DIALECT
INTELLIGIBILITY
TESTING

Eugene H. Casad
DIALECT INTELLIGIBILITY TESTING

by
Eugene H. Casad

SUMMER INSTITUTE OF LINGUISTICS
Dallas, Texas
DIALECT
INTELLIGIBILITY
TESTING
CONTENTS

Illustrations ........................................ ix
Tables .................................................. xi
Preface ................................................... xiii
1 INTRODUCTION ......................................... 1
2 OVERVIEW OF A DIALECT SURVEY .................. 3
  2.1 How the survey is organized ...................... 4
  2.1.1 Selecting a language area .................... 4
  2.1.2 Studying background information ............. 4
  2.1.3 Selecting test points ........................ 4
  2.1.4 Preliminary trip to prepare test materials .... 8
  2.2 How the data are collected ...................... 22
  2.2.1 Adapting the test set ........................ 23
  2.2.2 Administering the tests ...................... 24
  2.3 How the data are processed ..................... 29
  2.3.1 Archiving the materials ...................... 29
  2.3.2 Displaying the data in matrices .............. 30
  2.3.3 An optimization model ....................... 36
  2.4 How the conclusions are used ................... 46
  2.4.1 Intelligibility tests distinguish dialect groupings 46
  2.4.2 A display of dialect groups ................ 47
  2.4.3 Feasibility .................................. 49
  2.4.4 Priorities .................................. 50
3 HISTORICAL BACKGROUND .......................... 52
  3.1 Voegelin and Harris ............................. 52
  3.1.1 Olmsted .................................. 53
  3.1.2 Hickerson, Turner, and Hickerson ............ 53
  3.1.3 Pierce ................................... 54
  3.1.4 Biggs ................................... 56
  3.2 Wolff's criticism of the early studies .......... 57
  3.3 Crawford's adaption ............................ 58
  3.3.1 In relation to literacy ...................... 58
  3.3.2 How an index of extendability is derived .... 59
  3.3.3 Expected field conditions ................... 59
  3.3.4 The first test: the Mixe study .................. 60
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2.3</td>
<td>A modal of language maintenance and language shift</td>
<td>128</td>
</tr>
<tr>
<td>8.3</td>
<td>Data collection</td>
<td>130</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Degree of bilingualism</td>
<td>130</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Extralinguistic data</td>
<td>137</td>
</tr>
<tr>
<td>8.3.3</td>
<td>Group behavior toward the languages in contact</td>
<td>140</td>
</tr>
<tr>
<td>8.3.4</td>
<td>Tests for evaluating bilingualism</td>
<td>144</td>
</tr>
<tr>
<td>8.4</td>
<td>Summary</td>
<td>146</td>
</tr>
<tr>
<td>APPENDIXES</td>
<td></td>
<td>149</td>
</tr>
<tr>
<td>A</td>
<td>Sample texts and sets of questions</td>
<td>149</td>
</tr>
<tr>
<td>B</td>
<td>The Trique sentence tests</td>
<td>151</td>
</tr>
<tr>
<td>C</td>
<td>Sample score sheet</td>
<td>155</td>
</tr>
<tr>
<td>D</td>
<td>Standard summary charts</td>
<td>157</td>
</tr>
<tr>
<td>E</td>
<td>Ethnographic questionnaire</td>
<td>159</td>
</tr>
<tr>
<td>F</td>
<td>Summary of studies carried out in Mexico</td>
<td>161</td>
</tr>
<tr>
<td>G</td>
<td>Vowel shifts between the San Lorenzo and Santa Maria dialects of Mazatec</td>
<td>163</td>
</tr>
<tr>
<td>H</td>
<td>Statistical measures</td>
<td>167</td>
</tr>
<tr>
<td>I</td>
<td>Significance testing and Kirk’s reliability study</td>
<td>171</td>
</tr>
<tr>
<td>J</td>
<td>A general theory of intelligibility</td>
<td>185</td>
</tr>
</tbody>
</table>

**BIBLIOGRAPHY** | 195 |
### ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Illustration</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 The Choapan Zapotec Area</td>
<td>7</td>
</tr>
<tr>
<td>2 The Choapan Zapotec Network</td>
<td>43</td>
</tr>
<tr>
<td>3 The Ocotlán Zapotec Network</td>
<td>45</td>
</tr>
<tr>
<td>4 The Mazatec Network</td>
<td>47</td>
</tr>
<tr>
<td>5 The Trique Area</td>
<td>80</td>
</tr>
<tr>
<td>6 The Standard Normal Curve</td>
<td>172</td>
</tr>
<tr>
<td>7 Significance Test with Parametric Measure</td>
<td>182</td>
</tr>
<tr>
<td>8 Variables Underlying Intelligibility</td>
<td>186</td>
</tr>
<tr>
<td>9 Auxiliary and Main Theories: Trique</td>
<td>190</td>
</tr>
</tbody>
</table>
TABLES

