CHAPTER VII

PROBLEMS

Overview

This chapter will outline the problems that were studied in the present investigation. Before we do that, we will present a general summary of the extensive review of vocabulary research. This will show where the present study fits in the framework of the past and present vocabulary research paradigm.

Summary

It seems obvious that lexicon and word are receiving increasing attention in several disciplines. In linguistics, there has been a growing doubt about the psychological realism of syntactical transformations. This has led to a growing interest in devising close-to-surface lexical-interpretative theories of language and text (e.g., Halle, Bresnan & Miller, 1978; Melchuk & Zolkovsky, 1974). This view has tended to dispel some of the awe and sense of mystery concerning syntactical development. It seems likely that some of the complication in language development will be shifted to the lexical component (e.g., Maratsos, 1978). In this vein, Bolinger (1976) has argued that languages have a great amount of "prefabricated" elements and thus, for instance, idiomaticity is "a vastly more pervasive phenomenon than we ever imagined, and vastly harder to separate from the pure freedom of syntax, if indeed any such fiery zone as pure syntax exists" (p.3). Similarly, Fillmore (1979) suggests that a large portion of people's ability to get along in a language consists in the mastery of formulaic expressions. Wilkins (1972) crystallized this way of thinking by noting that while it is difficult to say

much with a limited command of syntax, it is practically impossible to say anything without adequate vocabulary (and other fixed, formulaic expressions, we might add).

Research done in Finland showed that vocabulary knowledge had a decisive influence on school grades in foreign languages and that intensive vocabulary review was possible in terms of student cooperation and led to clear improvement in grades.

It has also been shown consistently that vocabulary is a very good predictor of overall verbal ability measures. Recent experiments with intensive vocabulary training in the primary school have also indicated that it can result in increased comprehension of text. On the other hand, the number of words in written school language in the case of first language is so high that direct teaching of vocabulary seems unable to cope with the learning task of such magnitude. Thus, word analysis skills and the ability to derive and learn word meanings form context are important prerequisites for the acquisition of an adequate vocabulary. There is some indication, however, that the quality of contexts in textbooks is not nearly as good as it might be.

In foreign language teaching, the choice of vocabulary has occupied a prominent position for a long time. Several methods have been used as criteria for word selection, including word frequency, availability (disponibilite), usefulness, familiarity, and difficulty (e.g., Bongers, 1947; Carpay, 1975; Fries & Traver, 1940; Richards, 1971a; Sciarone, 1979). The determination of the minimum number of vocabulary needed to be able to function in a foreign language has also been a central concern in pedagogical lexicography. It seems that about 5,000 words are needed in order to be able to read

literary text with easy comprehension of the content.

Finally, it appears that while there has been considerable interest in making frequency lists and estimating how many words are needed to comprehend "normal" text, there have been relatively few studies that have attempted to estimate how many words students have acquired. This applies both to first language and foreign language teaching. One of the reasons for this neglect is obviously the amount of work that a reliable estimation of vocabulary sizes requires. Yet, as several researchers have recently pointed out, accurate estimation of vocabulary size is one of the most important preconditions for progress in research on vocabulary learning in general.

In recent years there has been a revival of interest in studying various aspects and issues related to vocabulary learning. Anderson and Freebody (1981) have summarized what is known about role of vocabulary knowledge in reading comprehension. Their discussion is related to first language acquisition but its relevance for the present study is still obvious. In the introduction to the review Anderson and Freebody (1981) state that

> An assessment of the number of meanings a reader knows enables a remarkably accurate prediction of this individual's ability to comprehend discourse. Why this is true is poorly understood. Determining why is important because what should be done to build vocabulary knowledge depends on why it relates strongly to reading. The deeper reasons why word knowledge correlates with comprehension cannot be determined satisfactorily without improved methods of estimating the size of people's vocabularies. Improved assessment

> methods hinge, in turn, on thoughtful answers to such questions as what is word, what does it mean to know the meaning of a word, and what is the most efficident way of estimating vocabulary size from an individual's performance on a sample of words. (p. 77)

Similar views have been expressed by other researchers as well. In a recent review of vocabulary learning in mother tongue, Jenkins and Dixon (1983) note that there are several problems that are in need of investigation. They suggest that it would be important to know if large amounts of vocabulary can be learned in an economical time frame; if direct teaching of skills of analyzing the morphological structure of words will enhance vocabulary growth; if effective ways can be developed to teach children not only to derive meanings from contexts but also to remember them. They also point out that more refined answers are needed to the perennial question about the relationship between knowledge of word meanings and listening and reading comprehension. For the present study, the most pertinent observation by Jenkins and Dixon (1983) is, however, the following:

> In addition, fundamental questions remain pertaining to the measurement of vocabulary knowledge: how many words are actually learned (and how many should be taught), what kinds of learning are involved, and what kinds of tests are indicative of these learnings? (p. 22)

The same point is made somewhat differently in another study by the same research team (Jenkins, Stein & Wysocki, 1983):

> A major unresolved issue that will continue to haunt researchers until they achieve better measurement procedures is that of vocabulary size. Because size estimates vary so greatly, it is difficult

> if not impossible to obtain reasonable estimates of the relative contribution of different vocabulary experiences in the development of word knowledge. Because we do not know how many words individuals know, we are seriously limited in accounting for changes in these totals. (p. 26)

It is in the spirit of such statements that the present investigation was initiated. The fact that the study is related to taught vocabulary implies that it belongs to the domain of program evaluation. It also means that it is, almost by definition, related to the domain-referenced mode of testing.

Research Problems

The main purpose of this study is to estimate the size of students' active and passive vocabulary in English after seven years of English in the Finnish comprehensive school. The students started learning English at the age of nine and had some 450 clock hours of classroom instruction, usually 2-3 lessons a week. The estimation is to be done so that the results apply to the whole student population as well as to the entire universe of taught vocabulary. Thus, a high degree of generalizability of the results is a central research objective as well as the possibility of assessing the quality of the data and the dependability of conclusions.

In addition to the main research question addressing the estimation of the overall size of active and passive vocabulary, the study is designed to provide answers to some more specific quantitative questions:

(1) How many words are known passively and actively of the ones taught during different stages in the seven-year course (lower stage, upper

stage, and upper stage extra vocabulary)?

(2) What is the relationship between the taught and learned vocabulary?

(3) What is the relative contribution of students vs. items to observed variation in scores of vocabulary items?

The study also has some objectives related to methodological questions. It seeks to draw on recent advances made in test and sampling theory (multiple matrix sampling, generalizability theory) and in test construction (criterion-referenced measurement). The hope is that this exercise will increase our knowledge about their applicability in general but especially in L2 research. Specifically, we want to get at least tentative answers to questions like the following:

(4) How does multiple matrix sampling work in vocabulary research?

(5) How do the variance components estimated with the generalized symmetrical sums approach, which allows an unbalanced multiple matrix sampling design, compare with the ones computed with standard analysis, which set more constraints on design?

(6) What is the optimal trade-off between the number of word items and the number of students in terms of measurement accuracy?

(7) To what extent do students' word-analysis skills and their ability to utilize context in inferring word meanings affect the estimates of vocabulary size?

Obviously, several problems had to be solved in designing the study, so that there was a high likelihood that sufficiently reliable answers would be obtained to the questions. It is to the design and methodology adopted in the study that we now turn.

CHAPTER VIII

DESIGN AND METHODOLOGY

Introduction

The purpose of the study was to get generalizable estimates of the entire student population's average passive and active vocabulary sizes, i.e., to estimate average universe scores. This means that several problems had to be addressed in the planning of the study. These had to do with (1) the sampling of students, (2) the sampling of the items, and (3) development of a valid, reliable and practicable way of measuring passive and active vocabulary. Problems and issues related to all three areas will be briefly discussed and the adopted solutions described and justified. Before the design of the study is described, it is appropriate to give a brief account of what considerations led to its adoption.

Cronbach and his associates (Cronbach, Gleser, Nanda & Rajaratram, 1972) developed the theoretical ground-work for generalizability theory. Since their pioneering work, a lot of work has already been carried out to develop the theory further and apply it in various circumstances. In Finland, Konttinen's contribution (Konttinen, 1980) has shown how generalizability theory can be applied to Finnish conditions. In his capacity as Head of the Department of Research Methdology at the national Institute for Educational Research, since the mid 1970's he carried out intensive work on new ideas in test theory and paved the way for the first National Assessment of Teaching in the Comprehensive School. His contribution to the design of the assessment, including this study, was of decisive importance. It is to this work that we now turn. Design problems will be discussed in some detail, partly

because they were major concerns in the planning of this investigation and partly in the hope that this study might contribute to a higher awareness of design problems and possibilities in L2 research.

Preliminary Work on Design Problems

Konttinen (1980) analyzed several empirical evaluation data collected by various researchers at the Institute for Educational Research in order to determine the size of the standard error of measurement in various study designs. Using mathematics data, he showed that about 80 items were needed in order to bring the generic true score within the confidence interval of plus/minus .10, alpha coefficient to .89, and the generalizability coefficient to .87. He also showed that with 24 items, it was not possible to bring the generic standard error down to a satisfactory level even with an infinite number of subjects. With 32 items and 256 subjects, the confidence interval with 95% level of confidence was found to be .15, and with 64 items and 256 subjects the same range was .10 (i.e., plus/minus .05 round the observed mean). Konttinen (1980) further showed that the generic standard error can be brought surprisingly low with very few subjects, provided that they are sampled using a simple random sampling method and provided that the number of items is large. If a random sample of 16 subjects were presented 256 items, the confidence interval of the mean would be .15, which level can also be obtained with 32 items presented to 256 subjects. If the number of subjects is doubled (up to 32) and the number of items is raised to 512, the confidence interval would be .10, which could not be achieved using 32 items however many subjects were tested.

Using data related to the measurement of students' knowledge of Swedish grammar, Konttinen (1980) also showed that the standard error of the p-values

of items decreased faster when the number of schools was increased than when the number of students was raised. In order to be able to estimate the difficulty level of items with a confidence interval of plus/minus .05 (and 95% level of confidence), it was necessary to sample about 30 schools and 30 students from each school. This means a sample of some 1,000 students.

Thus, when earlier empirical evaluation data collected in Finland were reanalyzed, it was found that no generalizable results can be obtained with few items and few schools. Provided that the sample contains a sufficient number of schools, it is not necessary to sample many students from each school. Konttinen (1980) concluded that several and partly contradictory requirements must be fulfilled if the same instrument is used for several purposes, for instance, to evaluate the performance of individual students, to estimate the means of different subtests, and the means of the whole test. The first requirement presupposes several tests to cover the whole area of the curriculum, the second one entails several items, and the third one only a limited amount of items. Konttinen (1980) states that:

> With only reasonable constraints on measurement error, this easily leads to a situation in which 30 schools are needed,and some 30 subtests with about 20 items in each. If school means are also to be estimated, this means that some 30 students from each school must be sampled. Altogether, this means that about 1,000 students should be sampled and some 600 items constructed. (p. 105)

Generalizability, not only to the student population, but also concerning curricular domains and the universe of items has proved to be the weakest point in the earlier evaluation research carried out at the Institute and

elsewhere. Konttinen (1980) conducted a number of studies on generalizability using earlier evaluation data. This was done by studying the relative size of variance components in different types of design.

Konttinen (1980) showed further that using a design of the p(s) x s x i(t) x t type (i.e., students are nested within schools and items are nested within subtests, such that each student belongs only to one school/class and each subtest has different items), the residual variance was the highest, varying from 55 to 85% in different school subjects. The second largest source of variation in student performance was due to subtests and items, with the latter being more important and of the order of 10%. Thus there were easier and harder domains in the subject studied but more important seemed to be how the items were formulated. The variance components related to schools/classes and students was found to be of the same order of magnitude as that of subtests/domains and items: the school/class usually about 2% and the students some 6%. On further analysis, it was found that the between-schools variance was usually around 25% of the between-students variance. In terms of correlation, this means that if we knew which school a student came from, we would be able to predict his or her performance as well as with a variable that correlated .50 (the square root of .25) with achievement. Thus, schools and classes vary considerably with regard to performance level.

As far as interaction components are concerned, the most salient finding was that the st-component (school-domain component) was very small. This indicates that all schools appeared to follow the national curriculum fairly closely. There were, however, differences in emphasis on items, as shown by the fact that the si and sit components were on the order of 2% (usually statistically significant). The interaction of students with domains and

items was also usually of the 2% order and usually statistically significant. Some students had learned certain tasks better than others had.

Generalizability Problems Related to Designs

Usually student sampling is not a problem and there are well-established methods developed within sampling theory (e.g., Kish, 1964). Thus, it is relatively easy to generalize to the entire population of subjects. By contrast, generalizing to the content domain has proved a difficult problem in research. The construction of a large amount of items is laborious but there is no way to avoid that if there is interest in generalizing the results to the entire domain (universe). This is typically the case when the attainment of curricula is evaluated. By means of generalizability studies it is, however, possible to estimate what kind of measures are needed for each particular research and conclusion purpose, and how many subjects, items, etc are needed in order to obtain the level of accuracy that is desired. This possibility facilitates considerably the design of studies and encourages careful planning.

Even if it were possible to produce a large number of items, their presentation to students creates problems. Recent developments in generalizability theory and sampling theory have suggested some feasible solutions. It is possible to apply a hierarchic or partially hierarchic subject x item design, which means that items are divided into several test forms and each student gets only one form. Thus the basic p x i (person, item) design becomes a p x i(f) x f design. This means that subjects are crossed with items within forms and with forms.

