GUME 5 ak	8	15	12	152	4	4	4	6
GUME A	7	adults	6	125	3	–	3	10

tests do not favour any particular method.

In all experiments roughly similar questionnaires, intended to disclose the student's attitudes towards the teaching procedures, were administered. Similarly, the teachers were asked to give their opinions on various aspects of the experiments in a teacher attitude test.

Our research design and statistical treatment implied various considerations. For instance, measuring progress by means of a difference score (Posttest – Pretest = Progress) has been criticized because of the unsatisfactory psychometric properties of raw gain scores. In general, analysis of variance of difference scores has lower precision than analysis of covariance and treatment x levels designs. Feldt, in comparing the three techniques, states that analysis of covariance is to be preferred when the correlation between the dependent variable and the covariate is. 60 or more, that the treatment x levels design is to be preferred when the correlation is between .20 and .60, partly because of its less stringent assumptions (no linear regression between x and y), and that analysis of variance of difference scores is generally inferior unless the correlation between the control and criterion variable is substantial. Having suitable data, and assuming that our experiments may provide an empirical check on this problem, we have performed the three types of computation in each of our experiments. In addition we have performed analyses of variance of a second type of difference score, the so-called Actual/Possible progress score. This type of score relates raw gain to the ceiling effect of the test in so far as it gives proportional credit to pupils with high pretest scores; the assumption is that an increase from, say 40 to 60 points, is relatively more difficult to achieve than an increase from 20 to 40.

As was mentioned previously the single school class is the sampling unit. If

the individual were used as the unit of statistical treatment, the error term would be spuriously low, assuming that each school class is more homogeneous than the population from which it was sampled. We have not calculated any intra-class correlations, i.e. we have no precise measure of the class homogeneity. However, in cases where the number of observations permits, we have made calculations utilizing the school class means as the statistical unit.

It may be argued that our teaching strategies, in so far as they appear to have no different effects on the treatment means may be differentially related to various levels of student ability, age, etc. Basically, this is a question of treatment-aptitude interaction. We have discussed this concept at some length and also tentatively investigated, by analysis of variance (two-way classification), whether our treatments tended to interact with different levels of scholastic aptitude and achievement, the latter defined by pretest scores.

What, then, is the main outcome of our studies? At the compulsory school level the pupils belonging to the easier course (ak) generally made very little progress, which of course minimizes the probability of obtaining treatment effects. Nor were any such discernible in the easier course. We have ventured the hypothesis, strengthened by various findings, that belonging to the easier course somehow causes low motivation; the pupils do not work up to their ability, at least as judged by their DBA ("intelligence") scores.

In the sk courses, where the progress is considerably greater, no differential treatment effects appear in any single experiment. However, when the four sk groups are considered as a whole, the results present a consistent pattern from experiment to experiment. The Explicit-Swedish method ranks first, the Implicit method second, and the Explicit-English method last. When three methods are compared, there are six possible premutations of rank orderings. If the three later experiments are regarded as replications of the first, the probability of obtaining exactly the Es > Im > Ee rank order in each of the three experiments is extremely low. Our finding is substantiated by the fact that exactly this ordering was found in the analyses of covariance, no matter whether they were performed with the individual or the school class means as the unit of analysis, and in the analyses of variance of raw gain scores. It thus appears that, in the sk courses, a teaching method utilizing the native language for explanations tends to facilitate learning. The Implicit method, consisting of structure drills but no explanations, is in turn better than a method where explanations are given English.

At the adult level the tendency is in a similar direction. The Es method proved significantly superior to the Implicit method in all comparisons made. Even in the case of an oral test, which might have been suspected to favour the Implicit method, the Es method excelled. Thus, at the adult level the results suggest that explanations in the mother tongue clarifying linguistic

patterns are effective in internalizing the English grammar even when supplied at the expense of practice. They also suggest that pattern drills are of limited value as long as the adult has not been provided insight into the structure of language.

The results at the adult level and in the sk course at the compulsory school level thus support the cognitive code-learning theory. We previously commented on the orientation of the present curriculum towards a mechanistic school of thinking. It should be noted that in the latest version of the curriculum, Lgr 69, this orientation is even more pronounced than in its predecessor, Lgr 62. Our research results apparently do not lend support to this development. Nor do they lend support to the somewhat categorical formulations in the curriculum on the necessity of using the English structure as starting-point for explanations or discussions about problems of syntax.

In view of the fact that the various teaching strategies did not produce any differences in the ak course and no dramatic differences in the sk course, we have investigated whether they had any effect on the group variances. That is to say, we investigated if the increase – or decrease – in scores from pretest to posttest varied between treatments. The general picture is one of no such differences at the compulsory school level. The only exception is found in GUME 4, that is the youngest sample (grade 6), where the Es method brought about a larger variation in scores from pre- to posttest than did the Implicit method. This result indicates that, in a comparatively young group of students, the method utilizing explanations in the mother tongue tends to favour the more able students and put a handicap on the less able. This finding thus suggests the existence of interaction between ability level and treatment. Otherwise our search for interactions between treatments and various levels of pupil aptitude or ability did not yield any notable results. Although a few statistically significant interaction terms were obtained, the findings appeared to be inconsistent and rather difficult to interpret.

The students' responses to the attitude test bore no clear relation to the outcomes in terms of learning effects. In the majority of cases the pupils were neutral or slightly positive to their respective teaching procedure. The teachers, on the other hand, proved to favour a method similar to our Es method. Incidentally, this predilection of the teachers corresponds to teacher opinions as evidenced in the Swedish foreign-language teaching debate.

A number of control classes were given the criterion test on three occasions dispersed over a two-year period of time. The results indicate that the pupils learn as much (of one specific structure) in one year of ordinary teaching as they did in our six project lessons. The results in the control classes further indicate that the difference between the sk and ak means tends to increase from one year to the next. This "gliding-apart effect", which is also marked in our grade 8 experimental sample proper, is regarded as indicating

the unwanted identification phenomenon (in the ak course) alluded to above. In view of this, we ventured to speculate on what effects lumping the two courses together would have on teaching. As far as we could find the negative effects, if any, would be outweighed by the positive effects of not leaving the choice of course to the pupil and/or his parents. This choice, although free in principle, partly reflects social handicaps.

In sum, our main results tend to support the cognitive code-learning theory at the upper stages of the compulsory school level and are decidedly in favor of it at the adult level. The findings do not suggest a return to a traditional grammar-translation method involving a lot of rule-cramming and practically no oral practice; our explicit methods simply do not resemble this type of obsolete procedure, not even the fairly traditional method used in the adult sample. Besides, we have not compared methods of teaching in a general sense, but rather specific variables related to two foreign-language learning theories.

APPENDICES

APPENDIX I

List of reports from the GUME project

Lindblad, T. Implicit and explicit. An experiment in applied psycholinguistics assessing different methods of teaching grammatical structures in English as a foreign language. June 1969.

Carlsson, I. Implicit and explicit. An experiment ..... September 1969.

Olsson, M. Implicit and explicit. An experiment .....September 1969.

Levin L. Implicit and explicit. En jämförande studie av olika metoder att lära ut grammatiska strukturer i engelska. Sammanfattande rapport. September 1969.

Levin, L. Implicit and explicit. A synopsis of three parallel experiments in applied psycholinguistics assessing three different methods of teaching grammatical structures in English as a foreign language. December 1969.

Lindblad, T & Levin, L. Teaching grammar. December 1970.

Levin, L & Olsson, M. Learning grammar. January 1971.

von Elek, T & Oskarsson, M. Teaching foreign-language grammar to adults: A comparative study. May 1972.

Levin, L. Comparative studies in foreign-language teaching. June 1972.

APPENDIX 2

Table I. *GUME 1, sk.* Means and Standard Deviations for the Total and for the Treatment Groups.

VARIABLE	N	$\bar{x}$	s	$\bar{x}$			s,(N)		
				Im	Ee	Es	Im	Ee	Es
Pre-test	227	71.01	16.23	70.87	70.39	71.90	15.58(69)	17.66(77)	15.50(81)
Post-test	227	81.76	17.02	81.72	79.95	83.52	16.16(69)	18.17(77)	16.63(81)
Progress(raw)	227	10.69	8.70	10.86	9.56	11.62	8.97(69)	8.39(77)	8.75(81)
Act/Poss Progress	227	23.57	21.52	23.42	19.73	27.36	20.89(69)	21.81(77)	21.38(81)
Re-test	206	84.16	17.36	84.83	80.88	86.66	15.39(64)	18.95(69)	17.15(73)
Ret. Progress	206	13.42	9.75	14.67	11.54	14.10	9.72(64)	9.79(69)	9.62(73)
Pup. Attitude	204	25.47	5.03	26.22	24.28	26.00	5.47(59)	4.42(71)	5.08(74)
Verbal DBA	221	5.74	1.63	5.56	5.53	6.09	1.33(68)	1.57(75)	1.87(78)
Inductive DBA	221	5.71	1.72	5.76	5.53	5.83	1.82(68)	1.84(75)	1.90(78)
Spatial DBA	220	5.42	1.86	5.26	5.28	5.69	1.60(68)	2.04(74)	1.87(78)
Total DBA	215	16.89	3.77	16.58	16.39	17.64	3.03(67)	4.04(72)	4.02(76)
Grades Engl.	227	3.68	0.84	3.70	3.58	3.77	0.83(69)	0.88(77)	0.81(81)
Grades Sw.	225	3.65	0.84	3.57	3.61	3.77	0.79(69)	0.90(75)	0.81(81)
Grades Maths.	225	3.57	0.88	3.71	3.35	3.67	0.82(69)	0.92(75)	0.87(81)
Grades Total	225	10.92	2.18	10.97	10.56	11.20	2.01(69)	2.36(75)	2.14(81)
Social Class	208	1.87	0.79	1.88	1.88	1.84	0.82(65)	0.84(68)	0.72(75)

APPENDIX 2

Table II *GUME 1, ak.* Means and Standard Deviations for the Total and for the Treatment Groups.

VARIABLE	N	$\bar{x}$	s	$\bar{x}$ Im	$\bar{x}$ Ee	$\bar{x}$ Es	s.(N) Im	s.(N) Ee	s.(N) Es
Pre-test	104	48.35	12.29	48.17	46.71	50.21	11.56(23)	12.98(42)	12.00(39)
Post-test	104	51.65	13.00	55.61	49.17	52.00	13.64(23)	12.35(42)	13.02(39)
Progress (raw)	104	3.31	8.08	7.43	2.45	1.79	7.51(23)	7.35(42)	8.52(39)
Act/Poss Progress	104	4.15	12.07	10.65	2.74	1.85	10.89(23)	10.74(42)	12.98(39)
Re-test	91	54.85	12.59	55.95	53.17	55.94	13.64(21)	11.12(36)	13.54(34)
Ret. Progress	91	5.87	8.17	7.33	5.64	5.21	8.34(21)	8.30(36)	8.05(34)
Pup. Attitude	83	26.02	4.72	29.47	24.50	25.90	4.68(17)	3.65(36)	5.01(30)
Verbal DBA	99	3.52	1.37	3.38	3.83	3.24	1.16(21)	1.46(41)	1.34(37)
Inductive DBA	99	3.70	1.78	3.62	3.63	3.81	1.12(21)	2.12(41)	1.70(37)
Spatial DBA	101	4.12	1.82	4.00	4.07	4.24	1.80(22)	1.93(41)	1.75(38)
Total DBA	96	11.23	3.37	10.85	11.50	11.14	2.32(20)	3.73(40)	3.49(36)
Grades Engl.	103	2.12	0.74	2.00	2.10	2.21	0.60(23)	0.77(41)	0.80(39)
Grades Sw.	103	2.20	0.68	2.09	2.20	2.28	0.60(23)	0.75(41)	0.65(39)
Grades Maths.	103	2.13	0.74	2.00	2.10	2.23	0.52(23)	0.77(41)	0.81(39)
Grades Total	103	6.45	1.79	6.09	6.39	6.72	1.28(23)	1.80(41)	2.01(39)
Social Class	98	2.62	0.55	2.59	2.79	2.47	0.59(22)	0.41(38)	0.60(38)

APPENDIX 2

Table III *GUME 2, sk.* Means and Standard Deviations for the Total and for the Treatment Groups.

VARIABLE	N	$\bar{x}$	s	$\bar{x}$ Im	$\bar{x}$ Ee	$\bar{x}$ Es	s,(N) Im	s,(N) Ee	s,(N) Es
Pre-test	247	64.77	17.16	64.32	66.58	62.96	17.17(84)	15.61(92)	19.01(71)
Post-test	247	81.31	17.76	81.20	81.80	80.79	18.63(84)	16.67(92)	18.30(71)
Progress (raw)	247	16.54	10.68	16.88	15.23	17.83	10.03(84)	11.41(92)	10.41(71)
Act/Poss Progress	247	26.12	17.39	26.69	24.09	28.08	16.15(84)	18.03(92)	17.92(71)
Re-test	229	83.01	17.64	82.44	84.81	81.45	17.84(80)	17.71(83)	17.37(66)
Ret. Progress	229	18.97	11.71	18.50	18.98	19.53	10.25(80)	12.89(83)	11.97(66)
Pup. Attitude	223	28.92	4.53	28.07	30.03	28.43	4.65(76)	4.24(86)	4.55(61)
Verbal DBA	233	5.47	1.52	5.51	5.58	5.27	1.41(79)	1.46(90)	1.71(64)
Inductive DBA	233	5.56	1.94	5.43	5.58	5.69	1.79(79)	2.15(90)	1.82(64)
Spatial DBA	234	5.23	1.91	5.28	5.40	4.94	2.01(79)	1.78(90)	1.98(65)
Total DBA	230	16.31	4.06	16.29	16.56	15.98	3.88(77)	4.06(90)	4.30(63)
Grades Engl.	239	5.53	0.87	3.47	3.61	3.49	0.82(78)	0.95(92)	0.83(69)
Grades Sw.	239	3.50	0.81	3.51	3.60	3.36	0.79(78)	0.84(92)	0.80(69)
Grades Maths.	239	3.48	0.94	3.55	3.38	3.52	0.88(78)	0.91(92)	1.05(69)
Grades Total	239	10.51	2.26	10.54	10.59	10.38	2.13(78)	2.34(92)	2.33(69)
PACT	235	53.54	7.81	52.44	54.83	53.11	8.07(80)	7.93(90)	7.13(65)
Social Class	202	2.33	0.72	2.51	2.37	2.11	0.64(65)	0.61(73)	0.84(64)

APPENDIX 2

Table IV *GUME 2. ak.* Means and Standard Deviations for the Total and for the Treatment Groups.