1 Raw Scores for Three Samples of Ten: Mazatec Study ........................................ 33
2 Raw Sample Mean Scores: Mazatec Study .............................................................. 34
3 Adjusted Sample Mean Scores: Mazatec Study ...................................................... 35
4 Choapan Optimization Table: Zero State ............................................................... 38
5 Initial Allocation of Resources from x ............................................................... 39
6 Cost Reduction by Quantity Reallocation ......................................................... 41
7 Optimal Allocation: Abbreviated Table .............................................................. 42
8 Optimal Allocation for Ocotlán Zapotec Area ..................................................... 44
9 Percentages of Cognates Between Trique Dialects ............................................ 79
10 Intelligibility scores: Trique Study ................................................................. 80
11 Informant Opinion of Relative Degree of Intelligibility .................................... 84
12 Raw Data Matrices: Choapan Zapotec ............................................................... 93
13 Subject Scores on Sentence Tests ................................................................. 94
14 Differences in Text and Sentence Test Scores Using Chi Square and Wilcoxon T Tests ......................................................... 95
15 The Singular Pronouns of Comaltepec and Xochiapán ...................................... 96
16 Relative Consistency of Subjects’ Errors: Trique Sentence Tests .................... 115
17 Samples of Three and Six Compared with Samples of Ten .............................. 116
18 Language Choice in Diglossia ............................................................................. 136
19 Factors Producing Language Shift in Coastal Popoluca .................................... 139
20 Internal Characteristics of Popoluca Communities ............................................ 139
21 Mean, Median, and Modal Scores: Mazatec ...................................................... 168
22 Computation of Standard Deviation ................................................................ 169
23 Standard Deviation: Raw Score Formula .......................................................... 170
24 Ranges and Standard Deviations: Mazatec ....................................................... 170
25 Confidence Intervals for Mazatec Data ............................................................. 173
26 6 x 2 Contingency Table: Choapan Sentence Tests ........................................ 175
27 Wilcoxon T for Xochiapán Test Scores .............................................................. 176
28 Raw Score Matrix A ........................................................................................... 177
29 Ranked Score Matrix B ....................................................................................... 177
30 Some Critical Values of the Chi Square ............................................................ 178
31 Some Critical Values of the Wilcoxon T .............................................................. 179
32 Some Critical Values of the Kendall W .............................................................. 179
33 Summary of Wilcoxon T Test of Significance .................................................. 183
34 Estimated Variables and Predicted Intelligibility for Trique Data .................. 192
PREFACE

Studies of interdialectal intelligibility have attracted considerable attention since they were first suggested by Voegelin and Harris in 1951. Although these authors seemed to suggest the method primarily as a kind of dialect geography, it has been used mainly to measure dialect distance. In 1959 Hans Wolff questioned the validity of using a measure of intelligibility to determine genetic relationships among languages. He pointed out that intelligibility more appropriately signals societal relationships. On the strength of Wolff’s arguments, John Crawford adapted the method for the dialect survey program of the Summer Institute of Linguistics in Mexico. Preliminary studies were carried out in 1964. Since then, we have applied the method widely and have made many improvements in data collection techniques.

This monograph makes available the methods we use for collecting intelligibility data, the ways in which we treat and interpret these data, and the rationale for the methodology. We hope that these materials will be useful for researchers outside of the circle of SIL as well as to SIL colleagues, many of whom are just beginning their own dialect survey programs.

I have tried to be comprehensive, partly because the subject is complex, and partly to provide something of interest for a broad spectrum of readers. As a result the monograph is something of a conglomerate. However, the chapter divisions provide handy starting and stopping points. To some extent I have tried to keep discussions of techniques separate from those about theoretical issues. Thus the reader who is not interested in theory can avoid most of it by not reading certain chapters. Likewise, the theoretically inclined can generally skip over the sections on techniques. However, although theory and technique are analytically distinct, they are not separate in practice. Some mixture of the two was unavoidable.

The discussions touch on several important but undeveloped topics. For example, what to do with sociological data once it is collected. I hope that this will stimulate the reader to do some of his own independent research. To answer the basic questions that remain will require much more than a one-man effort.

xiii
This monograph itself has not been a one-man effort. The main impetus came from a conference of field workers held in Cuernavaca, Morelos, in April 1967, headed up by John Crawford. Those attending were C. Henry Bradley, Eugene Casad, Joseph E. Grimes, Conrad Hurd, Richard Hyde, Paul Kirk, Peter Landerman, Paul Mellema, Laurie McIntosh, Leo Skinner, Ronald Stoltzfus, and Morris Stubblefield. Paul Wright, of the University of North Dakota, graciously served as consultant to the conference. In addition to Crawford's introductory lecture (summarized briefly in Section 3.3) and individual reports on the Mixe, Mixtec, Chol, Mazatec, Chinantec, and Zapotec surveys, the conference considered questions of data processing, the collection of ethnographic data, informant techniques, recording techniques, reliability, and validity. Thus the topics discussed by the conference provided a principal source for the ideas and content of this monograph.