Such a partly nested design is usually called a matrix sampling design (e.g., Shoemaker, 1973). The problem with the early systems of estimating generalizability by computing variance components (e.g., Lord & Novick, 1968; Shoemaker, 1973) was that balanced and complete designs were required. No missing data were allowed and each subtest had to have an equal number of items. However, Sirotnik and Wellington (1977) have presented an integrated theory of matrix sampling (which they prefer to call "incidence sampling"), which makes the original designs special cases and allows arbitrary incidence matrices.

Drawing on the earlier work of Hooke (1956), Sirotnik and Wellington (1977) developed a method of defining "generalized symmetric sums" (gss's) and "generalized symmetric means" (gsm's), which can be used for estimating variance components (including error variance components) as in simpler designs. The usefulness of generalized symmetric means depends on two factors: (1) All moments (mean, variance, skewness, and kurtocity) can be expressed as linear combinations of gsm's. (2) The sample gsm is an unbiased estimate of the population gsm for all kinds of incidence matrices. Thus, if we know the symmetrical sums of a sample, we can obtain unbiased estimates, for instance, for the mean and for its standard error (based on the use of variance, i.e., the second moment). However, the problem with this gss/gsm method was that the computation was very time consuming and exceedingly complex. In their article published in 1977, Sirotnik and Wellington suggested that it might be possible to computerize the process if the algorithms for computing gsm's could be programed efficiently. Work on this problem was started at the Institute immediately after the publication of the article,

and Tormakangas (in press) developed a computer program which can compute gsm's with reasonable speed. Large data, like in the present study, still require large amounts of computer time. The analysis of the data was possible only because the Institute has unrestricted access to the University of Jyvaskyla computer time (i.e., "free" computer time) and the computer is not yet seriously overloaded.

Because of its obvious advantages, the gss/gsm approach to data analysis was adopted in the present study. Different students answered partly different items, but not according to any earlier balanced designs that would have required that "each item is paired with every other item the same number of times, every item pair being administered to the same number of examinees, and every examinee responding to the same number of items" (Sirotnik & Wellington, 1977, p.343). The student x item matrix in the present study is not balanced and missing data are allowed.

Considerations in Defining the Design of the Study

Shoemaker (1973) has noted that the methodology used in the evaluation of individual differences is neither appropriate nor efficient for the assessment of group performance. This is a very important point in terms of its implications for program evaluation. In educational planning and in national assessments, we are typically interested in how well groups of students or entire populations of students perform in particular types of programs.

In Finland, as well as in several other countries, there seems to have been a similar trend in evaluation. In the first stage, standardized tests were used to measure educational achievement. In the second stage, there was

an attempt to make the tests more congruent with instruction. The work done in connection with the First National Assessment was an attempt to move to a third stage of evaluation in Finland, which could be characterized by more careful domain and item universe definition and a more detailed specification of item generation rules. The first step was work on the common core curricula and the second step was work on implementing a new type of design and collecting data in accordance with it.

The finding by Konttinen (1980) that the variance components of items and sub-domains was normally the second largest (after the residue component) and that the components related to schools or to the schools x items or schools x sub-domains were clearly smaller, suggested that getting an adequate school and student sample was not the biggest problem. When the purpose of the present study was to get a good estimate of students' active and passive vocabulary sizes and to compare the two, the greatest source of uncertainty seemed to be items: different items might give varying results. Thus, getting a good sample of items appeared to deserve high priority. This led to a conclusion that matrix sampling was the most appropriate design.

Shoemaker (1973) and Wolf (1979) and several other scholars have discussed the merits of multiple matrix sampling. First, the standard error in estimating e.g., group mean test scores is reduced. Second, it makes it possible to cover large item universes. Third, it is economical since it allows a maximum amount of information with a minimum amount of testing time. This is particularly important in large-scale assessments of broad curricular areas. The obvious effectiveness of the testing procedure will also help increase the willingness of school principals and teachers to respond positively to invitations to take part in evaluation studies.

Matrix sampling was perceived to involve a few problems, however. The first had to do with the equating of test forms. The possibility of equating different test forms was regarded as necessary since, at a later stage, the achievement scores would be related to a number of other variables collected with questionnaires. Scores on different test forms were equated with the one-parameter logistic latent trait model (i.e., the Rasch model; see e.g., Lord, 1980). Equating worked quite well. This was probably due to the fact that equating was done horizontally, not vertically (across several grade levels), which does create problems (e.g., Slinde & Linn, 1978).

Another problem was more directly connected with the design of the present study. In spite of all earlier work done by Konttinen, which has been described in the above, the design could not be based on solid grounds. There simply was not available any earlier work that had used a sufficient number of vocabulary items, which would have made it possible to operate with relatively exact figures.

Summary of Sampling Rationale

Earlier work on sampling theory and generalizability estimation (e.g., Cronbach, Gleser, Nanda & Rajaratram, 1972; Konttinen, 1980; Shoemaker, 1973; Sirotnik & Wellington, 1977) had shown that matrix or incidence sampling has several advantages. Tormakangas' work on estimating standard errors with the gss/gsm approach to variance components analysis appeared promising and has subsequently led to a computer algorithm (Tormakangas, in press), which cuts down the needed computing time considerably. Thus, multiple matrix sampling was chosen as the sampling method.

Work on earlier empirical evaluation data by Konttinen (1980) had provided information about the relative size of variance components of schools/classes, students, sub-domains, and items, and also about relative pay-offs for meas-urement accuracy of increasing the number of schools, students, or items. His work suggested that there would be relatively few problems with getting an adequate school and student sample. On the other hand, the work by Lord and Novick (1968) and Konttinen (1980) indicated that the most important factor for generalizability was an adequate item sample. Getting a good item sample was accorded a high priority in the study.

One of the tasks of the study was to improve our knowledge of design problems in vocabulary research. For this reason, it was decided to estimate the active vocabulary size in two ways. The "intensive sample" was designed to have fewer items presented to relatively many students, whereas the "extensive sample" was to have more items but fewer students. It was hoped that by following rigorous principles in the design of the study, it would be possible, a posteriori, to estimate how well these two sample designs work. Subsequent investigations could then be planned with a higher level of sophistication.

In spite of all preliminary work, the design of the study had to be settled by juggling different constraints. As mentioned earlier, equating was considered important and was thought to require a common set of anchor items. Guaranteeing generalizability seemed to presuppose a good word sample. On the other hand, the number of items had partly to be determined by considering how many items students could answer in one class period.

Sampling of Students

Population

For financial and practical reasons, it is usually necessary in evaluation research to measure only a sample of the total group about which we wish to draw conclusions that apply to the whole populatuion. The unit of sampling needs to be determined on the basis of whether the main focus of the study is to estimate the performance of individual students, whole classes, schools, regions, or entire school systems. Another important task in planning the sampling design is to define the population of the study.

The desired target population of the study was defined as "all students in the final grade of the comprehensive school". The excluded population consisted of students in the Swedish-speaking schools and special schools. The students in the special schools were excluded because they did not always follow the normal curriculum. The Swedish-speaking schools were excluded mainly for practical reasons. Preparing tests that would have been appropriate for another language group was beyond the resources of the project. Also, most Swedish-speaking students have a five-year course in English while most Finnish-speaking students have a seven-year course. Thus the final target population of the study was defined as "all Finnish-speaking students in the final grade of "normal" comprehensive school classes".

Sampling Method

It was decided that the size of the sample should be as small as possible without jeopardizing the reliability (generalizability) of the results. Some earlier work in Finland had indicated (e.g., Konttinen, 1980) that a good

sample would probably not need to include more than 60 schools and about 25 students from each school.

The most direct way of sampling students would be to draw a simple random sample of students. A list of all final year students would be needed to be able to do that. This is often difficult to obtain. Simple random sampling is often not practicable for other reasons as well. School systems divide students into schools and classes. It is costly and administratively difficult to take a few students from different classes and arrange testing sessions for them. Therefore, it was decided to use a multistage sampling design. It was further decided to use the school as the primary sampling unit rather than, for instance, province or class.

The sampling method was a two-stage stratified cluster sample. The primary sampling unit was the school and the secondary sampling unit was the class. In the case of the vocabulary test (but not the questionnaires), there was in fact a third stage of sampling. Different test forms were presented to different students. This was done in a totally random fashion, and does not create any new problems for the generalizability of the results.

The error constraints for parameter estimates were set such that the standard error of the country means must be within .03 and .06. It was further specified that a difference of one half of standard deviation should be detected with 95% level of confidence in the standard error of means for different sets (streams).

In earlier Finnish studies, it had been found that the standard deviation of the schools was about 35% of the standard deviation of students (about 7 vs. 17 in a hundred-item test). Konttinen's analyses with earlier empirical assessment data, conducted during the period when the study was

being planned (Konttinen, 1980), showed that about 30 schools with some 30 students from each school were needed with some 40-60 items to achieve a confidence interval of .05 -.075 with means and the p-values on individual items.

These considerations led to the decision that although 30 schools seemed sufficient to satisfy the error constraints, it would be desirable to sample approximately 40 schools. This was done in order make sure that, in spite of the use of multiple matrix sampling (cf. section on item sampling below), there would be a sufficient number of students answering each item.

The final sampling frame consisted of four strata. The size of the school and the degree of urbanization of the school community were used as the two bases of stratification. Small schools were defined to have up to 349 students and large school 350 or more students. In the other stratum, the schools were divided also into two groups: those located in rural areas and those in urban areas.

The first stage of sampling, the selection of schools, was performed by means of a simple random sampling by stratum (using tables of random numbers). This was based on a national register which contained several kinds of statistical data on schools. In the second stage, one class representing each of the three sets in English was selected. This was done in the following way: local educational officers in those communities whose schools were drawn into the sample were asked to provide information on the number of classes in each set and the number of students in each class. A random selection of one class representing each set was then made by the author. The distribution of schools in the four strata is shown in Table 12.

Table 12

Designed and Executed Sample of Schools

Degree of urbanization	Size of school: Small (349 or less) Designed	Small (349 or less) Executed	Large (350 or more) Designed	Large (350 or more) Executed	Total Designed	Total Executed
Urban	5	4	11	10	16	14
Rural	14	13	12	12	26	25
Total	19	17	23	22	42	39

The achieved sample of schools varies to some extent from the designed sample. This is mainly due to the fact some schools failed to send in the data and there was no time to draw new substitute schools, because data collection took place late in the spring term.

Item Sampling

Introduction

In most studies evaluating student achievements, strict sampling have been applied only to the sampling of students. Several scholars (e.g., Cronbach, Gleser, Nanda & Rajaratram, 1972; Lord & Novick, 1968; Popham, 1978; Wolf, 1979) have pointed out that in evaluation research the sampling of content is equally, if not more, important. There has been a movement in measurement called "criterion-referenced" or "domain-referenced" measurement, which has tried to make advance in this respect. According to Popham (Popham, 1978, 1980), criterion-referenced measurement provides an exact description of a person's performance in an entire domain and not only on the

presented items. When the present study was being planned, the author reviewed current literature on criterion-referenced measurement and wrote his Master's thesis in Education on the topic (Takala, 1980).

It is in the specification of the content domain that the greatest challenge and the greatest merit of criterion-referenced measurement lies. The content universe has to be defined with similar rigor as the student population. In traditional norm-referenced tests the content limits are only partially specified. Short instructional and behavioral objectives are used as the basis of item generation. As Bormuth (1970) and Anderson (1972) and several others after them have shown, there is so much room left for interpretation that the items may reflect the characteristics of the test constructor more than those of the instructional program.

As far the present study is concerned, the "test specification" method, as advocated by Popham (1978, 1980) seemed most appropriate. Test specification, which defines stimulus and response characteristics, item generation rules, scoring criteria, etc., is extremely difficult in areas like reading and listening comprehension and speaking and writing. Some clear progress has been made in the case of writing (e.g., Baker, 1982; Takala, 1982e; Vahapassi, 1982). It is easier when the domain is more limited and when the elements can be identified with some rigor.

Once the rules for domain specification and for item generation have been worked out, it is necessary to consider specific items. Unlike in norm-referenced measurement, it is necessary in criterion-referenced measurement to know what the universe of items is that represents the defined domain content. This universe can be finite or infinite. As Millman (1973) has

pointed out, it is not necessary that the population of items actually exists. What is necessary, however, is that the domain is so well described that a high agreement can be reached about what items are and what are not members of the item universe. If generalizability to the defined content domain is to be achieved, items must be randomly sampled from the content domain following equally strict criteria as are applied in the sampling of subjects.

When the present study was being planned, it seemed that measuring vocabulary, as well as measuring mastery of grammatical structures, lent itself well to an attempt to apply principles of criterion-referenced measurement in foreign language research. There are several reasons why studying the size of students' active and passive vocabulary using new ideas developed in criterion-referenced measurement, in sampling theory (especially multiple-matrix sampling) and measurement theory (generalizability theory and latent trait theory) seemed promising.

It seemed possible, if laborious, to define the domain and even identify and count the items in the domain. This was possible because of the specific nature of English teaching in Finland. First, there is a national curriculum, and textbooks are written on the basis of the curriculum. They are checked by the National Board before they are approved for school use. There were only two major textbook series used in English teaching at the time of the study. Thus it was possible to list the words that were likely to have been taught to students. Like in the earlier study in Sweden by von Mentzer (1968), it seemed relatively safe to assume that the textbook was, in fact, practically the only source of input in English classes.