VARIABLE	N	$\bar{x}$	s	$\bar{x}$			s.(N)		
				lm	Ee	Es	lm	Ee	Es
Pre-test	98	46.77	14.09	48.05	47.18	43.82	14.78(38)	13.36(38)	14.32(22)
Post-test	98	59.61	15.95	59.61	60.71	57.73	16.68(38)	15.43(38)	16.11(22)
Progress (raw)	98	12.85	13.57	11.55	13.53	13.91	12.88(38)	11.91(38)	17.42(22)
Act/Poss Progress	98	14.95	16.06	13.87	16.13	14.77	16.35(38)	14.67(38)	18.35.22)
Re-test	88	58.30	15.32	59.48	58.33	56.16	15.27(33)	14.54(36)	17.39(19)
Ret. Progress	88	11.59	12.44	12.00	11.75	10.58	12.29(33)	12.71(36)	12.80(19)
Pup. Attitude	81	29.05	5.27	28.52	29.06	29.71	5.34(27)	5.27(33)	5.38(21)
Verbal DBA	90	3.70	1.53	3.38	3.94	3.81	1.72(34)	1.55(35)	1.08(21)
Inductive DBA	90	3.72	1.61	3.88	3.51	3.81	1.55(34)	1.85(35)	1.25(21)
Spatial DBA	90	4.07	1.65	4.03	4.26	3.81	1.71(34)	1.70(35)	1.47(21)
Total DBA	90	11.49	3.55	11.29	11.71	11.43	3.54(34)	4.20(35)	2.31(21)
Grades English	94	2.23	0.72	2.29	2.16	2.27	0.71(35)	0.76(37)	0.70(22)
Grades Sw.	94	2.37	0.66	2.26	2.49	2.36	0.61(35)	0.73(37)	0.58(22)
Grades Maths.	94	2.31	0.76	2.23	2.32	2.41	0.88(35)	0.71(37)	0.67(22)
Grades Total	94	6.91	1.73	6.77	6.97	7.05	1.78(35)	1.80(37)	1.59(22)
PACT	88	42.83	8.65	44.76	41.15	42.33	10.25(34)	6.88(33)	8.09(21)
Social Class	80	2.71	0.48	2.69	2.81	2.60	0.54(29)	0.40(31)	0.50(20)

APPENDIX 2

Table V *GUME 3, sk.* Means and Standard Deviations for the Total and for the Treatment Groups.

VARIABLE	N	$\bar{x}$	s	$\bar{x}$			s,(N)		
				Im	Ee	Es	Im	Ee	Es
Pre-test	170	86.09	15.39	86.92	82.32	89.53	14.98(50)	15.72(63)	14.71(57)
Post-test	170	97.27	16.19	98.92	91.89	101.77	12.27(50)	17.72(63)	16.00(57)
Progress (raw)	170	11.18	10.05	12.00	9.57	12.25	11.16(50)	8.62(63)	10.45(57)
Act/Poss Progress	170	25.55	24.80	24.18	21.49	31.25	24.39(50)	20.50(63)	28.58(57)
Re-test	148	99.28	16.34	100.20	96.19	101.72	14.47(45)	18.21(53)	15.62(50)
Ret. Progress	148	13.23	11.28	13.47	13.91	12.30	13.13(45)	9.96(53)	10.96(50)
Pupil Attitude	154	25.06	5.56	24.38	25.23	25.49	4.91(47)	5.74(56)	5.96(51)
Verbal DBA	157	5.58	1.76	5.53	5.34	5.89	1.65(45)	1.79(59)	1.82(53)
Inductive DBA	161	5.76	1.73	6.13	5.79	5.42	1.63(45)	1.80(61)	1.70(55)
Spatial DBA	161	5.34	1.90	5.16	5.26	5.56	2.08(45)	1.77(61)	1.90(55)
Total DBA	155	16.59	4.07	16.82	16.38	16.63	4.07(45)	4.15(58)	4.03(52)
Grades Engl.	168	3.63	0.81	3.64	3.57	3.68	0.75(50)	0.90(61)	0.76(57)
Grades Sw.	168	3.58	0.88	3.72	3.46	3.58	0.81(50)	0.91(61)	0.91(57)
Grades Maths.	168	3.53	0.94	3.54	3.41	3.65	0.93(50)	0.99(61)	0.90(57)
Grades Total	168	10.74	2.25	10.90	10.44	10.91	2.00(50)	2.46(61)	2.21(57)
PACT	157	54.71	7.29	53.24	53.75	57.02	6.83(45)	7.87(59)	6.53(53)
Social Class	161	2.23	0.72	2.07	2.41	2.18	0.74(46)	0.65(58)	0.73(57)

APPENDIX 2

Table VI *GUME 3, ak.* Means and Standard Deviations for the Total and for the Treatment Groups.

				$\bar{x}$			s,(N)		
VARIABLE	N	$\bar{x}$	s	Im	Ee	Es	Im	Ee	Es
Pre-test	57	65.44	10.46	63.38	64.40	68.00	12.70(16)	7.46(20)	11.00(21)
Post-test	57	69.58	11.16	66.56	69.30	72.14	10.30(16)	9.15(20)	13.25(21)
Progress (raw)	57	4.14	8.44	3.19	4.90	4.14	8.95(16)	7.25(20)	9.38(21)
Act/Poss Progress	57	5.67	13.44	3.13	7.10	6.24	15.52(16)	10.14(20)	14.84(21)
Re-test	50	72.92	14.43	69.47	71.06	77.56	15.26(15)	10.74(17)	16.22(18)
Ret. Progress.	50	8.04	9.30	7.27	6.53	10.11	9.75(15)	7.21(17)	10.72(18)
Pupil Attitude	55	25.76	6.02	23.07	25.75	27.80	6.71(15)	6.32(20)	4.43(20)
Verbal DBA	55	3.56	1.60	4.00	2.80	4.00	1.52(14)	1.24(20)	1.73(21)
Inductive DBA	54	3.00	1.30	3.15	2.85	3.05	1.41(13)	1.18(20)	1.40(21)
Spatial DBA	54	3.93	1.95	3.85	4.00	3.90	1.82(13)	2.34(20)	1.70(21)
Total DBA	54	10.48	2.93	11.00	9.65	10.95	2.24(13)	3.18(20)	3.01(21)
Grades English	56	2.14	0.75	2.19	1.80	2.45	0.66(16)	0.52(20)	0.89(20)
Grades Swedish	56	2.30	0.74	2.56	1.90	2.50	0.73(16)	0.64(20)	0.69(20)
Grades Maths.	56	2.38	0.80	2.69	2.15	2.35	0.87(16)	0.75(20)	0.75(20)
Grades Total	56	6.82	1.82	7.44	5.85	7.30	1.67(16)	1.50(20)	1.89(20)
PACT	55	45.36	9.17	46.43	43.20	46.71	11.20(14)	7.84(20)	8.94(21)
Social Class	49	2.57	0.54	2.75	2.47	2.50	0.45(16)	0.51(19)	0.65(14)

APPENDIX 2

Table VII *GUME 4.* Means and Standard Deviations for the Total and for the Treatment Groups.

VARIABLE	N	$\bar{x}$	s	$\bar{x}$ Im	$\bar{x}$ Ee	$\bar{x}$ Es	s,(N) Im	s,(N) Ee	s,(N) Es
Pre-test	574	51.61	20.89	48.83	53.14	52.27	21.46(180)	21.25(194)	19.91(200)
Post-test	574	68.67	27.16	65.35	70.79	69.58	25.70(180)	28.14(194)	27.30(200)
Progress, raw	574	17.26	12.32	16.52	17.64	17.54	11.95(180)	12.35(194)	12.65(200)
Act/Poss Progr.	574	17.63	14.42	16.29	18.36	18.12	13.48(180)	14.79(194)	14.85(200)
Pupil Attitude	529	22.94	4.41	23.01	23.53	22.34	3.98(167)	4.24(173)	4.85(189)
Verbal DBA	561	5.30	1.79	5.09	5.43	5.36	1.83(172)	1.85(193)	1.67(196)
Inductive DBA	561	5.79	1.93	5.65	5.95	5.76	1.85(172)	1.91(193)	2.02(196)
Spatial DBA	561	5.56	1.97	5.70	5.54	5.46	1.91(172)	1.89(193)	2.09(196)
Total DBA	561	16.66	4.33	16.44	16.92	16.59	4.12(172)	4.43(193)	4.43(196)
Grades Engl.	576	3.09	1.03	3.13	3.09	3.04	1.07(181)	0.98(194)	1.04(201)
Grades Sw.	573	3.15	0.92	3.12	3.19	3.13	0.95(178)	0.90(194)	0.91(201)
Grades Maths.	576	3.08	0.97	3.01	3.14	3.09	0.94(181)	0.98(194)	0.98(201)
Grades Total	570	9.32	2.58	9.25	9.42	9.29	2.63(177)	2.52(193)	2.60(200)
National Test	569	53.48	18.68	50.94	56.04	53.20	17.87(175)	19.06(195)	18.77(199)
PACT	550	34.29	8.77	33.55	34.61	34.63	8.83(172)	8.83(190)	8.65(188)
Social Class	548	2.20	0.93	2.21	2.26	2.13	0.88(168)	0.89(189)	1.01(191)

APPENDIX 2

Table VIII *GUME 5, sk.* Means and Standard Deviations for the Total and for the Treatment Groups.

VARIABLE	N	$\bar{x}$	s	$\bar{x}$ Im	$\bar{x}$ Ee	$\bar{x}$ Es	s.(N) Im	s.(N) Ee	s.(N) Es
Pre-test	235	59.11	14.83	60.69	60.22	56.19	14.80(70)	14.13(92)	15.48(73)
Post-test	235	66.99	14.13	68.44	67.86	64.49	13.37(70)	14.09(92)	14.73(73)
Progress (raw)	235	7.88	8.04	7.76	7.64	8.30	7.14(70)	8.80(92)	7.96(73)
Act/Poss Progr.	235	19.92	20.60	20.24	19.71	19.88	18.32(70)	23.39(92)	19.14(73)
Pupil Attitude	200	22.55	4.43	24.03	21.97	21.73	4.19(64)	4.24(73)	4.59(63)
Verbal DBA	214	5.87	1.63	6.20	5.78	5.66	1.54(65)	1.61(85)	1.72(64)
Inductive DBA	214	5.89	1.73	6.22	5.79	5.70	1.66(65)	1.71(85)	1.79(64)
Spatial DBA	214	5.59	2.08	5.86	5.26	5.75	1.85(65)	2.16(85)	2.17(64)
Total DBA	214	17.37	3.94	18.28	16.82	17.11	3.76(65)	3.71(85)	4.31(64)
Grades Engl.	233	3.33	1.00	3.35	3.40	3.22	0.94(68)	0.98(92)	1.07(73)
Grades Sw.	233	3.47	0.84	3.50	3.49	3.42	0.76(68)	0.86(92)	0.90(73)
Grades Maths.	233	3.21	1.06	3.21	3.28	3.14	1.11(68)	1.10(92)	0.95(73)
Grades Total	233	10.02	2.45	10.06	10.17	9.78	2.39(68)	2.49(92)	2.47(73)
National Test	228	49.41	12.81	50.09	51.18	46.51	11.83(70)	12.30(88)	14.00(70)
PACT	227	50.19	3.10	49.99	50.90	49.51	3.06(67)	2.52(89)	3.61(71)
Social Class	185	2.10	0.86	1.82	2.17	2.29	0.97(57)	0.76(70)	0.79(58)

APPENDIX 2

Table IX *GUME 5, ak.* Means and Standard Deviations for the Total and for the Treatment Groups.

				$\bar{x}$			s.(N)		
VARIABLE	N	$\bar{x}$	s	Im	Ee	Es	Im	Ee	Es
Pre-test	152	31.52	7.08	33.12	30.80	30.68	6.72(50)	7.01(49)	7.35(53)
Post-test	152	35.24	8.45	35.70	33.18	36.72	7.61(50)	9.10(49)	8.35(53)
Progress (raw)	152	3.72	6.80	2.58	2.39	6.04	6.77(50)	5.94(49)	7.07(53)
Act/Poss Progr.	152	5.39	10.42	3.66	3.59	8.68	11.00(50)	9.30(49)	10.26(53)
Pupil Attitude	140	21.34	4.38	21.60	19.86	22.34	4.20(48)	4.12(42)	4.51(50)
Verbal DBA	120	3.67	1.19	3.63	3.45	3.95	1.05(43)	1.32(40)	1.15(37)
Inductive DBA	120	3.81	1.73	3.72	3.57	4.16	1.79(43)	1.60(40)	1.77(37)
Spatial DBA	120	4.54	1.89	4.44	4.25	4.97	1.69(43)	2.23(40)	1.66(37)
Total DBA	120	12.02	3.56	11.79	11.27	13.08	3.41(43)	3.63(40)	3.49(37)
Grades Engl.	148	2.86	0.86	2.92	2.71	2.96	0.70(49)	1.00(49)	0.86(50)
Grades Sw.	148	2.30	0.60	2.37	2.22	2.30	0.57(49)	0.59(49)	0.65(50)
Grades Maths.	148	2.66	0.89	2.55	2.51	2.92	0.91(49)	0.96(49)	0.75(50)
Grades Total	148	7.82	1.84	7.84	7.45	8.18	1.68(49)	1.99(49)	1.79(50)
National Test	148	39.18	10.49	38.39	36.42	42.55	10.36(49)	11.99(48)	8.10(51)
PACT	139	43.19	6.72	43.65	40.54	45.34	6.46(46)	7.71(46)	4.94(47)
Social Class	150	2.51	0.83	2.60	2.61	2.33	0.73(50)	0.70(49)	0.99(51)

APPENDIX 2

Table X *GUME A.* Means and Standard Deviations for the Total and for the Treatment Groups.