I am indebted to many of my colleagues for their encouragement and help. Henry Bradley, Sarah Gudschinsky, Bruce and Barbara Hollenbach, and Ronald Stoltzfus all read earlier drafts of the manuscript and made suggestions, many of which I have adopted. Peter Landerman, Mildred Larson, Larry Lyman, Paul Mellema, David Persons, and Stoltzfus have all stimulated my thinking in our too infrequent discussions. In addition, Stoltzfus gave me free access to the manuscripts in the survey files, including some of his own. Bradley allowed me to use his unpublished paper on the Mixtec study. Bruce Hollenbach also made some helpful suggestions about Appendix J. Lawrence Clark was kind enough to permit me to include a summary of his pre-publication version of a paper about Popoluca language shift. Allan Jamieson, Paul Kirk, and Eunice Pike all read and commented on the appendix about Mazatec vowel shifts (Appendix G).

I am grateful to my director, Frank Robbins, for his kindness in letting me operate at my own pace, unhindered by other responsibilities, during eighteen months of research, writing, and revision to bring this manuscript to completion.

My extreme gratitude goes to Joseph E. Grimes for his hard-nosed, thorough critique of a late draft of the entire manuscript. He corrected many of my erroneous statements and cleared up some incoherent arguments as well as opening my eyes to some questions I have not yet answered. The monograph is very much the better for Grimes's capable and kind supervision.

Needless to say, none of my colleagues agrees with everything I say. I assume full responsibility for whatever faults and mistakes the manuscript contains. The major fault may have been the attempt to apply my scant knowledge of statistics to a very complicated problem. My only claim is that I have tried to do my homework well and avoid novel interpretations.

Finally, I am very grateful to my wife, Betty, who has kindly typed the entire manuscript at least three times through (and some sections more often than that) as a result of seemingly countless revisions. She also managed to endure my grumpiness as I tried to think through many difficult sections of the monograph. Finally, she has on numerous occasions corrected my unorthodox grammar.

Eugene H. Casad

Mexico, D. F.
January 18, 1971
1 INTRODUCTION

Nations often face decisions about which language to use in mass communications, educational programs, or vernacular language translations. On the one hand, it may not be feasible to initiate such projects in every one of a country's languages and dialects. On the other hand, since that country needs the cooperation of every group within it, it must carry out its program in languages that are both understood and accepted by all groups concerned. An urgent need, then, is to develop sociolinguistic surveys that are broad enough to do the following: first, define and group languages and dialects rationally; second, specify which of these languages or dialects are acceptable to speakers of other languages or dialects; third, predict directions and probable limits on trends toward dialect levelling; and finally, assign priorities to the languages in which programs will be established.

A team from the Summer Institute of Linguistics has been developing a methodology for sociolinguistic surveys. While it is not the only approach to this kind of survey, we feel that it is the most reasonable one yet proposed and is capable of leading to useful solutions that relate to language use in national or regional programs. It measures intelligibility among dialects or languages in an area. There are circumstances in which the method cannot be applied, e.g., when the investigator neither speaks the language nor is able to find a bilingual informant, or when it is impossible to obtain cooperation in an area. At such times it may be necessary to fall back on less useful approximations like word lists and cognate counts. Nevertheless, since the method is feasible, and since enough data are now available to suggest that intelligibility is a valid measure for the purposes mentioned, we present this monograph in order to show others how to use the method and to encourage additional research into the nature of intelligibility.

In brief, we view any multilingual area as a network of primary and secondary dialects interrelated in such a way that communication from the smallest possible set of primary dialects to all of the secondary ones is adequate. The procedures followed by a survey team, therefore, constitute a strategy for approximating the optimal network for an area.

To begin with, any local or social dialects that can be distinguished from each other are impressionistically divided into likely primary centers of
communication and likely secondary ones. Because intelligibility tests are administered in each dialect, we refer to all the dialects as test points. Of these, the ones we guess might turn out to be primary centers are termed reference points, and the remainder subsidiary points.

During a preliminary visit to each reference point, the team establishes public relations and constructs intelligibility tests. The tests are made from tape recordings of short autobiographical or non-folkloristic texts. A test tape consists of a dubbed copy of a text together with ten questions based on information the text contains. During this trip the team usually finds out more about dialect relationships and revises its initial inventory of test points. It may add or drop test points on the basis of this closer look.