Second, English and Finnish are not related languages. This applies to the structure as well as the lexicon. Thus Finnish-speakers do not benefit from the existence of cognates, except to a very small extent. Going back to roughly a hundred years, when Finnish emerged as an official language equal to Swedish, there has been an attempt to "purify" Finnish from "alien" influence, especially that of Swedish. When the present writer went to school, "Sveticisms" were rigorously expurgated from student compositions. Thus, there has been an attempt to create native words to code new concepts in science, technology and other fields. English has come to have any appreciable impact on Finnish life and language only after World War II. While English can be regularly heard on TV (with subtexts in Finnish; no or little dubbing is used) and pop culture has a strong English dominance, this would hardly seriously distort the vocabulary estimates. Also, travel to language schools in Britain or staying a year abroad usually happen after students have left the comprehensive school.

Third, the problem of polysemy, often referred to in literature on vocabulary teaching and learning, is less acute in foreign language teaching. Students are typically taught the core meanings, or "central meaning normal to the most frequent and nonspecial set of contexts" as Pike (1982) puts it. They are seldom taught the "marginal meanings" which occur when the central meaning "is modified by other words in the context". Foreign language textbooks in Finland have always, with the exception of a short period in the heyday of audiolingualism, had bilingual wordlists, in which these central meanings are explained in Finnish. It is the knowledge of these meanings, and not all possible context-determined meanings, that is evaluated.

In sum, it appeared possible to define and even list what lexical items had been taught and what meanings students should be familiar with. This was done so that the vocabulary lists of the two textbook series (14 books in all) were transferred onto a computer tape. A number of codes were assigned to each lexical item, e.g., the year when the word was first taught, the type of text it belonged to (core text vs. extra/optional text), set (stream), and part of speech. This was done mechanically by the clerical staff. In the second stage, the author screened the file and removed those items that were not considered to require a separate entry in the lexicon. This applied especially to adverbs derived in the regular way from adjectives as well as regular inflected forms of nouns and verbs(cf. Aronoff, 1976). Also the forms of irregular verbs were excluded, although learning them does constitute an extra memory burden. Excluding them was based on the tradition in foreign language teaching in Finland. Irregular verb forms are reviewed regularly and knowing the forms of irregular verbs is considered an integral part of knowing the basic form. On the other hand, all irregular adverbs, all compound words entered in the word lists, all proper names (mainly names of countries and nationalities), and all phrases and idioms were included in the word population. So were all so-called structural words (e.g., prepositions and conjunctions) with the exception of articles. This edited list constituted the vocabulary population from which stratified word samples were drawn, as shown in Tables 13 and 14.

Word population and word samples. As in the case of student sampling, stratification was used in item sampling to improve its efficiency. Stratification was used, first, in order to reduce the standard error in the estimation of the average proportion correct values for items (analogously with

student sampling; e.g., Kish, 1964) and second, to guarantee that the item sample covered the word universe adequately. The following strata were used:

- textbook (textbook 1 and textbook 2)
- vocabulary taught in different sets /streams (vocabulary taught to sets A & B, vocabulary taught to set C)
- period when vocabulary was taught (vocabulary taught during the lower stage - grades 3 through 6, and vocabulary taught during the upper stage - grades 7 through 9 , either as core material or as extra material).

Students who received different test forms can also be regarded as random samples from the whole student sample. This is due to the fact that test forms were randomly rotated in each class.

The number of words and the number items in different strata in Textbook 1 in sets A, B and C are shown in Table 13.

Table 13

Word Population and Word Samples by Strata, Textbook 1

Vocabulary Stratum	Textbook 1: Sets A & B					
	Population N	%	Active items	Passive items	Total N	%
Lower stage vocabulary	1,011	40.5	117	24	141	40.5
Upper stage/ core vocabulary	1,164	46.6	143	22	165	47.4
Upper stage/ extra vocabulary	323	12.9	42	-	42	12.1
Total	2,498	100.0	302	46	348	100.0

Table(13 cont.)

	Textbook 1: Set C					
	Population N	%	Active items	Passive items	Total N	%
Lower stage vocabulary	1,011	68.8	156	22	178	68.0
Upper stage/ core vocabulary	405	27.6	62	12	74	28.2
Upper stage/ extra vocabulary	54	3.6	10	-	10	3.8
Total	1,470	100.0	228	34	262	100.0

The number of vocabulary population and samples in different strata in Textbook 2 are shown in Table 14.

Table 14

Word Population and Word Samples by Strata, Textbook 2

Vocabulary stratum	Textbook 2: Sets A & B					
	Population N	%	Active items	Passive items	Total N	%
Lower stage vocabulary	812	28.5	72	6	78	25.4
Upper stage/ core vocabulary	1,690	59.2	152	37	189	61.6
Upper stage/ extra vocabulary	352	12.3	36	4	40	13.0
Total	2,854	100.0	260	47	307	100.0

table continues

Table 14 (cont.)

	Textbook 2: Set C					
	Population		Active items	Passive items	Total	
	N	%			N	%
Lower stage vocabulary	812	34.7	66	12	78	36.1
Upper stage/ core vocabulary	1,078	46.1	84	24	108	50.0
Upper stage/ extra vocabulary	450	19.2	6	24	30	13.9
Total	2,340	100.0	156	60	216	100.0

The following two tables show the distribution of items among the "intensive" vs. "extensive" samples.

Table 15

Distribution of Active and Passive Items by Type of Sample and Vocabulary Stratum for Textbook 1 (Figures without Parentheses Refer to Sets A & B and Those within Parentheses to Set C)

Vocabulary stratum	Intensive sample		Extensive sample	
	Active	Passive	Active	Passive
Lower stage vocabulary	59 (100)	0 (22)	58 (56)	24 (0)
Upper stage/ core vocabulary	69 (39)	0 (12)	74 (23)	22 (0)
Upper stage/ extra vocabulary	20 (10)	0 (0)	22 (0)	0 (0)
Total	148 (149)	0 (34)	154 (79)	46 (0)

Table 16

Distribution of Active and Passive Items by Type of Student Sample and Vocabulary Stratum for Textbook 2 (Figures without Parentheses Refer to Sets A & B and Those within Parentheses to Set C)

Vocabulary	Intensive sample		Extensive sample	
stratum	Active	Passive	Active	Passive
Lower stage vocabulary	24 (30)	0 (12)	48 (36)	6 (0)
Upper stage/ core vocabulary	54 (48)	0 (0)	98 (36)	37 (24)
Upper stage/ extra vocabulary	12 (6)	0 (0)	24 (0)	4 (24)
Total	90 (84)	0 (12)	170 (72)	47 (48)

The system of sets (streams) was such that at the end of the lower stage students and their parents could choose from among three sets: Set A was meant for the students who had shown the greatest ability for English and thus could be taught according to the most demanding syllabus, Set B was meant for those students who had made average progress, whereas Set C was designed to cater for the needs of the slow learners. Sets A and B were taught using the same textbook, whereas Set C had a separate, less demanding textbook. Setting was not, however, only a device designed to make it possible to tailor-make instruction to students' needs and abilities. Setting also carried serious implications for further studies, since only the choice of Sets A and B would make it possible for students to be eligible for all kinds of post-compulsory schools. Set C led to a partial educational blind alley.

Choice of Test Format

It was recognized that the choice of the test format was closely dependent on the research problem. Since the main research task was to estimate the size of students' active and passive vocabulary in English, the test types should provide as direct and valid information about such knowledge as possible. Besides such a requirement of construct validity, it was necessary for the test format to provide reliable scores. In addition to these typical requirements, some practical considerations had to be taken into account. The test format should be sufficiently familiar to students. It should take a minimum amount of student time. Of less importance was the time that scoring might take. It is the author's firm belief that researchers should not ask students and teachers to sacrifice any more time and effort to provide data than is absolutely necessary. They should be asked for information (data) that only they can provide. The researchers should, on their part, undertake to do everything that they can do themselves without shifting the work on the shoulders of students and teachers.

On the basis of the above-mentioned principles, it was decided that the best pay-off between validity, reliability, and practicality was shown by test types which ask students to write foreign language or native language equivalents to written decontextualized stimulus words. The following considerations were used in what amounted to an elimination procedure in the selection of the test format:

1) The multiple choice format was not acceptable. Some ten years' experience with multiple choice test construction had led to a view that it is extremely difficult to prepare good multiple choice items. As Anderson and

Freebody (1981) note, the way distractors are formulated (e.g., broad vs. fine distinctions in meaning) largely determines whether the students are able to find the correct answer or not. The multiple choice format has an additional vexing characteristic in the case of foreign language teaching. When students' L2 vocabulary is limited, it becomes extremely difficult to find sensible distractors in L2. Same words must be repeated several times and this may lead to inferential learning during the testing. Also, the difficulty might not be connected with the stimulus word but with the distractors. The use of native language equivalents would largely remove that problem, but it would not remove the objection raised by Anderson and Freebody. An additional objection to the multiple choice format was that it does not reflect how language is used in realistic situations.

2) The check list format, which Anderson and Freebody (1981) prefer to call the "yes-no format", was also considered but not adopted. The work by Oskarson (1978) on self-assessment in foreign language learning , which uses the yes-no format, was known to the author. It was considered interesting and was tried on a small scale in testing students' knowledge of some language functions (e.g., asking, telling the way, apologizing). It was not considered safe to use the format with the vocabulary section, since Oskarson did not provide any clear evidence of its reliability and validity. The work by Anderson and Freebody (1981) was not available at the time when the study was planned and carried out. These authors have shown that the yes-no format is, indeed, a very promising method for measuring vocabulary knowledge in general and has several advantages when the aim is to try to estimate the absolute vocabulary size. For an early study on the self-estimation of vocabulary size, see also Whipple (1908).

3) Matching where the students pair off words with other words (synonyms, antonyms, definitions, etc) was not chosen mainly for the same reason as the multiple choice format was rejected. The list from which pairs have to be selected gives undue hints to the students. It seemed to the author that matching would have been acceptable, provided that the list from which pairs were to be chosen contained all the words covered in the course. This would, however, have been totally unrealistic. If a smaller list is provided, the test constructor risks making the test reflect more his or her own intuitions and predilections than instruction (cf. Anderson, 1972; Bormuth, 1970). Another objection to matching was that such a task is most unlikely in real-life language use situations.

4) This left the constructed answer format, in which students are asked to produce an answer on their own. Anderson and Freebody (1981) note that this format has two problems: scoring and response bias. The measure is confounded with spelling ability, neatness of handwriting, sentence construction ability, and in the case when stringent criteria are used, the expository ability may also play a role. If strict criteria are used, partial knowledge of word meaning is discounted. If looser standards are applied, the subjective judgement of raters may introduce error into the scores. The authors also point out that the constructed answer format is inefficient per unit of testing and scoring time.

The objections raised by Anderson and Freebody (1981) are well taken. These authors clearly prefer the yes-no format. While they do not rank the other three formats, it would appear that they consider the constructed answer second best. While planning the study, the author came to regard the

constructed format the most appropriate, since there was no empirical evidence and no personal experience with the yes-no format available at that time.

The constructed answer format has some special attractions in the foreign language context. When students learn a new language, they can use the other language as a kind of a "metalanguage". While it is by no means always easy to establish a one-to-one equivalence between the words of two languages, this can be done with a fair degree of approximation in the case of many words. It would appear that this is especially true of the type of vocabulary that is used in the early stages of foreign language learning. Many concepts in Western culture are roughly similar. Thus students only learn a new verbal label for already existing concepts. In many cases, then, the most direct way to probe the knowledge of a word meaning is to ask its equivalent in the other language.

Th interfering factor of spelling can be reduced by proper rater instructions and training, as was done in the present study. Raters were to focus only on semantic equivalence. The effect of sentence forming and expository ability can be minimized in using isolated words and requiring the production of their equivalents only.

The use of decontextualized words in language teaching is sometimes strenuously objected to (e.g., van Parreren, 1967; Carpay, 1975). Yet, Carpay (1975), who strongly recommends teaching words in contexts and fostering the ability of learning words incidentally from contexts, defends his use of decontextualized words in testing for practical reasons.

The present author believes that the role of context is even at present time poorly understood in spite of much work on it. For the purposes of the

present study, suffice it to point out that it is an empirical fact that when asked to give a foreign language equivalent to such words as "run", "speak", "house", "table", "happy", "small", "three", "Sunday", etc., most people do not need to ask for a context to clarify what meaning is meant. It is the "central", non-specialized, the "prototype" meaning that is meant.

The fact that most words have several meanings is not a major problem in foreign language instruction, either. Only the central meanings can be taught in the small amount of time that is available. Also in testing, students are only tested on their knowledge of central meanings. They are expected to know only those meanings that they have been taught. Naturally, raters must accept all non-specialized meanings as well.

In summary, it appeared that the constructed answer format was the most appropriate. It also appeared that context was not needed to "determine" what meanings were asked. Context was also undesirable for two other reasons. Putting words , say, in a sentence context would have increased reading time and made it possible to test fewer words. Yet, design considerations clearly indicated that a large sample of words was needed in order to minimize measurement error. Also, context might have introduced difficulty rather than facilitation (cf. Beck et al., 1983). Finally, the purpose was to obtain "robust" estimates of vocabulary size. Therefore, it was considered desirable to get an estimate of how many words students have easy and fluent access to in their long term semantic memory.