				$\bar{x}$		s,(N)	
VARIABLE	N	$\bar{x}$	s	Im	Es	Im	Es
Pre-test	125	54.72	15.93	56.56	53.18	18.32(57)	13.57(68)
Post-test	125	74.10	20.58	69.93	77.60	20.03(57)	20.53(68)
Progress (raw)	125	19.38	13.33	13.37	24.43	10.12(57)	13.67(68)
Act/Poss Progress	125	26.96	19.20	19.39	33.32	15.62(57)	19.71(68)
Oral Test	95	34.72	9.98	32.67	36.14	10.97(39)	9.06(56)
Pupil Attitude	119	24.80	3.99	22.60	26.69	4.21(55)	2.59(64)
Other Subjects	122	1.86	1.00	2.04	1.72	1.00(55)	0.98(67)
Work	122	1.64	0.48	1.69	1.60	0.47(55)	0.49(67)
Age	125	32.98	9.11	30.68	34.90	8.08(57)	9.53(68)
F-test Verbal	111	51.40	9.22	51.27	51.49	10.19(48)	8.49(63)
PACT	124	31.06	10.85	32.84	29.54	11.35(57)	10.24(67)
Diagn. Engl.	125	30.75	9.46	31.00	30.54	10.13(57)	8.94(68)

APPENDIX 3

Pretest means, standard deviations, and reliabilities.

	N of items	$\bar{x}$	s_x	r_{xx}	N of subjects
GUME 1 sk	120	71.07	16.23	0.90	227
GUME 1 ak	120	48.35	12.29	0.82	104
GUME 2 sk	131	64.77	17.16	0.90	247
GUME 2 ak	131	46.77	14.09	0.86	98
GUME 3 sk	133	86.09	15.39	0.88	170
GUME 3 ak	133	65.44	10.46	0.72	57
GUME 4	160	51.61	20.89	0.93	576
GUME 5 sk	94	59.11	14.83	0.91	235
GUME 5 ak	94	31.52	7.08	0.59	152
GUME A	130	54.72	15.93	0.88	125

APPENDIX 4

Table I *GUME I, sk.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Pre-test	.864	−.175	.244	.833	−.175	−.108	.462	.301	.162	.415	.679	.502	.328	.588	−.330
2. Post-test		.345	.640	.923	.198	−.121	.485	.323	.162	.435	.688	.581	.459	.675	−.368
3. Progress			.797	.269	.733	−.036	.093	.074	.014	.079	.079	.200	.285	.222	−.098
4. Act/Poss Progr.				.592	.560	−.069	.290	.178	.111	.260	.331	.355	.347	.404	−.168
5. Re-test					.400	−.093	.484	.357	.184	.464	.725	.632	.506	.729	−.349
6. Ret. Progress						.028	.061	.099	.028	.087	.162	.269	.312	.292	−.094
7. Pupil Attitude							−.100	−.036	.070	−.023	−.115	−.177	−.050	−.131	.124
8. Verbal DBA								.273	.226	.666	.485	.591	.353	.556	−.118
9. Inductive DBA									.356	.748	.362	.377	.470	.473	−.221
10. Spatial DBA										.753	.208	.260	.312	.303	−.017
11. Total DBA		N≈227									.473	.561	.529	.612	−.183
12. Grades Engl.												.747	.494	.871	−.243
13. Grades Sw.													.535	.888	−.201
14. Grades Maths														.800	−.202
15. Grades Total															−.253
16. Social Class															

APPENDIX 4

Table II *GUME 1, ak.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Pre-test	.797	−.239	−.213	.785	−.299	−.201	.280	.301	.114	.349	.572	.267	.242	.439	−.113
2. Post-test		.396	.410	.815	.063	−.039	.276	.397	.004	.343	.562	.242	.220	.417	−.063
3. Progress			.984	.111	.548	.251	.022	.187	−.162	.026	.037	−.015	−.013	.004	.070
4. Act/Poss Progr.				.117	.527	.202	.059	.190	−.156	.047	.020	−.022	−.019	−.008	.080
5. Re-test					.356	.059	.221	.256	−.059	.201	.579	.242	.172	.403	−.034
6. Ret. Progress						.300	−.090	.019	−.138	−.116	.044	−.027	.054	.029	.128
7. Pupil Attitude							−.256	.017	−.106	−.179	−.147	−.076	−.165	−.163	.064
8. Verbal DBA								.153	.085	.529	.285	.422	.282	.393	−.119
9. Inductive DBA									.269	.739	.113	.110	.259	.195	−.066
10. Spatial DBA										.720	.062	.173	.334	.230	−.179
11. Total DBA		N≈ 104									.220	.326	.434	.394	−.165
12. Grades Eng.												.555	.509	.837	−.112
13. Grades Sw.													.518	.824	−.013
14. Grades Maths.														.821	−.123
15. Grades Total															−.102
16. Social Class															

APPENDIX 4

Table III *GUME 2, sk.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Pretest	.814	−.254	.252	.769	−.269	−.030	.452	.325	.182	.419	.518	.469	.254	.475	.558	−.190
2. Posttest		.356	.737	.823	.086	−.027	.528	.376	.243	.497	.582	.513	.319	.543	.614	−.217
3. Progress			.822	.146	.566	.005	.152	.103	.111	.153	.130	.095	.119	.134	.126	−.054
4. Act/Poss Progr.				.506	.439	.005	.368	.273	.173	.347	.371	.286	.220	.338	.393	−.153
5. Retest					.409	.032	.493	.310	.198	.428	.527	.483	.236	.473	.615	−.180
6. Ret. Progr.						.075	.109	−.001	.081	.069	.051	.047	−.043	.018	.166	.040
7. Pupil Attitude							−.026	−.042	−.114	−.095	−.004	−.127	−.125	−.100	.043	.147
8. Verbal DBA								.314	.295	.666	.544	.610	.415	.611	.427	−.174
9. Inductive DBA									.421	.799	.337	.352	.413	.433	.311	−.097
10. Spatial DBA										.782	.287	.278	.382	.373	.253	.023
11. Total DBA		N≈247									.500	.527	.530	.612	.428	−.105
12. Grades Eng.												.687	.553	.865	.458	−.145
13. Grades Sw.													.590	.872	.422	−.195
14. Grades Maths.														.843	.232	−.189
15. Grades Total															.433	−.205
16. PACT																−.161
17. Social Class																

APPENDIX 4

Table IV *GUME 2, ak.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Pretest	.598	−.335	−.103	.641	−.324	.083	.308	.236	.098	.285	.455	.329	.284	.440	.394	−.042
2. Posttest		.555	.721	.799	.323	.071	.426	.335	.232	.443	.604	.500	.408	.621	.571	−.046
3. Progress			.955	.235	.716	−.005	.187	.154	.175	.231	.222	.234	.175	.259	.270	−.009
4. Act/Poss Progr.				.384	.678	.025	.269	.221	.244	.330	.342	.328	.248	.376	.400	−.021
5. Retest					.519	.155	.282	.280	.074	.288	.540	.510	.357	.588	.423	.012
6. Ret. Progress						.121	.004	.124	.068	.091	.124	.278	.175	.238	.112	.038
7. Pupil Attitude							−.206	−.012	−.125	−.153	.052	.022	−.021	.020	.068	−.104
8. Verbal DBA								.299	.266	.690	.360	.444	.358	.478	.359	−.109
9. Inductive DBA									.406	.770	.164	.244	.366	.323	.352	−.260
10. Spatial DBA										.763	.086	.212	.419	.301	.236	−.017
11. Total DBA		N≈98									.269	.400	.512	.491	.423	−.171
12. Grades Engl.												.561	.393	.804	.380	.081
13. Grades Sw.													.499	.833	.251	.135
14. Grades Maths														.793	.127	.122
15. Grades Total															.310	.140
16. PACT																−.208
17. Social Class																

APPENDIX 4

Table V *GUME 3, sk.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Pre-test	.799	−.245	.216	.748	−.276	−.106	.542	.331	.220	.483	.628	.616	.457	.658	.510	−.098
2. Post-test		.388	.716	.848	.161	−.044	.592	.385	.235	.543	.645	.665	.483	.695	.459	−.216
3. Progress			.822	.211	.661	.090	.121	.115	.044	.133	.074	.124	.076	.107	−.043	−.197
4. Act/Poss Progr.				.514	.521	.082	.383	.263	.149	.358	.356	.370	.297	.398	.159	−.213
5. Re-test					.431	−.076	.598	.363	.245	.534	.638	.663	.405	.659	.493	−.259
6. Ret. Progress						.092	.136	.057	.076	.131	.109	.136	.021	.101	.003	−.263
7. Pupil Attitude							−.198	−.015	−.142	−.148	−.049	.075	−.075	−.019	.023	.071
8. Verbal DBA								.373	.340	.752	.502	.583	.531	.629	.443	−.244
9. Inductive DBA									.335	.748	.329	.399	.417	.448	.109	−.123
10. Spatial DBA										.757	.233	.247	.545	.406	.125	−.206
11. Total DBA		N≈170									.484	.551	.679	.671	.295	−.276
12. Grades Engl.												.715	.487	.844	.432	−.217
13. Grades Sw													.591	.896	.378	−.228
14. Grades Maths.														.825	.238	−.291
15. Grades Total															.402	−.289
16. PACT																−.157
17. Social Class																

APPENDIX 4

Table VI *GUME 3, ak.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Pre-test	.697	−.317	−.257	.765	.041	.090	.239	.430	.265	.506	.659	.409	.267	.554	.640	−.001
2. Post-test		.459	.498	.728	.313	.112	.337	.264	.079	.361	.697	.437	.093	.505	.605	.104
3. Progress			.977	−.138	.376	.032	.150	−.178	−.221	−.144	.091	.062	−.206	−.028	.009	.135
4. Act/Poss Progr.				−.129	.340	.039	.132	−.119	−.214	−.123	.115	.015	−.233	−.049	.030	.084
5. Re-test					.675	.172	.289	.265	.226	.441	.719	.450	.323	.623	.599	.098
6. Ret. Progress						.219	.259	−.059	.010	.127	.383	.257	.214	.358	.198	.203
7. Pupil Attitude							−.116	.089	−.320	−.235	.048	−.076	−.127	−.063	.048	.050
8. Verbal DBA								−.081	.025	.530	.553	.486	.206	.514	.488	.408
9. Inductive DBA									.134	.489	.404	.125	.382	.382	.149	−.013
10. Spatial DBA										.739	.160	.027	.416	.255	.054	−.012
11. Total DBA		N≈57									.583	.334	.556	.615	.378	.216
12. Grades Engl.												.546	.425	.819	.659	.266
13. Grades Sw.													.390	.801	.410	.301
14. Grades Maths														.772	.096	.140
15. Grades Total															.481	.293
16. PACT																.284
17. Social Class																

APPENDIX 4

Table VII *GUME 4.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. Pre-test	.902	.307	.627	.181	.586	.408	.220	.522	.697	.633	.492	.682	.729	.750	.676	.689	.832	.667	−.214
2. Post-test		.688	.885	.264	.604	.418	.208	.529	.735	.660	.517	.719	.763	.784	.734	.736	.882	.718	−.219
3. Progress			.896	.258	.345	.238	.069	.283	.448	.380	.316	.434	.457	.466	.477	.452	.542	.454	−.104
4. Act/Poss Progr.				.267	.467	.330	.139	.404	.591	.522	.419	.578	.624	.616	.579	.591	.704	.577	−
5. Pupil Attitude					.105	.077	.053	.097	.206	.160	.099	.171	.162	.143	.228	.246	.226	.246	−.126
6. Verbal DBA						.380	.264	.701	.564	.609	.478	.624	.636	.598	.576	.533	.684	.523	−.200
7. Inductive DBA							.446	.799	.372	.420	.554	.508	.404	.391	.374	.411	.461	.370	−.164
8. Spatial DBA								.768	.181	.218	.397	.301	.248	.211	.240	.237	.272	.203	−.043
9. Total DBA									.476	.534	.618	.616	.553	.517	.511	.508	.610	.474	−.176
10. Grades Engl.										.807	.610	.911	.637	.709	.664	.668	.784	.545	−.226
11. Grades Sw.											.608	.900	.600	.652	.586	.651	.729	.469	−.175
12. Grades Maths												.833	.503	.540	.489	.515	.599	.391	−.206
13. Grades Total		N≈574											.656	.715	.656	.691	.797	.530	−.233
14. Nat. test EL														.687	.637	.616	.848	.631	−.257
15. Nat. test EM															.577	.724	.886	.604	−.174
16. Nat. test EA																.618	.824	.660	−.189
17. Nat. test EU																	.863	.561	−.174
18. Nat. test Total																		.717	−.229
19. PACT																			−.238
20. Social Class																			