Domain Specification and Item Generation Rules

The above considerations were codified in the following domain specification and item generation rules:

Behavior

(1) When given a Finnish word in writing, the student can produce an acceptable English equivalent in writing (recall or active vocabulary). (2) When given an English word in writing, the student can produce an acceptable Finnish equivalent in writing (recognition or passive vocabulary).

Stimulus specification

The vocabulary presented in the core texts and extra (optional) texts in widely used English textbooks is listed. A stratified random sample is selected from the universe of such word lists. The words are presented without providing any context. Some of the words are used to measure both the passive and active knowledge of word meanings.

Response specification

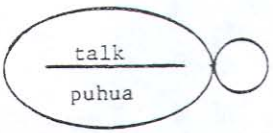
The student has to write the response in the space provided for that purpose.

Scoring

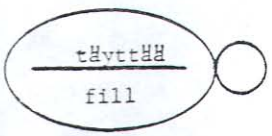
The responses are scored 0 - 1. A semantically acceptable and understandable response which may contain spelling errors is scored 1. In scoring active vocabulary, the decision is based on how the written English word would sound if read aloud. Thus, the student will get full marks if he/she has given the English equivalent of the Finnish word "talo" as "haus" instead of "house", since "haus" in Finnish orthography corresponds to the way "house" is pronounced in English.

Sample items

Intructions: "In this test you can show how well you know the English vocabulary included in your course work. Below are presented a number of Finnish words. Your task is to write the English equivalent on the line above the Finnish word. Write the word even if you may not be quite sure about the correct spelling, since spelling mistakes are a minor consideration in scoring."



"Write the Finnish equivalents of the following English words."



Instrumentation

The instruments prepared for the study consisted of a School Questionnaire, a Teacher Questionnaire, a Student Questionnaire, and of cognitive tests for the evaluation of students' performance in reading and listening comprehension, in grammar, and vocabulary. In order to give an idea of the size of the task, suffice it to mention that about half a million pages had to be printed for preparing the testing material needed to carry out the data collection. Since the present study focuses only on vocabulary learning, only the instruments developed for measuring vocabulary knowledge are discussed in this context.

As the section on item sampling has shown, there were several hundred items to be presented to students following the stratification plan of item selection. Two questions had to be addressed: (1) How many test forms should be formed? (2) How should items be allocated to the various test forms?

As Wolf (1979) has pointed out, the number of test forms depends on (1) the size of the item universe, (2) time available for testing, (3) the size of the student sample, (4) the assumed distribution of test scores, and (5) the estimated importance of certain items. On the basis of Wolf's criteria and the work done by Konttinen (1980) - described in the above - it was concluded that 10 forms were needed in order to satisfy the requirements of the intensive and extensive sampling strategies.

Words were divided into blocks on the basis of their stratum and these blocks were used to construct the different test forms. In the case of the "intensive sample" (fewer items, more students), the blocks appeared at least in two different test forms. In the case of the "extensive sample" (more items, fewer students), the blocks normally appeared only once. Typically, the active vocabulary was assigned to the intensive sample and the passive vocabulary to the extensive sample. The system of test form construction has been described in more detail in an earlier publication (Takala, 1982b).

Data Collection

All the test material was sent to schools at the end of March, 1979, and the English teachers administered the questionnaires and the cognitive tests in April according to carefully planned instructions. In spite of the new and complex system due to matrix sampling and the need to rotate several different test forms in class, the data collection worked well. There were very few

complaints and only a few clarifying questions. Several weeks had been spent on the preparation of the instructions, and that work clearly paid off (cf. Takala, 1982a).

The material was mailed back to the Institute, where the multiple choice parts of the assessment (listening and reading comprehension, and grammar) were scored with an optical reader. Equated test scores were quickly computed using a special program developed for the purpose and the schools were informed of student results, so that they could use the test results in May in student grading, if they so wished.

The number of students who took part in the tests is shown in Table 17.

Table 17

Number of Students in the English Assessment by Textbook and Set

Textbook	Set A	Set B	Set C	Total
Textbook 1 (Say it in English)	674	666	515	1,855
Textbook 2 (Welcome to English)	238	216	106	560
Total	912	882	621	2,415

Data Processing

The scoring was a formidable task even if it was simplified by focusing only on meaning equivalence and using a 0-1 scoring system. Altogether about 114,500 student answers had to be coded. This was done by a teacher of English employed at the Institute with funds allocated by the Ministry of Labor for unemployed academics. Coding was discussed with the coder and feedback was given to her after she had coded a few test booklet. She then

proceeded to score all student answers and entered the scores on optical reader answer sheets. This took several months.

The reliability of scoring was assessed by taking 50 - 100 test booklets randomly from each set of the two textbook series. The agreement was studied by using the percentage of agreement as the index of interrater agreement. The results are shown in Table 18.

Table 18

Interrater Agreement in Scoring Vocabulary Items

Set	Textbook 1		Textbook 2	
	2 raters	4 raters	2 raters	4 raters
A	96.1%	92.6%	96.1%	88.4%
B	97.3%	91.3%	95.4%	91.7%
C	97.7%	94.1%	97.7%	92.7%

The figures indicate that there was a high agreement among the raters. The chief scorer tended to score somewhat more strictly than the present author had intended. Thus the scores represent a relatively strict interpretation of students' knowledge of English vocabulary.

After the vocabulary items had been scored and entered on optical answer sheets, the items were run on a comprehensive data analysis program, which had been developed for the First Assessment. This was a logistic item analysis program (LOGIMA I). The analyses took a lot of time to carry out. The results of the item analysis were reported as the first stage in the study, creating a fairly large vocabulary item bank (Takala, 1982b, 1982c).

Before we turn to the results obtained in the study, it is appropriate to discuss briefly the method used in the analysis. Especially, since the results were obtained with a new computer program, it needs to be shown that they are reliable. In order to test the results obtained with the new program developed by Törmäkangas they were compared with the results of the standard SPSS Reliability program. This was done by taking a sample of 104 students all of whom had got the same 12 items and running the two programs on the same set of data. Using equations in Cronbach (Cronbach, Gleser, Nanda & Rajaratram, 1972), variance components were computed from the SPSS mean square indices. The results are shown in Table 19.

Table 19

Comparison between the Results Obtained with the Tormakangas Program and the SPSS Program

Index	Törmäkangas VARCOM/GSS	SPSS RELIABILITY
Mean	.4983974	.49840
σ^2 E (var. comp./subjects)	.0242110	.0242092
σ^2 I (var. comp./items)	.0676439	.0676229
σ^2 ExI (subj.x items + error)	.1641428	.16415
Alpha	-	.638997
Standard error for mean	.0774616	.0774675

The figures in Table 19 indicate that the results obtained with the two different programs are very similar, agreeing up to the fourth or fifth decimal point. Hence we have empirical evidence that the Tormakangas variance component program, which builds on the Sirotnik and Wellington (1977) system using generalized symmetrical sums (gss), produces results that are

highly comparable to other standard methods. Thus, the gss approach makes possible to do what standard programs do but it also allows the use of multiple matrix sampling (or incidence sampling in the Sirotnik & Wellington terminology), which a great advantage in many kinds of situations.

CHAPTER IX

RESULTS

Overview

The data that are reported here were provided by some 2,400 students who were chosen by randomized procedures to be representative of the Finnish-speaking students in the last grade (grade 9) of the Finnish comprehensive school system. Students had had seven years of English with a total of some 450 clock hours. The data were collected in 1979 in connection with a large-scale survey project called the "First National Assessment of Teaching in the Comprehensive School".

The fundamental purpose of this research was to estimate, on the basis of a sample of words, students' total passive and active vocabulary size in English. This was to be done so that the results would be generalizable not only to the entire student population but also to the subject content, i.e., the whole universe of taught vocabulary.

The purpose of this chapter is to review the results of the survey described in Chapters VII and VIII. First, descriptive statistical data based on variance component analysis will be presented to show the students' active and passive vocabulary size in English. That is followed by a discussion of the extent to which students' ability to use word-formation skills and context might affect the presented estimates of vocabulary size. The estimates of the passive and active vocabulary sizes covering all seven years of English study are presented first. Thus the figures cover the vocabulary taught both during the lower stage (grades 3-6) and the upper stage (grades

7-9). The estimates are given separately for the passive vocabulary size and for the active vocabulary size, which was estimated with two different types of sample: the intensive sample, which normally consisted of relatively few items answered by a fairly large number of students, and the extensive sample, which consisted of a greater number of items but fewer students answering each item.

Size of Overall Passive and Active Vocabulary

Students Using Textbook 1

The results for students using Textbook 1 are presented in Table 20.

Table 20

Estimated Total Means for Active and Passive Vocabulary Knowledge and their 95% Confidence Intervals in the Three Different Sets, Textbook 1

Vocabulary stratum, test and sample type	Set A (N= 674)		Set B (N= 666)		Set C (N= 515)	
	Mean	95% interval	Mean	95% interval	Mean	95% interval
	W= 2,498		W= 2,498		W= 1,470	
Passive	1,492	1,298 - 1,698	891	707 - 1,074	340	178 - 503
Active/intensive sample	1,441	1,258 - 1,625	821	655 - 987	379	293 - 466
Active/extensive sample	1,341	1,163 - 1,520	752	599 - 900	297	197 - 398

The results are presented separately for the three sets. Set A consists of the most able students (some 40% of the age group), Set B of intermediate students (also about 40% of the age group), and Set C includes those students who have made least progress in English (about 20% of the age group). As mentioned earlier, students and their parents choose the sets at the end of grade 6. Sets A and B give students full eligibility to apply to all post-

comprehensive schools, and they use the same textbook. Set C uses a different textbook, and gives eligibility to certain vocational schools only. Setting is now being phased out.

With the exception of the passive vocabulary in Set C, the estimates for the passive vocabulary size were larger than the ones for the active vocabulary size. In Set A, the estimated passive vocabulary size was 1,498 words, the size of the active vocabulary estimated with the intensive sample was 1,441 and that estimated with the extensive sample exactly one hundred words smaller (1,341). Due to measurement error, the range within which the estimates lie with 95% level of confidence varied from 300 words in the case of passive vocabulary, to 367 in the active intensive sample estimate, and 357 in the active extensive sample estimate. Thus, it can be stated with 95% level of confidence that the size of those students' passive vocabulary size who used Textbook 1 ranged between 1,298 and 1,698 words. Their active vocabulary size ranged between 1,163 and 1,625 words. If the confidence intervals did not overlap, we could conclude with 95% confidence that the passive and active vocabulary sizes are different. Since the upper limit of the active vocabulary size clearly overlaps with the lower limit of the passive vocabulary size, we cannot exclude with 95% confidence the possibility that the two vocabulary sizes are equal. Hence, the proper conclusion is that there is no clear difference in the passive and active vocabulary sizes of Set A students (i.e., the fast learners). A check with the z-test for proportions confirms the conclusion: a z-value of 1,96 is obtained when there is a small overlap of some 10 words or less. When there is no overlap, the z-value is more than two.

In Set B, the average size of passive vocabulary was 891, the active vocabulary estimated with the intensive sample was 821 and that estimated with the extensive sample 752. With 95% level of confidence the average passive vocabulary size ranged between 707 and 1,074 (a difference of 367 words), and the two active vocabulary estimates between 655 and 987 (332 words) and between 599 and 900 (301 words), respectively. Since the upper limit of the active vocabulary size estimate (987) clearly overlaps the lower limit of the passive vocabulary estimate (707), the conservative conclusion is that there is no clear difference between the passive and active vocabulary sizes of Set B students (i.e., students with average performance in English).

In Set C, the average size of passive vocabulary was 340, the active vocabulary as estimated with the intensive sample was somewhat higher at 379 and that estimated with the extensive sample 297. Using the 95% level of confidence as a guideline, the average passive vocabulary size ranged from 178 to 503, whereas the active vocabulary ranged 293-466 and 197-398 in the two different samples. The upper limit of the active vocabulary (466) overlaps clearly with the lower limit of passive vocabulary (178). Thus, we are again led to conclude that there is no clear difference between the passive and active vocabulary size of Set C students (i.e., slow learners).

Students Using Textbook 2

The corresponding results for students using Textbook 2 are presented in Table 21. The general pattern is exactly the same as that observed for Textbook 1: the highest average was found with the passive vocabulary and the active vocabulary estimate using the intensive sample was slightly higher

than that obtained with the extensive sample. This pattern applied fully for

Table 21

Estimated Total Means for Passive and Active Vocabulary with their 95% Confidence Intervals in the Three Different Sets, Textbook 2

Vocabulary stratum, test and sample type	Set A (N= 238)		Set B (N= 216)		Set C (N= 106)	
	Mean	95% interval	Mean	95% interval	Mean	95% interval
	W= 2,854		W= 2,854		W= 2,340	
Passive	1,598	1,255 - 1,940	999	628 - 1,369	538	234 - 842
Active/intensive sample	1,569	1,287 - 1,853	917	644 - 1,189	414	208 - 625
Active/extensive sample	1,515	1,270 - 1,760	891	671 - 1,112	282	150 - 419

Textbook 2. The only exception to this pattern appeared in the case of Textbook 1 users in Set C, where the passive vocabulary estimate was lower than the active vocabulary estimate. The overall estimates for Textbook 2 are higher than those for Textbook 1. The difference ranged from 118 to 174 for set A, from 96 to 139 for Set B, and from 35 to 84 for Set C. The only exception was Set C, where the size of active vocabulary estimated with the extensive sample was 15 words higher for Textbook 1 users.