APPENDIX 4

Table VIII *GUME 5, sk.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Pre-test	.847	−.356	−.047	−.171	.551	.276	.179	.452	.721	.571	.481	.675	.684	.658	.627	.748	.481	−.244
2. Post-test		.195	.452	−.094	.553	.340	.250	.517	.683	.557	.529	.672	.690	.656	.600	.745	.477	−.263
3. Progress			.880	.158	−.030	.096	.113	.087	−.130	−.076	.041	−.067	−.049	−.057	−.102	−.066	−.044	−.006
4. Act/Poss Progr.				.126	.122	.177	.177	.222	.077	.098	.208	.155	.149	.117	.074	.139	.098	−
5. Pupil Attitude					.024	.137	.042	.085	−.167	−.094	−.049	−.094	−.155	−.242	−.236	−.239	−.057	−.022
6. Verbal DBA						.248	.136	.601	.506	.616	.307	.546	.618	.517	.505	.626	.457	−.146
7. Inductive DBA							.437	.771	.231	.226	.344	.307	.352	.223	.236	.294	.181	−.050
8. Spatial DBA								.772	.114	.111	.367	.239	.199	.231	.181	.217	.108	−.089
9. Total DBA									.379	.410	.471	.489	.520	.444	.413	.515	.333	−.143
10. Grades Engl.										.696	.508	.834	.624	.647	.514	.691	.480	−.206
11. Grades Sw.											.530	.838	.561	.512	.415	.576	.324	−.145
12. Grades Maths.												.802	.448	.450	.369	.483	.215	−.259
13. Grades Total		N≈235											.637	.613	.507	.676	.377	−.211
14. Nat. test EL														.665	.677	.900	.494	−.207
15. Nat. test EM															.629	.885	.405	−.277
16. Nat. test EA																.841	.519	−.243
17. Nat. test Total																	.518	−.266
18. PACT																		−.087
19. Social Class																		

APPENDIX 4

Table IX *GUME 5, ak.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Pre-test	.629	−.259	−.175	−.064	.366	.127	.066	.219	.425	.324	.420	.510	.477	.440	.399	.506	.458	.240
2. Post-test		.587	.649	.054	.491	.272	.276	.443	.371	.371	.386	.488	.628	.487	.554	.652	.532	.165
3. Progress			.989	.134	.221	.203	.273	.317	.022	.127	.045	.080	.283	.145	.274	.283	.193	−.049
4. Act/Poss Progr.				.142	.236	.194	.268	.315	.057	.133	.068	.103	.305	.187	.291	.313	.224	
5. Pupil Attitude					.091	.073	.110	.123	−.078	−.005	.061	−.012	−.030	−.056	−.019	−.037	−.046	−.092
6. Verbal DBA						.284	.249	.604	.290	.330	.363	.419	.522	.382	.417	.523	.389	−.122
7. Inductive DBA							.373	.777	.014	.164	.301	.189	.287	.247	.286	.325	.178	−.091
8. Spatial DBA								.794	−.056	.147	.339	.189	.172	.158	.211	.216	.123	−.075
9. Total DBA									.073	.267	.448	.331	.400	.328	.386	.443	.283	−.124
10. Grades Engl.										.407	.391	.784	.461	.300	.378	.449	.455	.086
11. Grades Sw.											.418	.713	.449	.272	.333	.415	.225	−.008
12. Grades Maths.												.813	.486	.305	.428	.483	.233	−.018
13. Grades Total		N≈152											.612	.385	.495	.589	.396	.029
14. Nat. test EL														.565	.710	.900	.643	.013
15. Nat. test EM															.519	.758	.456	.135
16. Nat. test EA																.899	.688	.008
17. Nat. test Total																	.717	.047
18. PACT																		.036
19. Social Class																		

APPENDIX 4

Table X *GUME A.* Intercorrelations between the main variables.

	2	3	4	5	6	7	8	9	10	11	12
1. Pre-test	.762	−.019	.301	.714	.045	.211	.196	−.192	.333	.781	.785
2. Post-test		.633	.832	.749	.163	.186	.190	−.197	.347	.643	.729
3. Progress			.924	.313	.201	.033	.058	−.074	.140	.059	.187
4. Act/Poss Progr.				.463	.199	.063	.084	−.115	.229	.273	.418
5. Oral Test					.222	.309	.250	−.324	.287	.746	.701
6. Pupil Attitude						−.059	.028	.044	−.006	.007	−.016
7. Other subjects							.324	−.367	.006	.355	.161
8. Work								−.032	.005	.306	.244
9. Age		N≈125							.262	−.429	.063
10. F-test Verbal										.205	.418
11. PACT											.634
12. Diagn. Engl.											

APPENDIX 5

Step-wise multiple regression analysis. Dependent variable: posttest (the concomitant variables are presented in order as they contribute to increase in R).

	1	+	2	+	3	+	4	+	5	
GUME 1 sk	Pretest	.864	Grades tot.	.888	Grades Eng.	.889	DBA spatial	.890	DBA verbal	.890
GUME 1 ak	Pretest	.797	DBA inductive	.814	Grades Eng.	.826	DBA spatial	.837	Grades tot.	.838
GUME 2 sk	Pretest	.814	PACT	.836	Grades Eng.	.849	DBA verbal	.854	DBA inductive	.855
GUME 2 ak	Grades tot.	.621	PACT	.738	Pretest	.780	Grades Eng.	.782	DBA total	.784
GUME 3 sk	Pretest	.799	Grades tot.	.830	DBA verbal	.836	Grades math.	.840	DBA inductive	.842
GUME 3 ak	Grades Eng.	.697	Pretest	.765	Grades math.	.796	–	–	–	-
GUME 4	Pretest	.902	PACT	.915	Grades Eng.	.924	DBA verbal	.924	DBA spatial	.924
GUME 5 sk	Pretest	.847	DBA total	.860	Grades tot.	.867	PACT	.869	Grades Sw.	.870
GUME 5 ak	Pretest	.629	DBA total	.703	PACT	.734	Grades Sw.	.741	DBA verbal	.745
GUME A	Pretest	.762	Diagn. Eng.	.791	PACT	.793	F-test verb.	.795	–	–

APPENDIX 6

Critical scores for dividing each sample into three equal parts.

	Pretest scores			DBA scores		
	Lower	Middle	Upper	Lower	Middle	Upper
GUME 1	– 55	56–71	72 –	– 13	14–17	18 –
GUME 2	– 51	52–66	67 –	– 13	14–17	18 –
GUME 3	– 72	73–91	92 –	– 12	13–17	18 –
GUME 4	– 41	42–58	59 –	– 14	15–18	19 –
GUME 5	– 36	37–55	56 –	– 13	14–17	18 –
				F-test (verbal) scores		
GUME A	– 46	47–60	61 –	– 47	48–57	58 –

APPENDIX 7

Table 1 *GUME 1* Analysis of variance, two-way classification
Dependent variable: Posttest

Pretest	Im	Ee	Es	Tot.
U	94.58 (31)	94.89 (36)	95.89 (38)	95.16 (105)
M	73.66 (32)	68.47 (36)	71.36 (44)	71.09 (112)
L	56.17 (29)	49.79 (47)	52.87 (38)	52.44 (114)
Tot.	75.20 (92)	69.08 (119)	73.27 (120)	72.30 (331)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	1730334.000	*********	12032.555
A,B & G after M	8	101249.000	12656.125	88.009
Residual Error	322	46305.000	143.804	
Total	331	1877888.000		
Mean	1	1730334.000	*********	12032.555
A after M	2	100014.000	50007.000	347.743
B after M & A	2	793.000	396.500	2.757
G after M, A & B	4	442.000	110.500	0.768
Residual Error	322	46305.000	143.804	
Total	331	1877888.000		
Mean	1	1730334.000	*********	12032.555
B after M	2	2116.000	1058.000	7.357
A after M & B	2	98691.000	49345.500	343.143
G after M, B & A	4	442.000	110.500	0.768
Residual Error	322	46305.000	143.804	
Total	331	1877888.000		

Effects to be included in model: Row effects only

APPENDIX 7

Table 11 *GUME 2.* Analysis of variance, two-way classification
Dependent variable: Posttest

Pretest	Im	Ee	Es	Tot.
U	95.42 (40)	94.12 (41)	94.25 (28)	94.63 (109)
M	72.86 (36)	72.59 (49)	75.76 (29)	73.48 (114)
L	57.52 (46)	60.42 (40)	60.28 (36)	59.29 (122)
Tot.	74.48 (122)	75.64 (130)	75.33 (93)	75.14 (345)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	1948131.000	*********	10483.719
A, B & G after M	8	72865.000	9108.125	49.015
Residual Error	336	62437.000	185.824	
Total	345	2083433.000		
Mean	1	1948131.000	*********	10483.719
A after M	2	72393.000	36196.500	194.789
B after M & A	2	133.000	66.500	0.358
G after M, A & B	4	339.000	84.750	0.456
Residual Error	336	62437.000	185.824	
Total	345	2083433.000		
Mean	1	1948131.000	*********	10483.719
B after M	2	90.000	45.000	0.242
A after M & B	2	72436.000	36218.000	194.904
G after M, B & A	4	339.000	84.750	0.456
Residual Error	336	62437.000	185.824	
Total	345	2083433.000		

Effects to be included in model: Row effects only

APPENDIX 7

Table III *GUME 3.* Analysis of variance, two-way classification
Dependent variable: Posttest

Pretest	Im	Ee	Es	Tot.
U	104.04 (27)	110.53 (19)	114.28 (25)	109.38 (71)
M	93.31 (16)	92.80 (25)	91.85 (33)	92.49 (74)
L	74.30 (23)	70.64 (39)	71.40 (20)	71.85 (82)
Tot.	91.08 (66)	86.45 (83)	93.79 (78)	90.32 (227)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	1851681.000	********	14266.352
A, B & G after M	8	55726.000	6965.750	53.668
Residual Error	218	28295.000	129.794	
Total	227	1935702.000		
Mean	1	1851681.000	********	14266.352
A after M	2	54105.000	27052.500	208.427
B after M & A	2	202.000	101.000	0.778
G after M, A & B	4	1419.000	354.750	2.733
Residual Error	218	28295.000	129.794	
Total	227	1935702.000		
Mean	1	1851681.000	********	14266.352
B after M	2	2226.000	1113.000	8.575
A after M & B	2	52081.000	26040.500	200.630
G after M, B & A	4	1419.000	354.750	2.733
Residual Error	218	28295.000	129.794	
Total	227	1935702.000		

Effects to be included in model: Row effects and interaction

APPENDIX 7

Table IV *GUME 4.* Analysis of variance, two-way classification
Dependent variable: Posttest

Pretest	Im	Ee	Es	Tot.
U	95.73 (49)	99.72 (67)	97.98 (66)	98.02 (182)
M	62.80 (61)	66.35 (65)	66.49 (70)	65.30 (196)
L	46.30 (70)	43.80 (61)	44.98 (65)	45.09 (196)
Tot.	65.35 (180)	70.78 (193)	69.88 (201)	68.77 (574)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	2718976.000	*********	10069.359
A, B & G after M	8	269186.000	33648.250	124.612
Residual Error	565	152834.000	270.025	
Total	574	3140996.000		
Mean	1	2718976.000	*********	10069.359
A after M	2	267983.000	133991.500	496.219
B after M & A	2	263.000	131.500	0.487
G after M, A & B	4	940.000	235.000	0.870
Residual Error	565	152834.000	270.025	
Total	574	3140996.000		
Mean	1	2718976.000	*********	10069.359
B after M	2	3138.000	1569.000	5.811
A after M & B	2	265108.000	132554.000	490.896
G after M, B, & A	4	940.000	235.000	0.870
Residual Error	565	152834.00	270.025	
Total	574	3140996.000		

Effects to be included in model: Row effects only

APPENDIX 7

Table V *GUME 5.* Analysis of variance, two-way classification
Dependent variable: Posttest

Pretest	Im	Ee	Es	Tot.
U	77.00 (41)	75.07 (57)	75.71 (35)	75.83 (133)
M	50.28 (43)	55.05 (39)	52.32 (44)	52.47 (126)
L	34.92 (36)	32.07 (45)	36.21 (47)	34.39 (128)
Tot.	54.80 (120)	55.81 (141)	52.81 (126)	54.52 (387)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	1150303.000	*********	11845.867
A, B & G after M	8	113784.000	14223.000	146.469
Residual Error	378	36706.000	97.106	
Total	387	1300793.000		
Mean	1	1150303.000	*********	11845.867
A after M	2	112819.000	56409.500	580.907
B after M & A	2	66.000	33.000	0.340
G after M, A & B	4	899.000	224.750	2.314
Residual Error	378	36706.000	97.106	
Total	387	1300793.000		
Mean	1	1150303.000	*********	11845.867
B after M	2	613.000	306.500	3.156
A after M & B	2	112272.000	56136.000	578.091
G after M, B & A	4	899.000	224.750	2.314
Residual Error	378	36706.000	97.106	
Total	387	1300793.000		

Effects to be included in model: Row effects only

APPENDIX 7

Table VI *GUME ADULTS.* Analysis of variance, two-way classification
Dependent variable: Posttest

Pretest	Im	Es	Tot.
U	92.88 (17)	97.00 (20)	95.11 (37)
M	66.71 (21)	79.44 (25)	73.63 (46)
L	52.95 (19)	58.74 (23)	56.12 (42)
Tot.	69.93 (57)	77.60 (68)	74.10 (125)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	686425.313	*********	4035.053
A, B & G after M	5	32271.938	6454.387	37.941
Residual Error	119	20243.750	170.116	
Total	125	738941.000		
Mean	1	686425.313	*********	4035.053
A after M	2	29918.875	14959.438	87.937
B after M & A	1	1909.625	1909.625	11.225
G after M, A & B	2	443.438	221.719	1.303
Residual Error	119	20243.750	170.116	
Total	125	738941.000		
Mean	1	686425.313	*********	4035.053
B after M	1	1825.500	1825.500	10.731
A after M & B	2	30003.000	15001.500	88.184
G after M, B & A	2	443.438	221.719	1.303
Residual Error	119	20243.750	170.116	
Total	125	738941.000		

Effects to be included in model: Row and column effects

APPENDIX 7

Table VII *GUME I.* Analysis of variance, two-way classification
Dependent variable: Posttest

DBA scores	Im	Ee	Es	Tot.
U	88.61 (28)	82.71 (28)	89.51 (43)	87.33 (99)
M	75.97 (32)	75.61 (38)	73.13 (31)	74.96 (101)
L	62.48 (27)	55.07 (46)	55.42 (38)	56.99 (111)
Tot.	75.85 (87)	68.95 (112)	73.41 (112)	72.49 (311)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	1634040.000	*********	5632.242
A, B & G after M	8	51170.000	6396.250	22.047
Residual Error	302	87617.000	290.122	
Total	311	1772827.000		
Mean	1	1634040.000	*********	5632.242
A after M	2	49093.000	24546.500	84.607
B after M & A	2	880.000	440.000	1.517
G after M, A & B	4	1197.000	299.250	1.031
Residual Error	302	87617.000	290.122	
Total	311	1772827.000		
Mean	1	1634040.000	*********	5632.242
B after M	2	2484.000	1242.000	4.281
A after M & B	2	47489.000	23744.500	81.843
G after M, B & A	4	1197.000	299.250	1.031
Residual Error	302	87617.000	290.122	
Total	311	1772827.000		

Effects to be included in model: Row effects only

APPENDIX 7

Table VIII *GUME 2.* Analysis of variance, two-way classification
Dependent variable: Posttest

DBA scores	Im	Ee	Es	Tot.
U	89.61 (33)	87.49 (41)	94.00 (22)	89.71 (96)
M	77.80 (35)	77.51 (43)	76.22 (27)	77.28 (105)
L	61.51 (43)	63.20 (41)	63.23 (35)	62.60 (119)
Tot.	75.00 (111)	76.09 (125)	75.46 (84)	75.55 (320)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	1826345.000	********	6631.410
A, B & G after M	8	40252.000	5031.500	18.269
Residual Error	311	85652.000	275.408	
Total	320	1952249.000		

Mean	1	1826345.000	********	6631.410
A after M	2	39524.000	19762.000	71.755
B after M & A	2	127.000	63.500	0.231
G after M, A & B	4	601.000	150.250	0.546
Residual Error	311	85652.000	275.408	
Total	320	1952249.000		

Mean	1	1826345.000	********	6631.410
B after M	2	70.000	35.000	0.127
A after M & B	2	39581.000	19790.500	71.859
G after M, B & A	4	601.000	150.250	0.546
Residual Error	311	85652.000	275.408	
Total	320	1952249.000		

Effects to be included in model: Row effects only

APPENDIX 7

Table IX *GUME 3.* Analysis of variance, two-way classification
Dependent variable: Posttest

DBA scores	Im	Ee	Es	Tot.
U	101.30 (23)	105.50 (24)	110.85 (20)	105.66 (67)
M	94.05 (19)	81.48 (29)	96.48 (29)	90.23 (77)
L	72.38 (16)	71.84 (25)	75.08 (24)	73.17 (65)
Tot.	90.95 (58)	85.78 (78)	93.38 (73)	89.87 (209)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	1688043.000	********	9568.590
A, B & G after M	8	39585.000	4948.125	28.048
Residual Error	200	35283.000	176.415	
Total	209	1762911.000		
Mean	1	1688043.000	********	9568.590
A after M	2	34837.000	17418.500	98.736
B after M & A	2	2617.000	1308.500	7.417
G after M, A & B	4	2131.000	532.750	3.020
Residual Error	200	35283.000	176.415	
Total	209	1762911.000		
Mean	1	1688043.000	********	9568.590
B after M	2	2271.000	1135.500	6.437
A after M & B	2	35183.000	17591.500	99.717
G after M, B & A	4	2131.000	532.750	3.020
Residual Error	200	35283.000	176.415	
Total	209	1762911.000		

Effects to be included in model: Row and column effects and interaction

APPENDIX 7

Table X *GUME 4.* Analysis of variance, two-way classification
Dependent variable: Posttest

DBA scores	Im	Ee	Es	Tot.
U	83.15 (53)	85.47 (70)	84.39 (67)	84.44 (190)
M	64.00 (74)	71.26 (61)	69.94 (65)	68.14 (200)
L	48.20 (45)	53.65 (62)	55.85 (65)	53.05 (172)
Tot.	65.77 (172)	70.76 (193)	70.20 (197)	69.04 (562)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	2678442.000	*********	4599.031
A, B & G after M	8	93021.000	11627.625	19.965
Residual Error	553	322063.000	582.392	
Total	562	3093526.000		
Mean	1	2678442.000	*********	4599.031
A after M	2	89197.000	44598.500	76.578
B after M & A	2	2983.000	1491.500	2.561
G after M, A & B	4	841.000	210.250	0.361
Residual Error	553	322063.000	582.392	
Total	562	3093526.000		
Mean	1	2678442.000	*********	4599.031
B after M	2	2677.000	1338.500	2.298
A after M & B	2	89503.000	44751.500	76.841
G after M, B & A	4	841.000	210.250	0.361
Residual Error	553	322063.000	582.392	
Total	562	3093526.000		

Effects to be included in model: Row effects only

APPENDIX 7

Table XI *GUME 5.* Analysis of variance, two-way classification
Dependent variable: Posttest

DBA scores	Im	Ee	Es	Tot.
U	72.24 (42)	73.68 (41)	67.58 (33)	71.42 (116)
M	52.42 (31)	58.41 (37)	53.77 (31)	55.08 (99)
L	38.51 (35)	40.91 (47)	42.81 (37)	40.80 (119)
Tot.	55.62 (108)	56.84 (125)	54.27 (101)	55.67 (334)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	1035028.688	*********	4631.668
A, B & G after M	8	56877.313	7109.664	31.815
Residual Error	325	72627.000	223.468	
Total	334	1164533.000		
Mean	1	1035028.688	*********	4631.668
A after M	2	55136.313	27568.156	123.365
B after M & A	2	630.000	315.000	1.410
G after M, A & B	4	1111.000	277.750	1.243
Residual Error	325	72627.000	223.468	
Total	334	1164533.000		
Mean	1	1035028.000	*********	4631.668
B after M	2	370.188	185.094	0.828
A after M & B	2	55396.125	27698.063	123.947
G after M, B & A	4	1111.000	277.750	1.243
Residual Error	325	72627.000	223.468	
Total	334	1164533.000		

Effects to be included in model: Row effects only

APPENDIX 7

Table XII *GUME ADULTS.* Analysis of variance, two-way classification
Dependent variable: Posttest

F-test Verbal	Im	Es	Tot.
U	78.12 (17)	85.82 (17)	81.97 (34)
M	72.85 (13)	81.52 (25)	78.55 (38)
L	59.94 (18)	69.38 (21)	65.03 (39)
Tot.	69.88 (48)	78.63 (63)	74.85 (111)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	621827.563	*********	1653.700
A, B & G after M	5	8020.125	1604.025	4.266
Residual Error	105	39482.313	376.022	
Total	111	669330.000		
Mean	1	621827.563	*********	1653.700
A after M	2	6009.000	3004.500	7.990
B after M & A	1	1997.688	1997.688	5.313
G after M, A & B	2	13.438	6.719	0.018
Residual Error	105	39482.313	376.022	
Total	111	669330.000		
Mean	1	621827.563	*********	1653.700
B after M	1	2090.500	2090.500	5.560
A after M & B	2	5916.188	2958.094	7.867
G after M, B & A	2	13.438	6.719	0.018
Residual Error	105	39482.313	376.022	
Total	111	669330.000		

Effects to be included in model: Row and column effects

APPENDIX 7

Table XIII *GUME ADULTS.* Analysis of variance, two-way classification
Dependent variable: Posttest

			Im	Es	Tot.
Age	41–	U	58.57 (7)	66.38 (13)	63.65 (20)
	26–40	M	69.73 (30)	79.54 (46)	75.67 (76)
	–25	L	74.20 (20)	83.89 (9)	77.21 (29)
		Tot.	69.93 (57)	77.60 (68)	74.10 (125)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	686425.313	********	1728.542
A,B & G after M	5	5259.313	1051.862	2.649
Residual Error	119	47256.375	397.112	
Total	125	738941.000		
Mean	1	686425.313	********	1728.542
A after M	2	2651.500	1325.750	3.338
B after M & A	1	2593.000	2593.000	6.530
G after M, A & B	2	14.813	7.406	0.019
Residual Error	119	47256.375	397.112	
Total	125	738941.000		
Mean	1	686425.313	********	1728.542
B after M	1	1825.500	1825.500	4.597
A after M & B	2	3419.000	1709.500	4.305
G after M, B & A	2	14.813	7.406	0.019
Residual Error	119	47256.375	397.112	
Total	125	738941.000		

Effects to be included in model: Row and column effects

APPENDIX 7

Table XIV *GUME ADULTS.* Analysis of variance, two-way classification
Dependent variable: Posttest

	Im	Es	Tot.
Females	68.89 (28)	77.76 (55)	74.77 (83)
Males	70.93 (29)	76.92 (13)	72.79 (42)
Total	69.93 (57)	77.60 (68)	74.10 (125)

Source of variation	D.F.	Sum of sq.	Mean square	F-statistic
Mean	1	686425.313	*********	1640.686
A, B & G after M	3	1892.063	630.688	1.507
Residual Error	121	50623.625	418.377	
Total	125	738941.000		
Mean	1	686425.313	*********	1640.686
A after M	1	109.813	109.813	0.262
B after M & A	1	1732.125	1732.125	4.140
G after M, A & B	1	50.125	50.125	0.120
Residual Error	121	50623.625	418.377	
Total	125	738941.000		
Mean	1	686425.313	*********	1640.686
B after M	1	1825.500	1825.500	4.363
A after M & B	1	16.438	16.438	0.039
G after M, B & A	1	50.125	50.125	0.120
Residual Error	121	50623.625	418.377	
Total	125	738941.000		

Effects to be included in model: Column effects only

APPENDIX 8

Analyses of covariance for each sex in the GUME ADULTS sample.
Dependent variable: Posttest
Covariate: Pretest

	Adjusted means			ss'y				
	Im	Es	F-ratio	be-tween	with-in	df	b_w	p
Females	67.21	78.62	13.434	2403	14310	1/80	1.083	<.01
Males	70.19	78.57	6.816	627	3589	1/39	931	<.05