Taking a closer look at the figures, we note that in Set A the mean of the passive vocabulary was 1,598 words, the active vocabulary estimated with the intensive sample was 1,569 words and with the extensive sample 1,515 words. When an interval is set within which the means can be stated to be located with 95% level of confidence, it is seen that the mean ranged from 1,255 to 1,940 passive words (range 685 words), and from 1,287 to 1,853 (566) and from 1,270 to 1,760 (490) estimated with the intensive and extensive

sample, respectively.

In Set B, the mean size of passive vocabulary was found to be close to one thousand words (999). The two estimates of active vocabulary size were 917 and 891, respectively. With 95% level of confidence the passive vocabulary mean ranged from 628 to 1,369 (741 words), and the active vocabulary mean estimates varied between 644 and 1,189 (545) and between 671 and 1,112 (441).

In Set C, the mean for the passive vocabulary was 528 words and the two active vocabulary estimates were 414 and 282 words, respectively. When 95% level of confidence was used to compute the range within which the mean can be expected to be located with high likelihood, it was found that the range varied from 234 to 842 (608 words) for the passive vocabulary, and from 208 to 625 (417) and from 150 to 419 (269) in the two estimates of the active vocabulary size.

As in the case of Textbook 1, when a 95% level of confidence is used to estimate the range of the passive and active vocabulary, it is noted that the upper limits of the active vocabulary estimates substantially overlap with the lower limits of the passive vocabulary estimates in all three sets. Hence the same conclusion is arrived at as in the case of Textbook 1 users: there is no appreciable difference between students' passive and active vocabulary sizes within each of the three sets.

To summarize the main findings, the average size of total passive vocabulary in Set A was found to be about 1,550 (Textbook 1: 1,498/ Textbook 2: 1,598) words with 95% level of confidence ranging from 1,255 to 1940; in Set B about 950 words (T 1: 891/T 2: 999) with a 95% confidence level range

of 628 to 1,369, and in Set C about 450 words (340/538) with a 95% confidence level range from 178 to 842. The average size of total active vocabulary was found to be about 1,450 (T 1: 1,441/1,341/T 2: 1,569/1,515) words in Set A (about 100 less than the passive vocabulary) with a 95% confidence level range from 1,163 to 1,853. The corresponding figures for Set B were about 850 words (T 1: 752/821/T 2: 891/917) - again about 100 words less than the active vocabulary with a 95% confidence level interval ranging from 599 to 1,189. The average size of Set C students' active vocabulary was about 350 words (T 1: 297/379/T 2: 282/414) - about 50 words less than the passive vocabulary - with a 95% confidence level ranging from 150 to 625.

Within each set, the confidence interval of the active and passive vocabulary estimates overlapped considerably. Hence, the conservative estimate was that there is no reliable difference in students' passive and active vocabulary size at the end of the comprehensive school. The possible reasons for this somewhat unexpected result will be discussed later.

With the exception of Set C/Textbook 2, the 95% confidence intervals of both passive and active vocabulary size estimates of the three sets did not overlap each other. Consequently, we can conclude that the average performance of Set C students was definitely poorer than that of Set B and especially Set A students, and that Set A students also clearly outperformed Set B students. The three sets represent clearly different student populations. This result agrees with earlier findings made by the present writer and his colleagues.

Size of Passive and Active Vocabulary in Different Vocabulary Strata

This section will examine the results in greater detail by looking at the passive and active vocabulary estimates in the different strata and in different types of sample. As in the case of the total vocabulary size estimates, the results will first be presented separately for the two textbooks. After that the two will be compared, and finally the main findings will be summarized.

Students Using Textbook 1

As Table 22 shows, the number of words taught for Set A and B students during the lower stage (grades 3-6) and the upper stage (grades 7-9) is roughly of the same order of magnitude (1,011 vs. 1,164). The figures indicate that an average of 250-300 more words belonging to the lower stage stratum are known than words belonging to the upper stage stratum. There is no overlap between the 95% confidence interval estimates of the three different sets. Consequently, the earlier finding related to the estimates of total vocabulary sizes is replicated in the case of stratum-wise analysis: the students in different sets have clearly different achievement levels and thus "come from different populations". The results will now be reported by vocabulary stratum beginning with the words taught during the first four years of English, when students of different abilities studied together following the same syllabus and using the same teaching materials.

Lower stage vocabulary. On a closer examination of the figures we note that out of the 1,011 words taugh to all students using Textbook 1 during the lower stage, Set A students had learned to recognize the meanings of 895 words, taught during the lower stage and the estimate for the active vocabu-

lary was about one hundred words less: 795 words based on the intensive esti-

Table 22

Estimated Means with their 95% Confidence Intervals for Passive and Active Vocabulary Knowledge in Three Different Sets, by Vocabulary Stratum and Type of Sample, Textbook 1

Vocabulary stratum, test and sample type	Set A		Set B		Set C	
	Mean	95% interval	Mean	95% interval	Mean	95% interval
Lower stage vocabulary	W= 1,011		W= 1,011		W= 1,011	
Passive	895	832 - 956	659	556 - 762	266	167 - 365
Active/intensive sample	795	730 - 860	516	440 - 593	313	258 - 369
Active/extensive sample	733	663 - 803	505	439 - 572	234	173 - 296
Upper stage vocabulary	W= 1,164		W= 1,164		W= 405	
Passive	594	475 - 712	232	151 - 312	74	11 - 138
Active/ intensive sample	549	474 - 625	235	176 - 294	63	34 - 92
Active/extensive sample	464	396 - 533	183	130 - 235	63	24 - 102
Upper stage extra vocabulary	W= 323		W= 323		W= 54	
Active/ intensive sample	97	55 - 140	70	39 - 100	3	1 - 5
Active/ extensive sample	144	104 - 184	64	29 - 99	-	- -

mate and 733 words based on the extensive estimate. Due to measurement error, the range within which the average passive vocabulary varied was from 832 to 956 words (124 words), while the active vocabulary estimate with the inten-

sive sample varied from 730 to 860 words (130 words) and with the extensive sample from 663 to 803 words (140 words). There is only a slight overlap between the upper limit of the active vocabulary of the intensive sample and the lower limit of the passive vocabulary, whereas the extensive sample estimate does not overlap with that of the passive vocabulary. Thus, we can conclude that there seems to be some difference in students' passive and active vocabulary taught during the lower stage in favor of the passive vocabulary. This generalization concerns Set A-students (i.e., the fast learners, about 40% of the whole age group).

In Set B, the average size of passively known vocabulary covering those words that were taught during the lower stage was 656 words, active vocabulary estimated with the intensive sample 516 (140 less) and 505 (151 less) estimated with the extensive sample. The 95% confidence intervals ranged from 556 to 762 (206 words) for the passive vocabulary , from 440 to 593 (153 words) and from 439 to 572 (133 words) for the two active vocabulary size estimates. Since the ranges of the three estimates overlap, we conclude that in Set B the passive and active vocabulary sizes are essentially of the same size.

In Set C, the estimate for the passive vocabulary size related to the vocabulary taught during the lower stage was 266 words, and the two active vocabulary size estimates were 313 and 243 words for the intensive and extensive samples, respectively. With 95% confidence, the range of the average passive vocabulary ranged from 167 to 365 words (198 words), and the two active vocabulary size estimates from 258 to 369 (111 words) and from 173 to 296 (120 words). Since the ranges clearly overlap, the conclusion is that

in Set C the size of actively and passively known vocabulary of those words that were taught during the lower stage is essentially the same.

Upper stage vocabulary. Figures in Table 22 show that out of the 1,164 words taught to Set A and B students using Textbook 1 during the upper stage, students in Set A had learned to know passively 594 words and actively 549 words as estimated with the intensive sample and 464 measured with the extensive sample. Measurement error entails that the intervals within which the means of the passive and active vocabulary size estimates are located with 95% level of confidence overlap considerably: 475 - 712 (237 words) for the passive vocabulary and 474 - 625 (151 words) for the active intensive sample and 396 - 533 (137 words) for the active extensive sample. This warrants the conclusion that there is no appreciable difference between the passive and active knowledge of vocabulary taught during the upper stage among students who read Textbook 1 in Set A.

In Set B, the average size of passively known vocabulary out of the 1,164 words taught during the upper stage was 232 words. The corresponding figures for the active vocabulary estimated with the intensive and extensive sample was 235 and 183 words, respectively. The conclusion is that there is no real difference between the size of passive and active knowledge of this vocabulary stratum. This is confirmed by a look at the intervals within which the vocabulary estimates are located with 95% level of confidence: 151 - 312 (161 words) for the passive vocabulary, and 176 - 294 (118 words) for the active intensive sample and 130 - 235 (105 words) for the active extensive sample.

In Set C, the average size of passive knowledge of the 405 words taught during the upper stage was 74 words, and the estimates of active vocabulary sizes arrived at by means of the intensive and extensive samples was 63 words in both cases. Measurement error causes the intervals within which the means are located with 95% level of confidence to overlap: 11 - 138 (127 words) for the passive vocabulary, and 34 - 92 (58 words) and 24 - 102 (78 words) for the two active vocabulary estimates. Hence, once more we are led to conclude that in Set C, those students who studied the vocabulary taught during the upper stage in Textbook 1 learned about the same amount of words passively and actively.

Upper stage etxra vocabulary. In the case of Textbook 1, only the active command of this vocabulary stratum was measured. Looking at Set A, the figures in Table 22 show that out of the 323 words included in this stratum students had learned to know actively 97 to 144 words as estimated with the intensive and extensive sample, respectively. The 95% confidence levels for the estimates ranged from 55 to 140 words (85 words) for the intensive sample and from 104 to 184 (80 words) for the extensive sample.

As for Set B, the corresponding estimates were 70 and 64 words with 95% confidence intervals ranging from 39 to 100 (61 words) and from 29 to 99 (70 words) for the intensive and extensive samples, respectively.

Students in Set C had learned actively 3 words out the 54 included in their extra vocabulary stratum during the upper stage. The 95% confidence interval for the mean ranged from 1 to 5 words.

Students Using Textbook 2

As Table 23 shows, unlike the case of Textbook 1, where the number of words taught during the lower stage for all students and for Set A and B

students during the upper stage was of the same order of magnitude (1,011 vs. 1,164), during the lower stage the vocabulary taught for Set A and B students is about twice the number taught (812 vs. 1,690). This difference offers a "natural experiment", which will be discussed later on. Also, in contrast to Textbook 2, where the average for words taught during the lower stage was about 200-300 words higher than for words taught during the upper stage, students using Textbook 1 had learned about the same amount of words belonging both to the lower and upper stage vocabulary stratum. With the exception of the passive lower stage vocabulary for Set B, there is no overlap between the 95% confidence interval estimates of the three different sets. Set A did clearly better than Set B and especially Set C, and Set B outperformed Set C. This result is consistent with what was found also in the case of Textbook 1 users. After this overview, the findings will now be reported by vocabulary stratum, beginning with the words taught to all students during the lower stage (the first four years of English when the classes were mixed-ability classes).

Lower stage vocabulary. A closer look at the figures in Table 23 shows, that out of the 812 words taught during the lower stage, students in Set A had learned to recognize the meanings of 741 words (i.e., know them passively). The estimates for the actively known vocabulary are about 100 words lower: 615 for the intensive and 642 for the extensive sample. Measurement error caused the interval within which the average can be expected to be located with 95% level of confidence to vary from 655 to 827 words (172 words) for the passive vocabulary, and from 547 to 684 (137 words) for the active intensive sample and from 593 to 692 (99 words) for the active exten-

sive sample. Since the upper limit of the active vocabulary estimates overlap with the lower limit of the passive vocabulary estimates, the conclusion is that for Set A students using Textbook 2, the passive and active knowledge of lower stage vocabulary is of the same order of magnitude.

Table 23

Estimated Means with their 95% Confidence Intervals for Passive and Active Vocabulary Knowledge in the Three Different Sets, by Vocabulary Stratum and Type of Sample, Textbook 2

Vocabulary stratum, test and sample type	Set A		Set B		Set C	
	Mean	95% interval	Mean	95% interval	Mean	95% interval
Lower stage vocabulary	W= 812		W= 812		W= 812	
Passive	741	655 - 827	544	353 - 735	196	108 - 283
Active/intensive sample	615	547 - 684	348	271 - 424	134	91 - 177
Active/extensive sample	642	593 - 692	431	362 - 502	217	136 - 297
Upper stage vocabulary	W= 1,690		W= 1,690		W= 1,078	
Passive	771	595 - 946	424	265 - 583	134	27 - 221
Active/intensive sample	801	659 - 943	458	326 - 590	120	65 - 175
Active/extensive sample	761	608 - 913	389	269 - 509	65	8 - 122
Upper stage extra vocabulary	W= 352		W= 352		W= 450	
Passive	95	5 - 185	20	0 - 61	98	41 - 155
Active/intensive sample	153	81 - 226	111	47 - 175	160	46 - 274
Active/extensive sample	112	69 - 155	71	40 - 101	-	- -

In Set B, the average size of passively known lower stage vocabulary (812 words) was 544 words, while the active vocabulary estimates were 348 for the intensive sample and 431 for the extensive sample. The 95% confidence intervals for the averages ranged from 353 to 735 (382 words) for the passive vocabulary, from 271 to 424 words (153 words) for the active intensive sample estimate and from 362 to 502 (140 words) for the active extensive sample estimates. These figures again lead us to conclude that in Set B the passive and active knowledge of lower stage vocabulary is essentially the same.

In Set C, the passive knowledge of the 812 words taught during the lower stage was estimated to average 196 words. The estimated averages for the actively known vocabulary were 134 for the intensive sample and 217 for the extensive sample. Due to measurement error, we can conclude with 95% level of confidence that the passive vocabulary knowledge ranged from 108 to 283 (175 words), whereas the active vocabulary ranged from 91 to 177 (86 words) and from 136 to 297 (161 words) for the two samples, respectively. Overlap in the vocabulary estimate ranges entails the conclusion that in Set C, the passive and active knowledge of the lower stage vocabulary is comparable.

Upper stage vocabulary. Figures in Table 23 reveal that out of the 1,690 words taught to Set A and B students during the upper stage, Set A students had learned to know passively an average of 771 words and actively 801 words as estimated with the intensive sample and 761 words as estimated with the extensive sample. The similarity of the estimates suggests that there is no real difference in the passive and active knowlege of upper stage vocabulary in Set A. This is confirmed by a look at the 95% confidence level intervals: they ranged from 595 to 946 (351 words) in the case of the passive vocabula-

ry, from 659 to 943 (284 words) for the active intensive sample and from 608 to 913 (305 words) for the active extensive sample. The overlap is complete.

In Set B, the average size of passively known vocabulary out of the 1,690 included in the course was 424 words. The corresponding figures for the active vocabulary estimated with the intensive sample was 458 words and 389 estimated with the extensive sample. These figures suggest that there is no real difference in passive and active vocabulary size for the upper stage vocabulary. This is confirmed by an inspection of the intervals within which the averages can be expected to lie with 95% levels of confidence: 265 - 583 (318 words) for the passive vocabulary, and 326 - 590 (264 words) for the active intensive sample and 269 - 509 (240 words) for the active extensive sample. The overlap between the intervals is very substantial.

In Set C, the average size of passive knowledge of the 1,078 words included in the course material was 134 words. The estimates for the active knowledge were 120 words with the intensive sample and 65 words with the extensive sample. The difference do not appear to very large. This is confirmed by a look at the 95% confidence level intervals: they were 27 - 221 (194 words) for the passive vocabulary, and 65 - 175 (110 words) for the active intensive sample and 8 - 122 (114 words) for the active extensive sample. Substantial overlaps indicate that the passive and active knowledge of vocabulary taught during the upper stage is essentially the same.

Upper stage extra vocabulary. Looking at the figures in Table 23, we note that curiously enough more extra vocabulary was inluded in Set C material than in the material meant for Sets A and B. In Set A, students had learned to know passively an average of 95 words out of the 352 included in the course. Contrary to the general pattern, the estimates for the active

knowledge are higher than the estimate for the passive knowledge: 153 words with the intensive sample and 112 with the extensive sample. The 95% confidence level intervals for the estimates overlap considerably: 5 - 185 (180 words) for the passive vocabulary, 81 - 226 (145 words) for the active intensive sample and 69 - 155 (86 words) for the active extensive sample. Thus we are led to conclude that the passive and active knowledge of upper stage extra vocabulary is roughly the same in Set A.

In Set B, students had learned to know passively about 20 out of the 352 words included in the course materials. The estimates for the active vocabulary size were 111 words for the intensive sample and 71 for the extensive sample. The intervals within which the averages are expected to lie with 95% level of confidence were 0 - 61 (61 words) for the passive vocabulary, and 47 - 175 (128 words) for the active intensive sample and 40 - 101 (61 words) for the active extensive sample. The intervals overlap substantially and entail the conclusion that in Set B the passive and active knowledge of upper stage extra vocabulary is of the same order of magnitude.

In Set C, the average passive knowledge of the 450 words included in the course materials was 98 words and the active knowledge 160 words (intensive sample only). The 95% confidence level intervals were 41 - 155 (114 words) for the passive vocabulary and 46 - 274 (228 words) for the active vocabulary. Since they overlap substantially, the conclusion is again that there is no real difference in Set C in the passive and active knowledge of upper stage extra vocabulary.

Summary

To summarize the main findings of the stratum-wise analysis of passive and active vocabulary knowledge, it was found that Textbook 1 taught about the same amount of words during the lower stage and the upper stage (1,011 vs. 1,164). Yet, in all three sets the students knew passively and actively about 200 to 300 more lower stage words than upper stage words. Within each set, the confidence intervals of the active and passive vocabulary estimates overlapped considerably, leading to the conclusion that there is no reliable difference in students' passive and active knowledge of words taught at different stages of English study.

Of the 1,011 words taught during the lower stage, students in Set A had learned passively and actively about 800 words, Set B students about 550 words, and Set C students about 270 words. Of the upper stage vocabulary consisting of 1,164 words, Set A students had learned passively and actively about 535 words, Set B about 220 words. Out of the 405 words taught, Set C learned some 65 words. Of the 323 words included in the upper stage extra vocabulary, Set A had learned about 120 words, and Set B about 65 words. Of the 45 extra vocabulary, Set C had learned some 3 words. Thus, substantial and significant differences were found between sets in all vocabulary strata.

Unlike in Textbook 1, Textbook 2 taught more than twice more words for Sets A and B during the upper stage than during the lower stage (1,690 vs. 812). Also in contrast to Textbook 1, there was no major difference in students' vocabulary size related to the lower stage and upper stage words. No reliable difference was found between the passive and active vocabulary size estimates in any strata.

Of the 812 lower stage vocabulary, students in Set A had learned passively and actively about 665 words, Set B students about 440 words, and Set C students some 180 words. Of the 1,690 words taught to Sets A and B during the upper stage, Set A had learned some 775 words passively and actively, and Set B students about 425 words. Of the 1,078 words taught to Set C, about 105 were learned. Of the 352 upper stage extra vocabulary taight to Sets A and B, Set A had learned about 120 words, and Set B some 65 words. Of the 450 words included in the course material for Set C, about 130 words were learned.

The differences between the three sets were large and significant in all three vocabulary strata, except for the upper stage extra vocabulary.

Relationship Between Taught and Learned Vocabulary

This section will address the question of what was learned of the taught vocabulary. In keeping with the general quantitative orientation of the study, the focus will be on the quantitative relationship between taught and learned vocabulary. As in the foregoing report on the absolute vocabulary size estimates, this discussion on the relative vocabulary size estimates will begin with the total vocabulary estimates and move to deal with the three vocabulary strata. The two textbooks will be dealt with together.

Total Vocabulary

The data concerning the relationship between taught and learned vocabulary for students using Textbook 1 are presented in Table 24 and those for Textbook 2 in Table 25. The results are roughly similar for both textbooks. Set A students have a passive and active knowledge of about 55% of taught vocabulary. The corresponding figures for Sets B and C are about 32% and 20%, respectively.

Different Vocabulary Strata

An inspection of Tables 24 and 25 shows that the vocabulary first taught

Table 24

Estimated Mean Proportions for Passively and Actively Known Vocabulary in Relation to Taught Vocabulary in the Three Different Sets, by Vocabulary Stratum and Type of Sample, Textbook 1

Vocabulary stratum, test and sample type	Set A		Set B		Set C	
	Mean	95% interval	Mean	95% interval	Mean	95% interval
Lower stage vocabulary	W= 1,011		W= 1,011		W= 1,011	
Passive	.88	.82 - .95	.65	.55 - .75	.26	.16 - .36
Active/intensive sample	.79	.72 - .85	.51	.43 - .59	.31	.25 - .36
Active/extensive sample	.72	.66 - .79	.50	.43 - .57	.23	.17 - .29
Upper stage vocabulary	W= 1,164		W= 1,164		W= 405	
Passive	.51	.41 - .61	.20	.13 - .27	.18	.03 - .34
Active/intensive sample	.47	.41 - .54	.20	.15 - .25	.16	.08 - .23
Active/extensive sample	.40	.34 - .46	.16	.11 - .20	.15	.05 - .25
Upper stage extra vocabulary	W= 323		W= 323		W= 54	
Active/intensive sample	.30	.17 - .43	.22	.12 - .31	.06	.02 - .09
Active/extensive sample	.45	.32 - .57	.20	.09 - .31	-	- -
Total	W= 2,498		W= 2,498		W= 1,470	
Passive	.60	.52 - .68	.36	.28 - .43	.23	.12 - .34
Active/intensive sample	.58	.50 - .65	.33	.26 - .40	.26	.20 - .32
Active/extensive sample	.54	.47 - .61	.30	.24 - .36	.20	.13 - .27

during the lower stage was known relatively better than vocabulary taught during the upper stage, and the extra vocabulary included in the upper stage was known least well of all.

In Set A, the proportion of the lower stage vocabulary known passively was about 90% and about 75-80% known actively. About 45% of upper stage vocabulary was known both passively and actively. About 35-40% of upper stage extra vocabulary was known passively and actively.

In Set B, the proportion of the lower stage vocabulary that was known passively was about 65% while the corresponding figure for active knowledge was about 50%. The passive and active mastery of upper stage vocabulary was on the order of 20-25%. About 20% of the upper stage extra vocabulary was learned passively and actively.

In Set C, about 25% of lower stage vocabulary was known passively and actively. The share of upper stage vocabulary learned passively and actively was about 10-15%. The estimates for the upper stage extra vocabulary ranged from 6% for active knowledge for Textbook 1 to 22% for the passive knowledge of Textbook 2 words.

As Tables 24 and 25 indicate, there was an interesting difference between the two textbooks: Textbook 1 taught about 200 more words during the lower stage than Textbook 2 (1,011 vs. 812 words), whereas in the upper stage Textbook 2 included about 500 more words for Sets A and B than Textbook 1 (1,690 vs. 1,164) and for Set C more than twice the number of words (1,078 vs. 405). Textbook 2 contained ten times more upper stage extra vocabulary than Textbook 1 for Set C (450 vs. 45) whereas the the amount for Sets A and B was about the same (352 vs. 323).

Since the proportion of known vocabulary was roughly similar for both textbook users, it appears that Textbook 1 with its larger input during the lower stage was better adapted to students' learning capacity. By contrast, its low input during the upper stage was less than optimal, and the clearly higher input of new words by Textbook 2 led to a higher learning yield.

It was noted that a larger proportion of lower stage vocabulary was known than of upper stage vocabulary. Several reasons could be advanced to explain the observed trend. First, it is possible that the words were chosen by following frequency counts quite closely and were either naturally or by design repeated often, even during the upper stage. Second, it is possible that the lower stage vocabulary is somehow inherently more learnable than the upper stage vocbulary. It might, for example, be more concrete. A third possibility is that younger students (aged 9-13) learn foreign words better than older students (aged 13-16), either because of more appropriate processing or higher motivation or both. A fourth possibility is that some kind of a plateau exists in vocabulary learning. The data do not make it possible to test any of the above hypotheses.

Effect of Students' Word-formation and Context-utilization Skills on Vocabulary Size Estimates

This section gives an account of a small-scale study that was carried out in order to explore the extent to which students' ability to use word analysis skills and to utilize context for the recognition of word meanings might affect the estimates of students' passive and active vocabulary size that were reported in the above.

Such skills were not expected to be highly developed for several reasons. First, English and Finnish are not related languages and there are few cognates or words borrowed from English. Thus, there is not sufficient similarity to entice guessing and inference based on analogies. Second, students are still learning more advanced grammatical structures like past conditional and the passive voice. Due to the legacy of the audiolingual methodology, the emphasis in the early stages of second/foreign language learning can be expected to be on grammar rather than on vocabulary. Third, and related to point two, it is expected that work on written and spoken texts in class is intensive, i.e., extensive reading and listening is not expected to have been very common. Thus, students are not expected to have had much practice in inferring word meanings from context.

The test was arranged such that students first had to produce the English equivalents of Finnish basic words. The answer slips were collected. Students then produced either related derived or compound words. The answer slips were again collected before students were asked to write the Finnish equivalents of English words, which were either derived or compounded from the basic words. After the answer slips had been collected, students had the same task but this time the words were embedded in a sentence context.

Two different sets of words lists were constructed and rotated in classes in order to be able to cover a larger number of words. The two word lists were collapsed in the analysis. Before the results of this exercise are reported, some data are presented to demonstrate the extent to which the students participating in the check were comparable to the ones who participated in the main study. No great differences were expected since the students in the check came from the same schools that had used Textbook 1 in

the main testing. The time interval between the two measurements was three years.

The comparison between the two student groups is possible because the words used in the check (N= 173) were chosen from the larger word corpus used in the main testing. The results of the exercise are presented in Table 26. Quotation marks are used to indicate that the stimulus words were presented in Finnish and their English equivalents are used here.

Table 26

Comparison of Proportion Correct Scores for Some Words Used in the Main Testing and in the Word-formation and Context Utilization Skill Check

Word	Main testing	Word-formation and context check
"price"	.59	.59
"harm"	.23	.45
"boat"	.89	.97
"shed"	.07	not tested
"boatshed"	not tested	.25
"write"	.63	not tested
"letter"	.85	.85
"letter-writer"	not tested	.45
"ten"	not tested	.98
"hospital"	.89	.94
"play"	.80	.80
"player"	.52	.61
"record"	.74	.86
"real"	.15	.27
"add"	.06	.13
Mean	.57	.65

The mean proportion correct was somewhat higher in the check testing than during the main testing (.65 vs. .57; z= 1.354; critical value 1.96). Thus, the results that were obtained in the check testing are probably a good approximation of the main test students' ability to utilize word-formation

knowledge and contextual support in recognizing and producing derived and compounded words.