APPENDIX 9

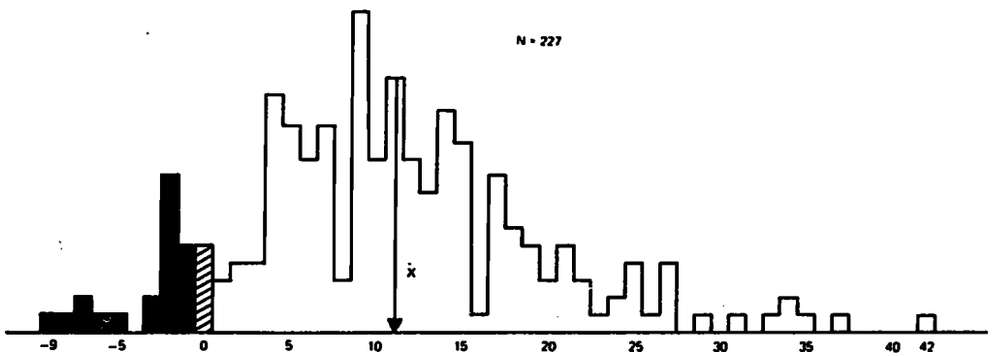


Figure I. Distribution of progress scores, GUME 1 sk

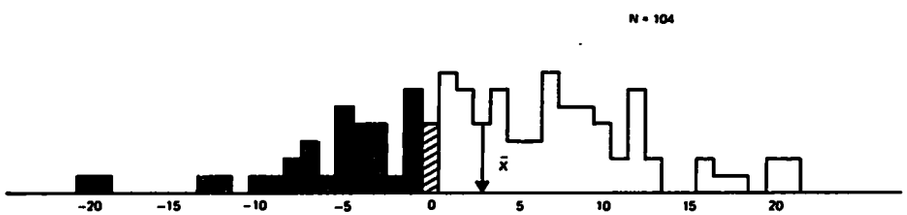


Figure II. Distribution of progress scores, GUME 1 ak

APPENDIX 9

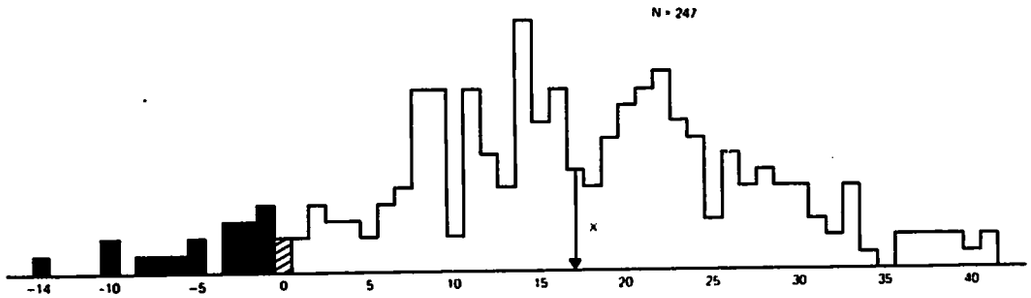


Figure III. Distribution of progress scores, GUME 2 sk

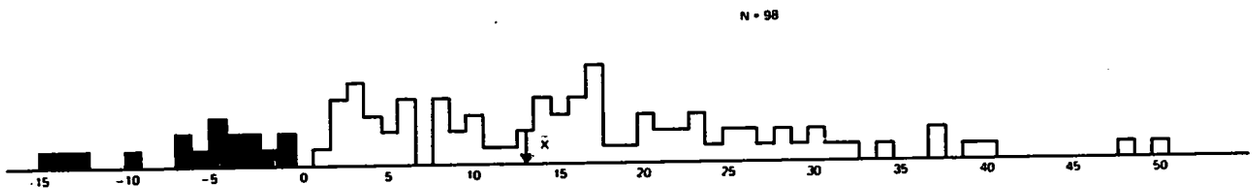


Figure IV. Distribution of progress scores, GUME 2 ak

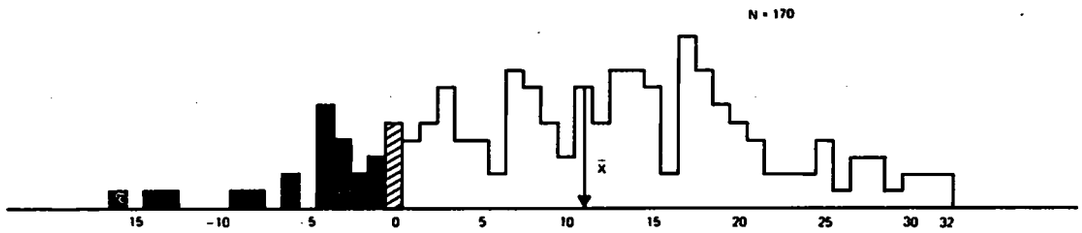


Figure V. Distribution of progress scores, GUME 3 sk

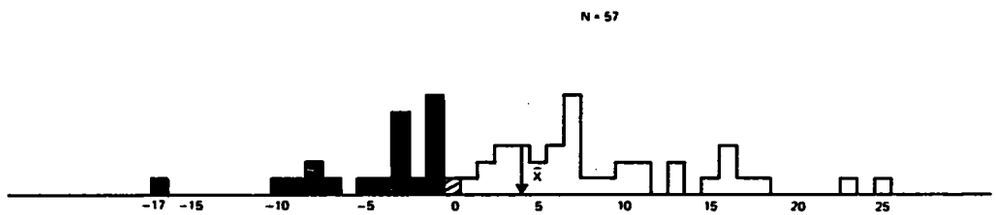


Figure VI. Distribution of progress scores, GUME 3 ak

APPENDIX 9

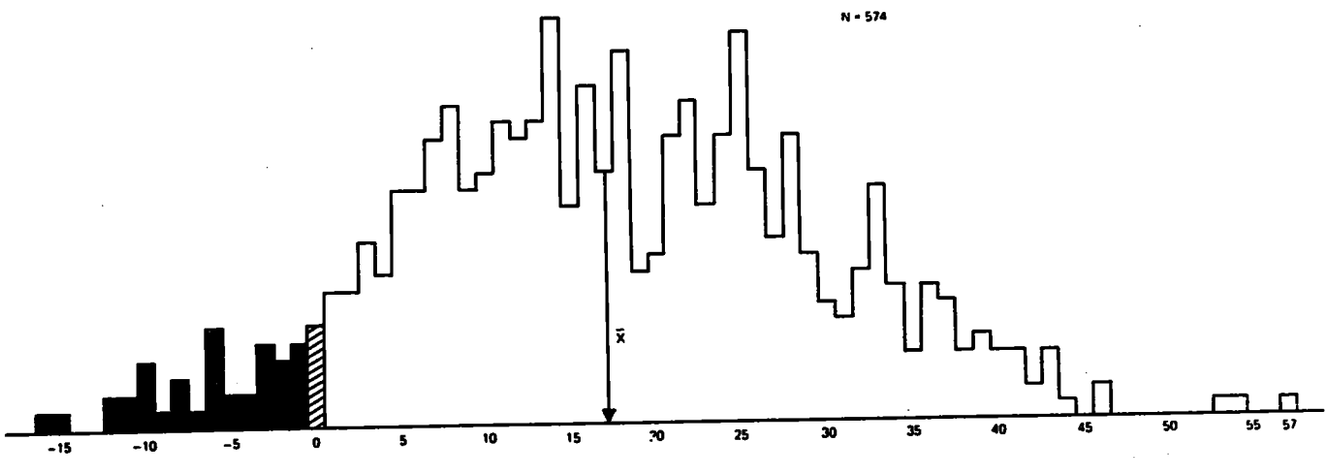


Figure VII. Distribution of progress scores, GUME 4

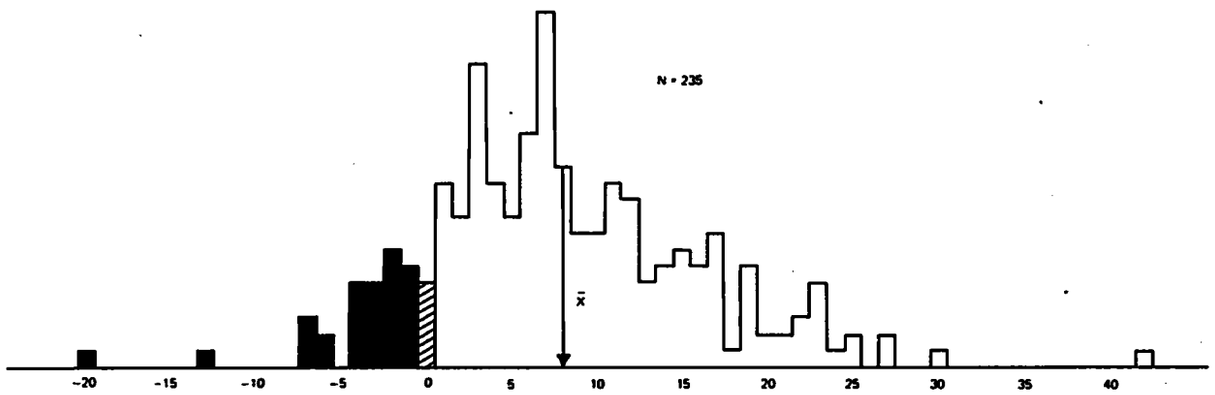


Figure VIII, Distribution of progress scores, GUME 5 sk

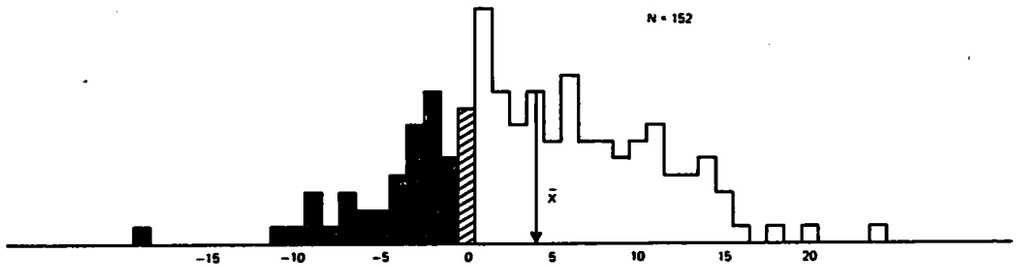


Figure IX. Distribution of progress scores, GUME 5 ak

APPENDIX 9

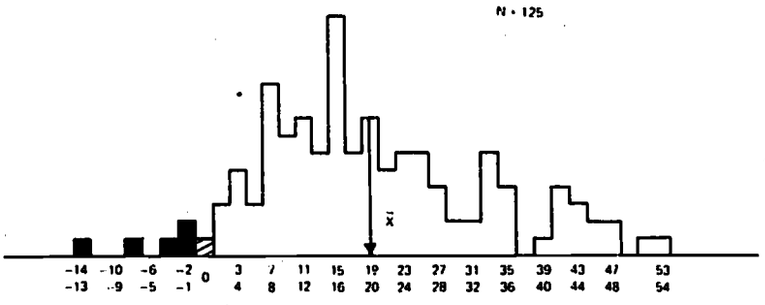
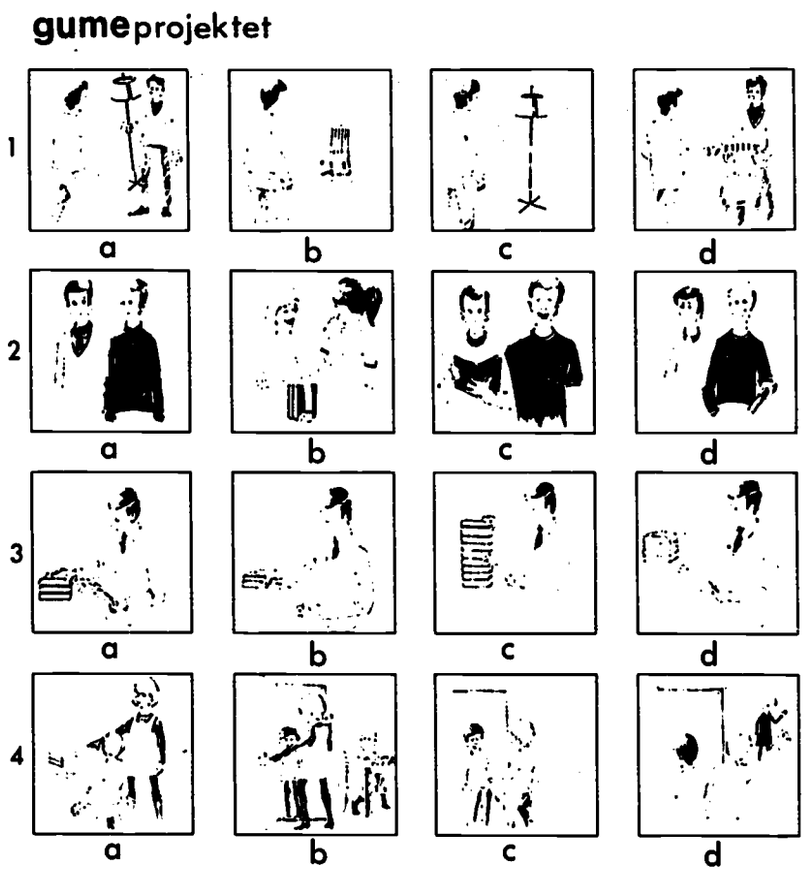


Figure X. Distribution of progress scores, GUME A

APPENDIX 10

Sample items illustrating the testing procedure in PACT (Pictoral Auditory Comprehension Test)



PACT

The four first items of the test are presented above. As a typical example the auditory stimulus of item No. 4 is given (the following is heard from the tape):

"He'll come when he's finished his homework."

The pupils mark their answer on a separate sheet. (It is B which is correct, of course.)

APPENDIX 11

Items number 1, 2, 5 and 6 in the student attitude test administered in GUME 1–3

Item No 1:

During the project lessons I learnt English
_____ much better than during ordinary lessons
_____ somewhat better than during ordinary lessons
_____ about as much as during ordinary lessons
_____ somewhat less than during ordinary lessons
_____ much less than during ordinary lessons

Item No. 2:

The project lessons were
_____ much more enjoyable than ordinary lessons
_____ somewhat more enjoyable than ordinary lessons
_____ about as enjoyable as ordinary lessons
_____ somewhat duller than ordinary lessons
_____ much duller than ordinary lessons

Item No. 5:

I think the headsets worked
_____ very well
_____ well
_____ not very good
_____ badly

Item No. 6:

In general I think the sound quality was
_____ very good (easy to hear)
_____ good
_____ rather bad
_____ very bad (difficult to hear)

APPENDIX 12

Recording manuscript for the explanations given in one lesson in GUME 1 (the figures 1-10 refer to the frames in the manuscript for slides).

Lesson 2

GROUP: Ee

Now we shall try to see what you really do when you ask a question in English. But first let us start with four English senteces (1). – Oh no, that can't be right, you can't say that in English. We must add something. (2). That looks better. Let's read these sentences: He looks, He can look. But then, no that is still not correct. We must add a little more (3) - like that. Now: He looks, He can look, He has looked, He is looking. They are four correct English sentences. But now we'll make them into questions. Let us start with the question marks (4) like that. We'll put one in front of the sentences too. Now we must change something because these are not correct questions. We'll put the red words in a frame (5) because it's with them that we must do something. We must move them to the beginning of the sentences (6) as the arrows show us. In English the black words can never change places. But now there is no red word in the first sentence, so we'll move the *s* first (7) as this arrow shows and then it looks like this (8). Now let us move the words in the frame to the beginning of the sentences, where the question mark is, like this (9). Now we have three fine sentences, three questions: *Can he look, Has he looked, Is he looking.* But the first one is no good, you can't say that: *s he look.* What we must do now is to add something to the *s*. Let us do as English people always do, let's take the word *do*. We'll have to spell it *d- -e* (10) and what we get is this: *Does he look.* Now we'll read these sentences together: Does he look, Can he look, Has he looked, Is he looking. Good.