After establishing that the check students were neither exceptionally poor nor good in vocabulary knowledge but in fact roughly comparable to the students who participated in the main testing, we can now move to examine the students ability to use word-formation and context utilization skills. The design of the experiment was described in the above. The results are shown in Table 27. The figures indicate the mean proportions of correct answers. In some cases the mean proportion correct scores have also been computed for the three sets. They are presented beneath the grand means in the order from Set A through Set B to Set C.

Table 27

Proportion Correct Means for the Active Knowlege of Basic Word Forms and of Related Derived and Compounded Words as well as for the Passive Knowledge of Other Derived or Compounded Words Presented with and without Context

Basic form	Derived or compunded forms		Derived or compounded forms	
			without context	with context
"play"	"player"	"playing"	People liked Dad's playful speech very much.	
.80	.61	.55	.05	.06
.86/.76/.58	.86/.40/.08	.56/.46/.33	.09/.00/.00	.12/.00/.00
"think"	"thinker"	"thinking"	Is it thinkable that we are wrong and he may be right?	
.75	.41	.60	.08	.18
.84/.69/.42	.64/.20/.00	.75/.49/.17	.18/.03/.00	.29/.06/.00
"letter"	"letter-writer"	"business-l."	Please, write more carefully. Your lettering is very bad.	
.85	.45	.52	.08	.52

table continues

Table 27 (cont.)

Basic form	Derived or compounded forms		Derived or compounded forms	
			without context	with context
"alarm"	"alarming"		Don't listen to him! He is such an alarmist.	
.49	.08		.15	.03
"live"	"lively"	"living"	This place is quite livable.	
.75	.19	.40	.04	.08
"cheer"	"cheerful"	"cheering"	The cheerlessness of our lives made us hate each other.	
.02 .04/.00/.00	.02 .04/.00/.00	.01 .02/.00/.00	.03 .04/.00/.00	.04 .02/.00/.00
"harm"	"harmless"	"harming"	Don't eat it! It may be harmful to you.	
.45 .68/.17/.08	.17 .25/.03/.00	.19 .25/.09/.00	.23 .30/.17/.00	.47 .68/.37/.08
"dirty"	"dirtiness"	"to dirty"	"How much did it cost?" - "I got it dirt-cheap."	
.56	.20	.14	.22	.30
lazy"	"laziness"		No use asking them to do it. They are such lazybones, all of them.	
.46 .48/.40/.00	.06 .09/.01/.00		.05 .25/.24/.00	.50 .64/.33/.00
"cheap"	"cheapness"		She never spends a penny. She is such a cheapskate.	
.43	.10		.00	.42
"easy"	"easiness"	"easily"	That's good news! That eased my mind a lot.	
.66	.08	.39	.11	.53
"awful"		"awfully"	The awfulness of it all made her cry.	
.65 .82/.47/.00		.28 .32/.19/.00	.07 .09/.00/.00	.16 .23/.03/.00
"real"	"unreal"	"reality"	Tell me honestly. Do we realistically have a chance of winning?	
.27	.16	.13	.21	.42

table continues

Table 27 (cont.)

Basic form	Derived or compounded forms		Derived or compounded forms	
			without context	with context
"a record"	"record-shop"		It's a very good recording. The sound is fantastic.	
.86	.82		.25	.35
"boat"	"boatshed"		Boating can be a very expensive hobby.	
.97	.25		.44	.73
"hospital"	"hospital-bed"		He was very ill. He was hospitalized for six months.	
.94	.78		.05	.54
"ten"	"tenfold"		I have money problems. Can you lend me a tenner for a week?	
.98	.01		.01	.27
"price"	"price-tag"	"low-priced"	Your help has been priceless to me.	
.58	.19	.33	.09	.17
.64/.44/.50	.23/.10/.08	.55/.27/.08	.18/.01/.00	.30/.01/.00
"low"	"to lower"		The machine needs only low-voltage electricity.	
.33	.16		.15	.33
.39/.24/.17	.21/.03/.16		.16/.07/.00	.46/.17/.00
"add"	"addition"		This is not enough. We need an additional five pounds.	
.13	.08		.14	.33
Total				
.59	.24		.13	.32

When students were asked to give the English equivalents of twenty Finnish basic words, the average proportion correct was .59. When they were to write either a common derived word (-er, -ful, -less, -ness, etc) or a compund word or both, the performance fell clearly to .24. This difference is statistically significant (z= 6.616; critical value 1.96). Even in the case

of producing a derived word from a verb to designate an actor (like "play" - "player", "think" - "thinker") the inability of Set B and especially of Set C students to use analogy caused a clear drop in the proportion correct score from .80 to .61 for "play" ($z= 4.294$) and from .75 to .60 for "think" ($z= 3.000$). The same trend was observed in all other derivation categories: verbal nouns formed with -ing (e.g., thinking, living; $z= 4.549$), abstract nouns with -ness (e.g., dirtiness, easiness; $z= 8.335$), and regular adverbs with -ly (easily, awfully; $z= 5.704$). The performance on compounds (e.g., letter-writer, hospital bed) was also lower than on the basic forms ($z= 6.680$).

By contrast, a context consisting of one or two sentences was found to help students to recognize and indicate the meanings of derived or compounded words (other than those discussed in the above). The mean proportion correct for decontextualized words was .13 and that for contextualized words .32 ($z= 4.241$). Especially the two-sentence contexts, in which one sentence preceded another in which the derived or compounded word was embedded (e.g., That's good news! That eased my mind a lot), were very helpful (mean proportion correct .39 vs. 13 with not context; $z= 5.509$) but one-sentence contexts were also useful (mean proportion correct .22 vs. .12; $z= 2.475$).

To summarize, the expectation that students would not be very proficient in word-formation and context-utilization skills was borne out. When active knowledge was used as the testing form, the means for proportion correct were considerably lower for derived and compounded words than for basic words (derived: .24 vs. basic: .59, $z= 6.616$). When passive knowledge was required and context was provided, the result was somewhat better (derived with context: .32 vs. derived with no context: .24, $z= 1.891$) but the difference

falls slightly below the level of significance. A context consisting of one or two sentences was helpful for the recognition of derived and compounded words as a comparison with no context help showed (.32 vs. 13, z= 4.241). Two-sentence contexts, in particular, proved effective.

The final conclusion is, however, that the estimates given in the main body of this chapter need not be upgraded twofold or fourfold, Nagy and Anderson (1982) suggest for SEM 2-level ability. By taking the figures in Tables 26 and 27 as a starting point, we tried to estimate the the increase through several ways to check out the outcome. Making a simplifying assumption that for each word in the original word universe, there is one simple and frequent derived and/or compound word, the following estimates appear defensible for Textbook 1, which was only used for this purpose: for non-context aided active vocabulary, the increase for Set A is about 32% (from 1,450 to 1,900 words), for Set B about 16% (from 850 to 990 words), and for Set C about 7% (from 350 to 375). For context aided passive knowledge (i.e., recognition) of words, the increase for Set A is about 45% (from 1,450 to 2,100 words), for Set B about 17% (from 850 to 1000 words). Probably due to motivation problems, Set C students usually left a blank answer slip when context aided meaning recognition was tested. Thus it is not possible to estimate what the increase is for them, but it can be conjectured that it is not much larger than the 7% for non-context increase, bringing their vocabulary to the region of 400 words.

For someone who is not familiar with vocabulary research or done research in this area, the revision of the estimate by up to 45% may seem a large change. The history of vocabulary research shows, however, that

variation in the estimates is often of the order of 5 to 10 times (i.e., 500% - 1000%).

Some Generalizability Considerations

Methods used in generalizability studies are usually based on variance analysis and the estimation of variance components. In a p x i (persons x items) design, which was used in this study, the variance of a person's observed score on one item consists of a person variance component, an item variance component, and the residual variance component (person and item interaction component plus error component (Cronbach, Gleser, Nanda & Rajaratram, 1972; Konttinen, 1980). This section will present the results that were obtained with the specially designed variance component analysis that uses generalized symmetrical sums (Törmäkangas, in press).

Tables 28, 29 and 30 show the variance components and their proportional contributions to explaining variance in scores. The values of s, the person variance component, indicate the variance of proportion correct scores in the whole student population in a vocabulary test where all students had received all items, i.e., the variance of proportion correct scores of a perfectly reliable test in the whole student population. The values of i, the item variance component, shows the difficulty variance of items in the whole student population. Thus, this component describes the homogeneity of items in terms of their difficulty. The values of the term si+e indicate the size of the subject and item interaction but it also includes the error variance.

In 13 our of 47 cases the variance components were negative. They were always related to the subject component and in all but two cases within the 95% confidence level interval from zero, and statistically speaking zero. As

is customary in generalizability studies, the components are treated as zeros in further analyses.

Another generalization is that in 44 cases out of the total of 47, the variance component for items was larger than the subject variance component. Usually the difference is sizable, the items component being twice or three times larger than the subjects component. This confirms the expectations that were held during the design stage: it is relatively easy to get an estimate of the "typical student" but it is much more difficult to talk about the "typical word", since difficulties vary so much across words. Even if the subject and item interaction component that also includes the error component is usually the largest of the three components, it is smaller than has usually been the case in earlier studies in Finland and can be regarded as relatively small. This means that students can be arranged in the order of ability with a relatively small number of items, since an easy item tends to be easy for all students and a difficult item tends to be difficult for everybody.

Word difficulty seems to be stable across students but words differ greatly in terms of their difficulty. Several questions can be raised to deal with this observation. First, to what extent is the difficulty variation likely to be due to the way words have been taught? Has there been substantial difference in the amount that different words have been repeated in the teaching material and in classroom discourse? Are recently taught words known and remembered better than words taught at an early stage?

Table 28

Variance Components and their Relative Contributions, with Students (s), Items (i), and Student and Items Interaction plus Error (si+ e) as Sources of Variation, Textbook 1, Sets A and B

Vocabulary stratum, test and sample type	Set A			Set B		
	s	i	si + e	s	i	si + e
Lower stage Passive	.0209	.0221	.0599	.0569	.0626	.1101
	20.3%	21.5%	58.2%	24.8%	27.3%	47.9%
Active/ intensive sample	.0141	.0590	.0962	.0262	.0863	.1389
	8.3%	34.9%	56.8%	10.4%	34.3%	55.3%
Active/ extensive sample	.0105	.0635	.1266	.0194	.0582	.1735
	5.3%	31.6%	63.1%	7.7%	23.2%	69.1%
Upper stage Passive	-.0042	.0575	.1993	.0185	.0235	.1184
	0.0%	22.4%	77.6%	11.5%	14.7%	73.8%
Active/ intensive sample	.0317	.0674	.1511	.0170	.0449	.0997
	12.7%	27.0%	60.3%	10.5%	27.2%	61.7%
Active/ extensive sample	.0322	.0645	.1440	.0136	.0368	.0821
	13.4%	26.8%	59.8%	10.3%	27.8%	61.9%
Upper stage, extra vocab.						
Active/ intensive sample	-.0014	.0855	.1308	.0238	.0441	.1038
	0.0%	39.5%	60.5%	13.9%	25.7%	60.4%
Active/ extensive sample	-.0004	.0941	.1575	.0109	.0677	.0841
	0.0%	37.4%	62.6%	6.7%	41.6%	51.7%

Table 29

Variance Components and their Relative Contributions, with Students (s), Items (i), and Student and Item Interaction plus Error (si+ e) as Sources of Variation, Textbook 2, Sets A and B

Vocabulary stratum, test and sample type	Set A			Set B		
	s	i	si + e	s	i	si + e
Lower stage Passive	-.2280	.0176	.2938	-.2115	.0825	.3647
	0.0%	5.7%	94.3%	0.0%	18.4%	81.6%
Active/ intensive sample	.0183	.0410	.1261	.0437	.0501	.1532
	9.9%	22.1%	68.0%	17.7%	20.3%	62.0%
Active/ extensive sample	.0129	.0379	.1155	.0246	.0810	.1451
	7.8%	22.8%	69.4%	9.8%	32.3%	57.9%
Upper stage Passive	-.0139	.0989	.1658	-.0165	.0824	.1241
	0.0%	37.4%	62.6%	0.0%	39.9%	60.1%
Active/ intensive sample	.0110	.0978	.1424	.0120	.0846	.1023
	4.4%	38.9%	56.7%	6.0%	42.6%	51.4%
Active/ extensive sample	.0329	.1259	.0907	.0254	.0758	.0774
	13.2%	50.5%	36.3%	14.2%	42.4%	43.4%
Upper st., extra Passive	-.0543	.0587	.2092	.0000	.0107	.0473
	0.0%	21.9%	78.1%	0.1%	18.4%	81.5%
Active/ intensive sample	.0104	.1338	.1126	.0134	.1026	.1088
	4.1%	52.1%	43.8%	6.0%	45.6%	48.4%
Active/ extensive sample	-.0098	.0913	.1395	-.0035	.0419	.1251
	0.0%	39.6%	60.4%	0.0%	25.1%	74.9%

Table 30

Variance Components and their Relative Contributions (%), with Students (s), Items (i), and Students and Item Interaction plus Error (si+ e) as Sources of Variation, Set C

Vocabulary stratum, test and sample type	Textbook 1			Textbook 2		
	s	i	si + e	s	i	si + e
Lower stage Passive	.0165	.0491	.1308	-.0550	.0116	.2294
	8.4%	25.0%	66.6%	0.0%	4.8%	95.2%
Active/ intensive sample	.0227	.0684	.1236	.0165	.0117	.1101
	10.6%	31.9%	57.5%	11.9%	8.5%	79.6%
Active/ extensive sample	.0310	.0442	.1038	.0288	.0591	.1101
	17.3%	24.7%	58.0%	14.6%	29.8%	55.6%
Upper stage Passive	-.0269	.0757	.1077	.0120	.0400	.0519
	0.0%	41.3%	58.7%	11.5%	32.5%	50.0%
Active/ intensive sample	.0117	.0469	.0741	.0055	.0282	.0654
	8.8%	35.4%	55.8%	5.5%	28.5%	66.0%
Active/ extensive sample	-.0044	.0569	.0808	-.0029	.0218	.0385
	0.0%	41.3%	58.7%	0.0%	36.1%	63.9%
Upper stage, extra vocab. Passive	-	-	-	.0487	.0815	.0447
	-	-	-	27.9%	46.6%	25.5%
Active/ intensive sample	.0062	.0031	.0390	.0396	.0943	.1121
	12.7%	6.3%	81.0%	16.1%	38.3%	45.6%

Second, to what extent are some words or word classes inherently more difficult to learn than other words or word classes? For instance, are concrete nouns easier to learn than abstract nouns, and both in turn easier than verbs, adjectives, adverbs, and especially structural words (e.g., conjunctions)? Third, are culturally divergent words harder than culturally convergent words? Such questions can only be raised at this point. Some answers may be forthcoming when the data are subjected to further analyses.