GROUP: Es

Nu skall vi ta och se litet på hur man gör när man skall ställa frågor på engelska. Vi skall också jämföra med hur vi gör på svenska. (1) Här är fyra engelska meningar. Men så där kan de ju inte se ut. Vi måste lägga till litet

APPENDIX 12

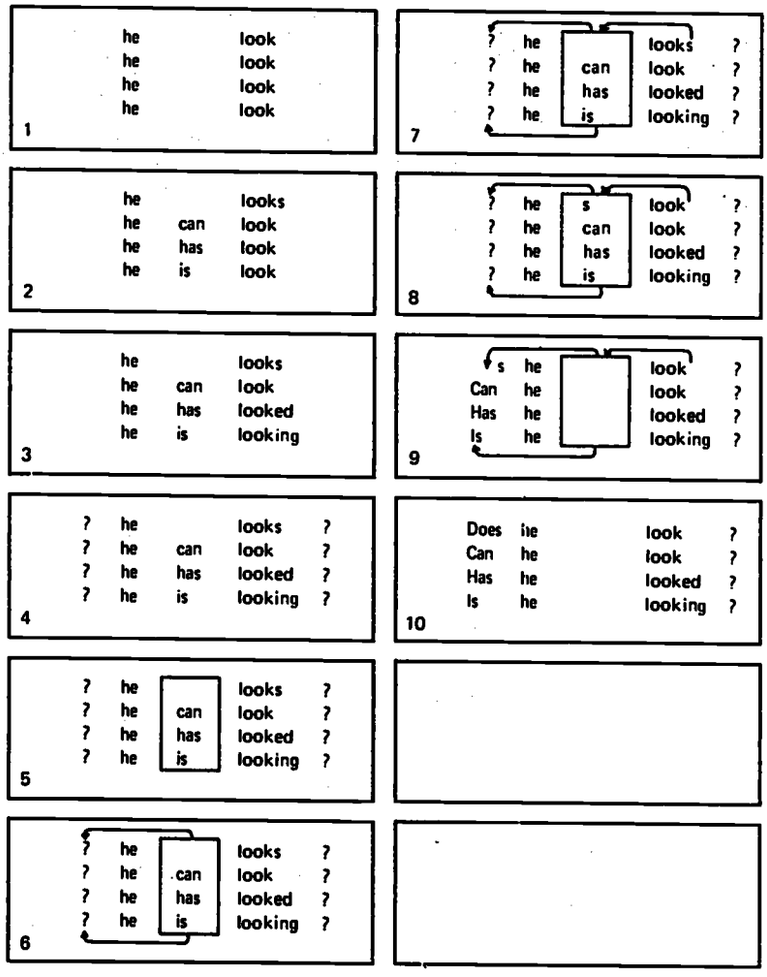
(continued)

grand. (2) Så, nu ser åtminstone två av dem bra ut; "He looks, He can look". Men de andra låter inte så bra. Vi tar och lägger till litet mer. (3) Så där, nu blev det bra. "He looks, He can look, He has looked, He is looking." På svenska skulle de heta: "Han tittar, Han kan titta, Han har tittat, Han tittar." Nu skall vi göra frågor av dem. Låt oss sätta in frågetecken. (4) Vi sätter ett framför också. De ord som nu intresserar oss är de som är röda. Vi tar och sätter en ram om dem (5) – så där. Det som nu skall hända är att de där orden skall flyttas längst fram i meningen. De svarta orden däremot får aldrig flyttas på engelska. Vi tar och markerar med pilar. (6) Men i första raden finns ju inget i rutan. Vi tar och flyttar in det röda "s"-et. (7) Här kommer vi nu till en stor skillnad mellan svenskan och engelskan som vi skall lägga noga märke till. På engelska måste de två svarta orden alltid stå kvar som det står. Det är bara "s"-et som flyttar på sig. På svenska däremot kan man ju flytta hela ordet "tittar" och säga "Tittar han". (8). Och nu skall vi alltså göra frågor genom att flytta orden i rutan som pilarna visar (9). Och detta är vad vi får. Vi börjar med den andra raden: "Can he look, Has he looked, Is he looking". Det låter bra och är också riktigt. Men den första ser konstig ut. Så kan man ju inte säga: "s he look". Vi måste lägga till något. Vi gör väl som engelsmännen själva brukar göra, vi lägger till verbet "do". (10). Som du ser får vi stava det med ett extra "e", men så får vi också fram en fin mening nu: "Does he look?" Om du tänker dig den svenska meningen, "Tittar han?", så märker du skillnaden: på svenska kan de två svarta orden helt enkelt byta plats, något som aldrig kan inträffa på engelska. Ordet "does" får man alltså lägga till på engelska för att markera att det är en fråga, det betyder liksom inget här. Låt oss läsa de engelska meningarna högt tillsammans: "Does he look, Can he look, Has he looked, Is he looking". Bra.

APPENDIX 12

(continued)

Frames 1 – 10 referring to the text.



REFERENCES

Agard, B. & Dunkel, H. (1948) *An investigation of second-language teaching.* New York: Ginn and Company.

Aleamoni, L.M. & Spencer, R.E. (1969) An evaluation of the Pennsylvania foreign language project. *Modern Language Journal, 53,* No. 6, 421–428.

Alkin, M.C. & Johnson, M. (1971) *Comments on the research and development program of byrå l. 4:1.* School Research Newsletter No. 24/71 from the National Board of Education, Stockholm.

Allen, J.P.B. & van Buren, P. (Eds.) (1971) *Chomsky: Selected readings.* London: Oxford University Press.

Anastasi, A. (1958) *Differential psychology.* New York: MacMillan.

Anderson, R.C. (1969) The comparative field experiment: An illustration from high school biology. Educational Testing Service. Princeton, New Jersey: *Proceedings of the 1968 invitational conference on testing problems,* 3–30.

Andrakammarvalet i Göteborg 1968. Stickprovsundersökning. (A sample study of the Gothenburg electorate in the parliamentary election in 1968). Göteborgs stads statistiska kontor, U 1969:2.

Anisfeld, M. (1966) Psycholinguistic perspectives on language learning. *In* Valdman, A. (Ed.), 1966.

Atiqullah, M. (1964) The robustness of the covariance analysis of a one-way classification. *Biometrika, 51,* 365–372.

Ausubel, D.P. (1964) Adults versus children in second-language learning: Psychological considerations. *Modern Language Journal, 48,* No. 7, 420–424.

Baker, R.L. (1969) Curriculum evaluation. *Review of Educational Research, 39,* No. 3, 339–358.

Barrutia, R. (1966) Some pedagogical dangers in recent linguistic trends. *IRAL, 4,* No. 3, 157–164.

Bloom, B.S. (Ed.) (1956) *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: Longmans, Green and Co.

Bloomfield, L. (1942) *Outline guide for the practical study of foreign languages.* Baltimore: Linguistic Society of America 1942.

Bolinger, D. (1968) The theorist and the language teacher. *Foreign Language Annals, 2,* No. 1, 30–41.

Bosco, F.J. & Di Pietro, R.J. (1970) Instructional strategies: their psychological and linguistic bases. *IRAL, 8,* No. 1, 1–19.

Bracht, G.H. (1970) Experimental factors related to aptitude-treatment interactions. *Review of Educational Research, 40,* No. 5, 627–645.

Brooks, N. (1960) *Language and language learning.* New York: Harcourt,

Brace & World.

Campbell, T. & Stanley, J.C. (1967) Experimental and quasi-experimental designs for research on teaching. *In* Gage, N.L. (Ed.), 1967.

Carroll, J.B. (1965) The contributions of psychological theory and educational research to the teaching of foreign languages. *Modern Language Journal, 49*, No. 5, 273-281.

Carroll, J.B. (1967) Research on teaching foreign languages. *In* Gage, N.L. (Ed.), 1967.

Carroll, J.B. (1968) The psychology of language testing. *In* Davies, A. (Ed.), 1968.

Carroll, J.B. (1969 a) Modern languages. *Encyclopedia of Educational Research.* Fourth edition, 1969. MacMillan.

Carroll, J.B. (1969 b) What does the Pennsylvania foreign language research project tell us? *Foreign Language Annals, 3*, No. 2, 214-236.

Carroll, J.B. (1971) Current issues in psycholinguistics and second language teaching. *Tesol Quarterly, 5*, No. 2, 101-114

Casey, D.J. (1968) *The effectiveness of two methods of teaching English as a foreign language in some Finnish secondary schools.* Research Bulletin No. 24 from the Institute of Education, University of Helsinki.

Chastain, K.D. & Woerdehoff, F.J. (1968) A methodological study comparing the audio-lingual habit theory and the cognitive code-learning theory. *Modern Language Journal, 52*, No. 5, 268-279.

Chastain, K. (1969) The audio-lingual habit theory versus the cognitive code-learning theory: some theoretical considerations. *IRAL, 7*, No. 2, 97-106.

Chastain, K.D. (1970) A methodological study comparing the audio-lingual habit theory and the cognitive code-learning theory – a continuation. *Modern Language Journal, 54*, No. 4 257-266.

Chomsky, N. (1957) *Syntactic structures.* The Hague.

Chomsky, N. (1959) Review of Skinner's 'Verval Bahavior'. *Language, 35*, No. 1, 26-58.

Chomsky, N. (1965) *Aspects of the theory of syntax.* The MIT Press. Cambridge, Mass.

Chomsky, N. (1971) Language teaching. *In* Allen, J.P.B. & van Buren, P. (Eds.), 1971.

Clark, J.L.D. (1969) The Pennsylvania project and the "audio-lingual" vs "traditional" question. *Modern Language Journal, 53*, No. 6, 388-396.

Coleman, A. (1929) *The teaching of modern foreign languages in the United States.* New York: American and Canadian Committees on Modern Languages 1929.

Cook, V.J. (1969) The analogy between first and second language learning.

IRAL, 7, No. 3, 207-216.

Cox, D.R. (1957) The use of a concomitant variable in selecting an experimental design. *Biometrika. 44*, 150-158.

Cronbach, L.J. (1963) Course improvement through evaluation. *Teachers College Record, 64*, 672-683.

Cronbach. L.J. & Snow, R.E. (1969) *Individual differences in learning ability as a function of instructional variables.* Final report No. OEC 4-6-061269-1217, School of Education, Stanford University, California.

Cronbach, L.J. & Suppes, P. (Eds.) (1969) *Research for tomorrow's schools.* London: Macmillan.

Cronbach, L.J. & Furby, L. (1970) How we should measure "change" – or should we? *Psychological Bulletin, 74*, No. 1, 68-80.

Crothers, E. & Suppes, P. (1969) *Experiments in second-language learning.* New York & London: Academic Press.

Darian, S. (1959) Backgrounds of modern language teaching: Sweet, Jespersen, and Palmer. *Modern Language Journal, 53*, No. 8, 545-550.

Davies, A. (Ed.), 1968 *Language testing symposium.* London: Oxford University Press.

Draper, N.R. & Smith, H. (1966) *Applied regression analysis.* Wiley & Sons.

DuBois, P.H. (1962) The design of correlational studies in training. *In* Glaser, R. (Ed.), 1962.

Edwardsson, R. (1970) *Språkdebatten 1969-1970.* (The language teaching debate, 1969-1970). Stockholm: Utbildningsförlaget.

Elashoff, J.D. (1969) Analysis of covariance: A delicate instrument. *American Educational Research Journal, 6*, No. 3, 383-401.

Ellegård, A. (1968) Språkteori och språkinlärning. (Linguistic theory and language learning.) *Modern språk, 62*, No. 1, 1-24.

Ellegård, A. (1969) *Svensk-engelsk transformationsgrammatik.* (Swedish-English transformational grammar.) Lund: Gleerups.

Ellegård, A. & Lindell, E. (1970) *Direkt eller insikt?* (Direct or insight?) Lund: Gleerups.

Ellegård, A. (1971) *Språk, språkvetenskap, språkinlärning.* (Language, linguistic science, language learning.) Stockholm: Aldus.

Federer, W.T. (1955) Experimental design. New York: MacMillan.

Feldt, L.S. (1958) A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika, 23*, No. 4, 335-353.

Fletcher, H.J. (1968) Possible interpretive problems in analyses using group means as the experimental unit. *Psychological Bulletin, 69*, No. 3, 157-160.

Fodor, J.A. (1966) How to learn to talk: Some simple ways. *In* Smith, F. & Miller, G.A. (Eds.), 1966.

Fries, C.C. (1945) *Teaching and learning English as a foreign language.* Ann Arbor, Michigan: University of Michigan Press.

Gage, N.L. (Ed.) (1967) *Handbook of Research on Teaching.* Chicago. Rand McNally.

Gage, N.L. (1969) Teaching methods. *Encyclopedia of Educational Research.* Fourth edition, 1969. The MacMillan Company.

Gagné, R.M. (1965) *The conditions of learning.* New York & London: Holt, Rinehart & Winston.

Gagné, R.M. (1967) Curriculum research and the promotion of learning. *Perspectives of curriculum evaluation.* AERA Monograph Series on Curriculum Evaluation, No. 1, 19-38.

Glaser, R. (Ed.) (1962) *Training research and education.* University of Pittsburgh Press.

Glass, G.V. & Stanley, J.C. (1970) *Statistical methods in education and psychology.* Prentice-Hall.

Gourlay, N. (1953) Covariance analysis and its applications in psychological research. *British Journal of Statistical Psychology, 6,* 25-34.

Grittner, F.M. (1969) A critical re-examination of methods and materials. *Modern Language Journal, 53,* No. 7, 467-477.

Hall, P. (1969) *Hundra språklärares erfarenheter: Enkät om undervisningsmetodiken i främmande språk på grundskolans högstadium.* (The experiences of one hundred foreign language teachers: Enquiry about foreign language teaching methodology at the upper stage of the comprehensive school system. Malmö: Lärarhögskolan. Report No. 97.

Hanzeli, V.E. (1968) Linguistics and the language teacher. *Foreign Language Annals, 2,* No. 1, 42-49.

Hayes, A.S., Lamber, W.E. & Tucker, G.R. (1967) Evaluation of foreign language teaching. *Foreign Language Annals, 1,* No. 1, 22-44.

Hensjö. P-O. (1964) *Build up their English.* Stockholm: Svenska bokförlaget.

Holmberg, B. (1965) Educators or drill-sergeants? *Moderna språk, 59,* No. 1, 1-4.

Härnqvist, K. (1968) Relative changes in intelligence from 13 to 18. *Scandinavian Journal of Psychology, 9,* 50-82.

Härnqvist, K. (1969) *Nya normer för DBA årkurs 6.* (New norms for DBA, grade 6). Stockholm: Skandinaviska Testförlaget.