Evaluation of the Implemented Design

This section will focus on evaluating the outcome of the implemented design. Specifically, the standard errors of measurement related to the active intensive sample and the active extensive sample will be compared in order to see if there is an optimal trade-off between the number of students and the number of items. The discussion is based on the data contained in Tables 31 and 32. In addition to standard errors, alpha reliability coefficients and the 95% confidence intervals have also been given. Due to the fact that the person variance component was sometimes negative, the alpha coefficient in those cases is set to zero.

It will be recalled that the idea with the intensive and extensive sample used to estimate the size of active vocabulary was to see what effect the trade-off between the number of items and subjects would have on vocabulary size estimates and on the size of the standard error of measurement. The intensive sample was to have fewer items and more students answering each of them, whereas the extensive sample was to have more items and fewer students. Due to the complexity of the practical implementation of the design, this idea did not work in five out of eight cells for Textbook 1

and in one out of eight cells for Textbook 2. This unintended outcome does not seriously hamper the testing of the original idea. That can be partly done but it can also be tested to what extent a larger item sample combined with a larger student sample decreases the size of the standard error in comparison to a case where there are both fewer items and fewer students.

Taking up this latter point, where the original idea of the intensive and extensive samples was reversed mainly in Set C, the lower stage vocabulary for Textbook 1/Set C had 80 students answering 99 items in the intensive sample and 34 students answering 56 items. As expected, the standard error was smaller for the former (.0280 vs. .0308) and the 95% confidence interval for the mean was smaller (111 vs. 120 words). A similar case exists for the upper stage vocabulary of the same set: 80 students in the intensive sample answered 39 items while 33 students in the extensive sample answered 23 items. Again, as expected, the standard error was smaller for the former (.0361 vs. .0496) and the 95% confidence interval for the mean was was also smaller (58 vs. 78 words). Set C using Textbook 2 had another case of this kind for the upper stage vocabulary: 29 students in the intensive sample answered 48 items while 7 students in the extensive sample answered 36 items. As expected, the standard error was slightly smaller for the former (.0264 vs. .0270) but 95% confidence interval for the mean was slightly larger (110 vs. 108 words). In sum, not surprisingly, measurement is more accurate when there are more students and items than when there are less.

Table 31

Size of the Standard Error (S.E.), Alpha Reliability Coefficient (a), and Range in Words of the 95% Confidence Interval for the Mean, Obtained with Different Samples of Students (s) and Items (i), Textbook 1, Sets A, B and C

Vocabulary stratum, test and sample type	Set A					Set B					Set C				
	s	i	S.E.	a	95%	s	i	S.E.	a	95%	s	i	S.E.	a	95%
Lower stage															
Passive	52	24	.0314	.89	124	51	24	.0523	.93	140	36	22	.0502	.74	198
Active/ intensive	129	59	.0327	.90	130	124	59	.0394	.92	206	80	99	.0280	.95	111
Active/ extensive	56	58	.0353	.83	140	53	58	.0342	.87	133	34	56	.0308	.94	120
Upper stage															
Passive	52	22	.0524	.00	237	51	22	.0345	.78	161	33	12	.0801	.00	127
Active/ intensive	126	69	.0328	.94	151	122	69	.0265	.92	118	80	39	.0361	.86	58
Active/ extensive	55	74	.0301	.94	137	51	74	.0226	.93	105	33	23	.0496	.00	78
Upper stage, extra voc.															
Active/ intensive	126	20	.0668	.00	85	37	20	.0479	.82	61	79	10	.0189	.61	5
Active/ extensive	52	22	.0644	.00	80	51	22	.0545	.72	70	-	-	-	-	-

Table 32

Size of the Standard Error (S.E.), Alpha Reliability Coefficient (a), and Range in Words of the 95% Confidence Interval for the Mean, Obtained with Different Student (s) and Item (i) Samples, Textbook 2, Sets A, B and C

Vocabulary stratum, test and sample type	Set A					Set B					Set C				
	s	i	S.E.	a	95%	s	i	S.E.	a	95%	s	i	S.E.	a	95%
Lower stage															
Passive	17	6	.0544	.00	172	17	6	.1202	.00	382	7	12	.0548	.00	175
Active/ intensive	76	24	.0425	.78	137	67	24	.0484	.87	153	26	30	.0269	.82	88
Active/ extensive	19	48	.0307	.84	99	17	48	.0440	.89	161	14	36	.0506	.90	161
Upper stage															
Passive	19	37	.0531	.00	351	17	37	.0481	.00	318	7	24	.0461	.82	194
Active/ intensive	75	54	.0430	.81	284	67	54	.0402	.86	264	29	48	.0264	.80	110
Active/ extensive	27	98	.0457	.98	305	25	98	.0363	.97	240	7	36	.0270	.00	114
Upper stage, extra voc.															
Passive	17	4	.1307	.00	180	17	4	.0579	.00	61	7	24	.0645	.96	114
Active/ intensive	76	12	.1047	.53	145	63	12	.0926	.60	128	29	6	.1292	.68	228
Active/ extensive	19	24	.0621	.00	86	17	24	.0443	.00	61	-	-	-	-	-

Moving to another unintended design outcome, in which there were almost the same number of items but different numbers of students in the lower stage

vocabulary of Sets A and B using Textbook 1, we note that in Set A there were 129 students answering 59 items in the intensive sample and 52 students answering 58 items in the extensive sample. As expected, the size of standard error is smaller for the intensive sample (.0327 vs. .0353) but the 95% confidence interval for the mean is almost the same (139 vs. 140 words). The result in Set B is unexpected: in spite of having more than twice the amount of students (124 vs. 53), the standard error for the intensive sample is larger (.0394 vs. .0342). The 95% confidence interval for the mean is in line with the standard error outcome (153 vs. 133 words). In sum, it does not always seem to make much of a difference in terms of measurement error if you only increase the number of students.

Turning to the remaining ten cases where the original plan worked out (Textbook 1: upper stage vocabulary stratum, Sets A and B; upper stage extra vocabulary, Set A; Textbook 2: all strata for Sets A and B, and Set C lower stage vocabulary), we note that in eight out these ten cases the standard error and the 95% confidence interval for the mean were smaller for the extensive sample: thus we can conclude that it is usually a good trade-off to have more items and fewer students than fewer items and more students. This is in agreement with what Lord and Novick (1968) note about estimating means by multiple matrix sampling. If there are 36 items and 25,200 students (as there might be in a norming study) with given values for means, variance, etc., dividing the items into 6 subtests with 6 items in each and presenting them to 6 non-overlapping samples of 4,200 examinees would yeild a standard error of .036. If one item is left out to produce 7 subtests of 5 items each (3,600 examinees), the standard error would be .234. If 6 items are left out to make 6 subtests of 5 items each (4,200 examinees), the standard error

would be .610. This drastic increase in the last two standard errors in comparison to the first is due to the failure to administer all 36 items. Even omitting one item can have a very detrimental effect.

In order to estimate in greater detail the effect of item and student sample sizes on the accuracy of measurement, 49 different data sets were analyzed to estimate the generic standard error of the mean. The results are presented in Table 33.

Table 33

Generic Standard Errors of Means in a Vocabulary Test (Decimal Point Omitted)

Number of items	Number of Students										
	5	10	25	50	75	100	150	500	1,000	1,500	2,000
1	2943	2707	2556	2503	2485	2476	2468	2455	2452	2451	2451
5	1393	1253	1161	1129	1118	1112	1106	1099	1097	1097	1096
10	1049	0922	0837	0806	0795	0791	0785	0778	0776	0776	0775
15	0906	0781	0696	0665	0654	0649	0643	0636	0634	0634	0635
20	0825	0700	0613	0581	0570	0564	0559	0551	0549	0549	0548
25	0772	0647	0558	0524	0514	0508	0502	0494	0492	0491	0491
40	0686	0557	0463	0427	0414	0407	0401	0391	0389	0388	0388
50	0654	0523	0426	0388	0375	0368	0361	0351	0349	0348	0348
75	0610	0475	0372	0330	0315	0308	0299	0288	0285	0285	0285
100	0586	0449	0342	0297	0281	0272	0264	0251	0248	0247	0246
125	0572	0433	0322	0276	0258	0249	0239	0225	0222	0221	0221
150	0562	0422	0309	0260	0242	0232	0222	0207	0203	0202	0202
200	0550	0407	0291	0239	0219	0209	0198	0181	0177	0176	0175

The same information is shown in a more concrete form in Figure 3.

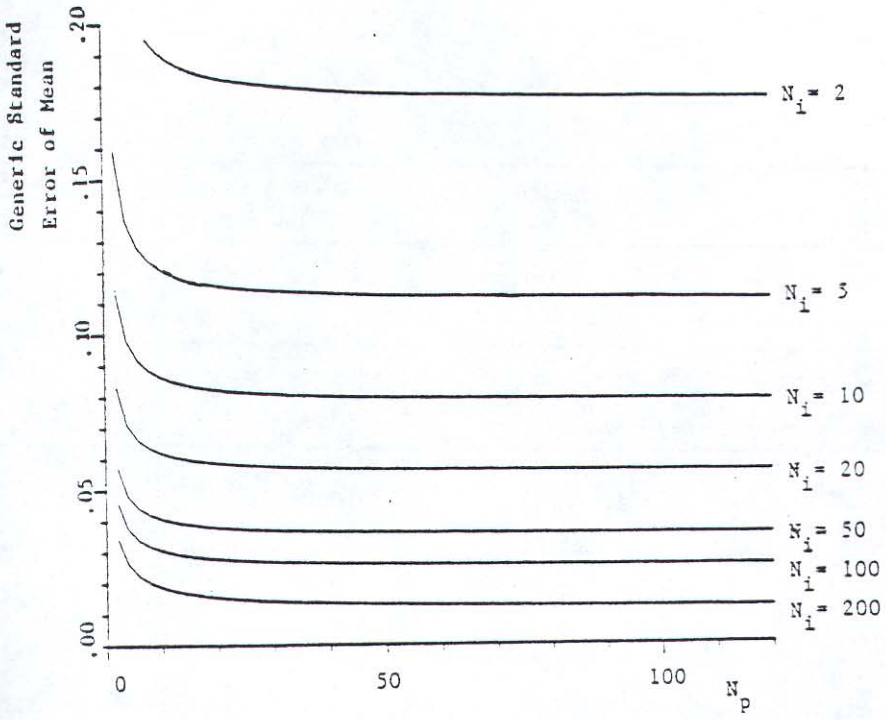


Figure 3. Size of the Generic Standard Error of the Mean as a Function of the Number of Items and Subjects

The reliability of vocabulary tests was measured by computing alpha coefficients in all vocabulary strata for such cases where the test had had 5, 10, 20, 50, and 100 items. The mean alpha coefficient were, respectively, .48, .64, .77, .89, and .94. Thus, about 50 items appear to be needed to obtain adequate reliability.

Another look at the data shows that there are 15 cases out of a total of 49 when the alpha is set at zero. In eleven cases the total number of responses (i.e., the number of students x number of items) is less than 750. More precisely, the outcome is as presented in Table 34.

Table 34

Size of Alpha Coefficient in Relation to the Number of Observations

Number of observations	Alpha = zero	Alpha > 0	Total
> 750	4	29	33
< 750	11	5	16
Total	15	34	49

A X^2 for these data (using Yates' correction for 2 x 2 tables), with 1 degree of freedom, is 13.7, significant at the .001 level. This result is another indication of the size of sample that is needed to guarantee sufficiently reliable measurement.

Another look at the data shows that there are 15 cases out of a total of 49 when the alpha is set at zero. In eleven cases the total number of responses (i.e., the number of students x number of items) is less than 750. More precisely, the outcome is as presented in Table 34.

Table 34

Size of Alpha Coefficient in Relation to the Number of Observations

Number of observations	Alpha = zero	Alpha > 0	Total
> 750	4	29	33
< 750	11	5	16
Total	15	34	49

A X^2 for these data (using Yates' correction for 2 x 2 tables), with 1 degree of freedom, is 13.7, significant at the .001 level. This result is another indication of the size of sample that is needed to guarantee sufficiently reliable measurement.