Härnqvist, K. (1972) *Training and career structures of educational researchers.* Report No. 18 from the Department of Education, University of Gothenburg.

Jacobsson, B. (1968) Transformationsevangeliets bärkraft. (The strength of the transformational gospel.) *Moderna språk, 62,* No. 4, 364-384.

Jakobovits, L.A. (1968) Implications of recent psycholinguistic developments for the teaching of a second language. *Language Learning,* 18, Nos. 1 & 2,

89-109.

James, C. (1969) Deeper contrastive study. *International Review of Applied Linguistics, 7,* 83-95.

Jarvis, G.A. (1968) A behavioral observation system for classroom foreign language skill acquisition activities. *Modern Language Journal, 52,* No. 6, 335-341.

Johansson, R. & Molander. L.G. (1970) *Vuxenstuderande 1968.* (Adult students 1968.) Report from the Department of Education, University of Uppsala.

Johnson, F.C. (1969) The failure of the discipline of linguistics in language teaching. *Language Learning, 19,* 235-244.

Katahn, M. & Koplin. J.H. (1968) Paradigm clash: Comment on "Some recent criticisms of behaviorism and learning theory with special reference to Breger and McGaugh and to Chomsky". *Psychological Bulletin, 69,* No. 2, 147-148.

Kerlinger, F.N. (1969) Research in education. *Encyclopedia of Educational Research.* Fourth edition, 1969. McMillan.

Kerlinger, F.N. (1970) *Foundations of behavioral research.* London: Holt, Rinehart & Winston.

Klausmeier, H.J. (1968) Research and development strategies in education. *In* Klausmeier, H.J., et.al., 1968.

Klausmeier, H.J. et.al. (1968) *Research and development strategies in theory refinement and educational improvement.* Theoretical paper No. 15. Madison: University of Wisconsin, Wis. Research and Development Center for Cognitive Learning.

Lado, R. (1961) *Language testing: The construction and use of foreign language tests.* London: Longmans.

Lenneberg, E. (Ed) (1964) *New directions in the study of language.* MIT Press.

Levin, L. & Marton, F. (1971) *Provteori och provkonstruktion.* (Test theory and test construction.) Stockholm: Almqvist & Wiksell.

Lim, K-B. (1968) *Prompting vs. confirmation, pictures vs. translation, and other variables in children's learning of grammar in a second language.* Research bulletin No. 68-43 from Educational Testing Service. Princeton, New Jersey.

Lindell, E. (1971) *Främmande språk i grundskolan* (Foreign languages in the comprehensive school.) Lund: Gleerups.

Lord, F.M. (1960) Large-sample covariance analysis when the control variable is fallible. *American Statistical Association Journal, 55,* 307-321.

Lumsdaine, A.A. (1967) Instruments and media of intruction. *In* Gage, N.L. (Ed.), 1967.

Läroplan för grundskolan (1969): "Lgr 62" (Authorized curriculum for Swed-

ish schools.) Stockholm: Skolöverstyrelsen.

Läroplan för grundskolan (1969): "Lgr 69" (ditto). Stockholm: Skolöverstyrelsen.

Läroplan för grundskolan, Lgr 69, Supplement Engelska (1969). (Ditto, Supplementary comments on English as a foreign language.) Stockholm: Skolöverstyrelsen.

Löfgren, R. (1950) *Engelsk grammatik* (English grammar). Stockholm: Natur och Kultur.

Löfgren, H. (1966) *Försök med en- och tvåspråkiga ordlistor* (Experiments with mono- and bilingual word lists). Malmö: Lärarhögskolan. Report No. 29.

Mackey, W.F. (1965) *Language teaching analysis.* London: Longmans.

Marklund, S. (1962) Folkskolebetygen som gallringsinstrument. (The elementary school grades as selection instruments.) *Pedagogisk Forskning, 4,* 169-177.

McCall, W.A. (1923) *How to experiment in education.* New York: Macmillan.

McKinnon, K.R. (1965) *An experimental study of the learning of syntax in second language learning.* Doctoral dissertation. Harvard University School of Education.

von Mentzer, C.H. (1970) *Projektet UME.* (The UME project.) Stockholm: Lärarhögskolan. Report No. 49.

Miller, G.A., Galanter, E. & Pribram, K. (1960) *Plans and the structure of behavior.* New York: Henry Holt & Co.

Mohrmann, C., et.al. (1961) *Trends in European and American linguistics, 1930-1960.* Spectrum, Vol. 1.

Moskowitz, G. (1968) The effects of training teachers in interaction analysis. *Foreign Language Annals, 1,* No. 3, 218-235.

Moulton, W.G. (1961) Linguistics and language teaching in the United States, 1940-1960. *In* Mohrmann, C., and others (Eds), 1961.

Mowrer, O.H. (1960) *Learning theory and behavior.* New York: Wiley.

Mueller, T.H. (1971) The effectiveness of two learning models: the audiolingual habit theory and the cognitive code-learning theory. *In* Pimsleur, P. & Quinn, T. (Eds), 1971.

Nachman, M. & Opochinsky, S. (1958) The effects of different teaching methods: a methodological study. *Journal of Educational Psychology, 49,* No. 5, 245-249.

Newmark, L. & Reibel, D.A. (1968) Necessity and sufficiency in language learning. *IRAL, 6,* No. 2, 145-164.

Nida, E.A. (1957) *Learning a foreign language.* Ann Arbor, Michigan: Friendship Press.

del Olmo, G. (1968) Professional and pragmatic perspectives on the audiolingual approach: Introduction and review. *Foreign Language Annals, 2,*

No. 1, 19–29.

Ornstein, J. & Lado, R. (1967) Research in foreign language teaching methodology. *IRAL, 5*, No. 1, 13-25.

Otto, F. (1969) The teacher in the Pennsylvania project. *Modern Language Journal, 53*, No. 6, 411-420.

Outhwaite, A.D. & Rutherford, A. (1955) Covariance analysis as an alternative to stratification in the control of gradients. *Biometrics*, 11, 431-440.

Pimsleur, P. & Quinn, T. (Eds.) (1971) *The psychology of second language learning.* Cambridge university Press.

Politzer, R.L. (1960) *Teaching French: an indroduction to applied linguistics.* Boston: Ginn.

Politzer, R.L. (1964) The impact of linguistics on language teaching: Past, present and future. *Modern Language Journal, 48*, No. 3, 146-151.

Politzer, R.L. (1967) *An investigation of the order of presentation of foreign language grammar drills in relation to their explanation.* Final report. Project No. 5-1096. Contract No. OE-6-14-025. Washington, D.C.: Office of Education.

Porter, A.C. (1967) *The effects of using fallible variables in the analysis of covariance. Doctoral dissertation, University of Wisconsin. University microfilms No. 67-12, 147, Ann Arbor, Michigan.*

Raths, J. (1967) The appropriate experimental unit. *Educational Leadership, 25*, 263-266.

Reibel, D.A. (1971) Language learning strategies for the adult. *In* Pimsleur, P. & Quinn, T. (Eds), 1971.

Rivers, W. (1964) *The psychologist and the foreign language teacher.* The university of Chicago Press.

Rivers, W. (1968) *Teaching foreign language skills.* Chicago: The university of Chicago Press.

Roddis, M.F. (1968) The contemporary relevance of three early works on language teaching methodology. *IRAL, 6*, No. 4, 333-347.

Saporta, S. (1966) Applied linguistics and generative grammar. *In* Valdman (Ed.), 1966.

Scherer, G.A.C. & Wertheimer, M. (1964) *A psycholinguistic experiment in foreign-language teaching.* New york and London: McGraw-Hill.

Schultz, R.E. (1968) Experimentation relating to formative evaluation. *In* Klausmeier, H.J., et.al., 1968.

Scriven, M. (1967)The methodology of evaluation. *Perpectives of curriculum evaluation.* AERA Monograph Series on Curriculum Evaluation, No. 1, 39-83.

Siegel, L. & Siegel, L.C. (1967) A multivariate paradigm for educational research. *Psychological Bulletin, 68*, No. 5,.306-326.

Siegel, S. (1956) *Nonparametric statistics for the behavioral sciences.*

McGraw-Hill.

Sjöberg, K. & Tropé, B. (1969) The value of external direction and individual discovery in learning situations: The learning of a grammatical rule. *Scandinavian Journal of Educational Research, 13,* 233-240.

Skinner, B.F. (1957) *Verbal behavior.* London: Methuen & Co. Ltd.

Slettengren, A. & Widén, B. (1966) *A modern English grammar,* Stockholm: A.V. Carlssons.

Smith, F. & Miller, G.A. (Eds) (1966) *The genesis of language.* The MIT Press.

Smith, Jr., P.D. & Berger, E. (1968) *An assessment of three foreign language teaching strategies utilizing three language laboratory systems.* Final report. Project No. 5-0683. Grant No. OE-7-48-9013-272. Washington, D.C.: U.S. Office of Education.

Smith, Jr., P.D. & Baranyi, H.A. (1968) *A comparison of the effectiveness of the traditional and audiolingual approaches to foreign language instruction utilizing laboratory equipment.* Final report. Project No. 7-0133. Grant No. OEC-1-7-070133-0445. Washington, D.C.: U.S. Office of Education.

Smith, Jr., P.D. (1969 a) *A comparison study of the effectiveness of the traditional and audiolingual approaches to foreign language instruction utilizing laboratory equipment.* Supplementary report. Project No. 7-0133. Grant No. OEC-1-7-070133-0445. Washington, D.C.: U.S. Office of Education.

Smith, Jr., P.D. (1969 b) The Pennsylvania foreign language research project: Teacher proficiency and class achievement in two modern languages. *Foreign Language Annals, 3,* No. 2, 194-207.

Smith, Jr., P.D. (1970) *The pedagogical unit.* Work paper No. 6/70 from The Center for Curriculum Development, Philadelphia, Pa.

Snedecor, G.W. & Cochran, W.G. (1967) *Statistical methods.* Iowa State University Press. Ames, Iowa.

Spolsky, B. (1966) A psycholinguistic critique of programmed foreign language instruction. *IRAL, 4,* No. 2, 119-129.

Stake, R.E. (1967) The countenance of educational evaluation. *Teachers College Record, 68,* 523-540.

Stanley, J.C. (1970) Controlled experimentation: Why seldom used in evaluation? Educational Testing Service, Princeton, New Jersey: *Proceedings of the 1969 invitational conference on testing problems,* 104-108.

Stephens, J.M. (1967) *The process of schooling.* New York and London: Holt, Rinehart & Winston.

Strasheim, L.A. (1970) A rationale for the individualization and personalization of foreign-language instruction. *The Britannica Review of Foreign Language Education.* Chicago: Encyclopedia Britannica, Inc.

Svensson, A. (1971) *Relative achievement.* Stockholm: Almqvist & Wiksell.

Tatsuoka, M.M. (1969) Experimental methods. *Encyclopedia of Educational Research.* Fourth edition, 1969. McMillan.

Thislethwaite, D.L. (1969) Some ecological effects of entering a field of study. *Journal of Educational Psychology, 60,* No. 4, 284-293.

Titone, R. (1968) *Teaching foreign languages. An historical sketch.* Georgetown University Press, Washington D.C.

Torrey, J.W. (1969) The learning of grammatical patterns. *Journal od Verbal Learning and Verbal Behavior, 8,* 360-368.

Valdman, A. (Ed.) (1966) *Trends in language teaching.* New York: McGraw-Hill.

Valdman, A. (1969) Evaluation of PDE/FL research projects. *In* Smith, Jr., P.D., 1969 (a).

Valdman, A. (1970) Toward a better implementation of the audio-lingual approach. *Modern Language Journal, 54,* No. 5, 309-319.

Valette, R.M. (1968) Testing. *The Britannica Review of Foreign Language Education, Volume I,* Chicago: Encyclopedica Britannica.

Valette, R.M. (1969) The Pennsylvania project, its conclusions and its implications. *Modern Language Journal, 53,* No. 6, 396-404.

Wallen, N. & Travers, R.M.W. (1967) Analysis and investigation of teaching methods. *In* Gage, N.L. (Ed.), 1967.

Wardrop, J.L. (1968) Controlled experimentation in multiclassroom settings. *In* Klausmeier, H.J., et.al., 1968.

Wardhaugh, R. (1969) *Teaching English to speakers of other languages: The state of the art.* Washington, D.C.: Center for Applied Linguistics.

Werdelin, I. (1968) The value of external direction and individual discovery in learning situations. II: The learning of a foreign alphabet. *Scandinavian Journal of Psychology, 9,* 248-251.

Wiley, D.E. & Bock, R.D. (1967) Quasi-experimentation in educational settings: Comment. *The School Review, 75,* No. 4. 353-366.

Wiley, D.E. (1969) A methodological review of the Pennsylvania foreign language research project. *Foreign Language Annals, 3,* No. 2, 208-213.

Wiley, D.E. (1970) Design and analysis in evaluation studies. *In* Wittrock, M.C. & Wiley, D.E. (Eds), 1970.

Wittrock, M.C. (1966) *The experiment in research on evaluation of instruction.* ERIC: ED 012 107.

Wittrock, M.C. & Wiley, D.E. (Eds) (1970) *The Evaluation of Instruction.* New York: Holt, Rinehart & Winston.

Wohl, M. (1967) *Classroom experiment to measure the relative efficiency of two different linguistic models in their application to the teaching of English as a foreign language.* Final report BR 6-8200, Contract No. OEC-3-6-068200-0668 PA-24, The university of Michigan, Ann Arbor, Michigan. ED 013449.

Wolfe, D.L. (1967) Some theoretical aspects of language learning and language teaching. *Language Learning, 42*, Nos 3 & 4, 173-188.

Wragg, E.C. (1970) Interaction analysis in the foreign language classroom. *Modern Language Journal, 54*, No.2, 116-120.