After the preliminary trip, the team specifies test sets of tapes to be played to subjects at each test point. A test set is composed of one introduction tape, a hometown test tape, and four or five reference point tapes. The hometown tape is intended for use only at the subsidiary point where it is constructed; a reference tape is to be used at several subsidiary points as well as at the corresponding reference point.

During the actual survey trip, the team adapts the questions on the test set to the local dialect at each test point. In addition, a hometown test is constructed at any test point that the team has not previously visited. Finally, ten subjects are selected at each test point. They are tested individually through earphones, and their responses are scored as correct, incorrect, or half-correct.

The mean score obtained by subjects from test point A on a test tape from reference point B provides a point estimate of the level at which speakers from dialect A understand speakers from dialect B. To interpret the scores, different levels of intelligibility are treated as possible threshold values. Dialect relationships then are indicated by grouping subsidiary points around the smallest possible number of reference points, using a mathematical criterion of economy. A contour map can be made by grouping dialects for a series of thresholds. The stability of dialect relationships is directly proportional to the number of contours that separate one set of dialects from other sets of dialects. One point within a grouping is the center for communicating to all other points within that group (Grimes, in press).

An index of the maximum extendability of a center is derived from the measure of intelligibility by selecting a particular set of threshold values. A grouping of subsidiary points around the center which is identical for each value in the set of thresholds constitutes an index of the maximal domain of that center. This index, combined with linguistic and sociological information, is the basis for administrative decisions about where to initiate programs in the vernacular.

The rest of this monograph deals with various details of the method. Chapter 2 describes the four stages of the study, from organizing the survey to basing decisions on it. Chapter 3 covers the history of the method. Chapters 4 through 6 concern technical questions of reliability and validity and include a comparison of two kinds of tests and a critical review of intelligibility testing. Chapter 7 briefly considers measures that could possibly be used when intelligibility testing is unfeasible. These measures in turn suggest a method that might be easier to apply than the present one, though equally useful. Chapter 8 discusses how degrees and trends of bilingualism might be measured.
2 OVERVIEW OF A DIALECT STUDY

2.0. This section of the monograph outlines the way dialect surveys are organized and carried out. In general, the discussion corresponds to the chronological sequence of steps a survey team follows when it does an intelligibility study. At several points we describe aspects of both the test and techniques, and discuss the reasons for adopting a particular constraint or technique; it is hoped that in this way the reader may obtain a more unified view of both the theory underlying surveys and the methods used to collect and analyze the survey data. (Even though problems of theory and problems of measurement, i.e., technique, are analytically distinct, they are frequently not methodologically separable.) Alternative viewpoints and techniques are also mentioned.

The survey team consists of two members. One is generally a linguist who already speaks at least one dialect of the language under study. The other is a technician familiar with both the mechanical aspects of preparing the test materials and the administering of the tests. The linguist's outlook is generally language-oriented; the technician's is experiment-oriented. Both need to be familiar with the methodology.

A dialect survey falls into four stages. The first consists of organizing the study. It entails the choice of an area for study, examination of background materials, initial selection of reference points and subsidiary points, and preparation of test materials during a preliminary trip to the reference points. During the second stage, the actual intelligibility data are collected. At each test point the test tapes are modified to the local dialect, subjects are selected, and the tests are administered. The third stage involves handling the materials which have been collected. This includes archiving the test tapes and data sheets and analyzing and interpreting the test results. The final stage of the study is one of decision-making. On the basis of intelligibility, linguistic, and ethnographic data, the number of primary dialectal centers is determined for the area. Priorities are then established for initiating programs that work out from each of these dialectal centers to cover an area linguistically.
2.1. How the Survey is Organized.

2.1.1. Selecting a language area. The problem is not so much one of determining the number and locations of distinct language families as it is of defining patterns of communication between dialects of a known language or language family. Most of the mapping of Mexican linguistic groups has already been done; for example, it is generally known where the boundaries are between the Mixtec, Trique, and Zapotec language families. A survey to determine the number and locations of distinct language families is therefore preliminary to the type of survey discussed in this paper; in fact, it would be premature to attempt intelligibility studies before a language survey has been made. The area for study is more appropriately one in which speakers from different dialects can understand one another with varying success.

2.1.2. Studying background information. After an area has been selected, it is necessary to gain a feel for the area and build an adequate base for carrying out the study. Principal sources of such information include maps, census publications, linguistic and anthropological publications, and the team linguist's own familiarity with the sociological and linguistic scene. Further information can be gathered by talking with people who know the area.

Maps may supply data about the number of towns in the area, the relative distances between them, their positions relative to one another, and the presence of roads and trails. These data may suggest the most economical routes for the trips that will be made through the area during the course of the survey. Census publications or even previous survey reports may give approximate populations for towns in the area, the distinct languages spoken in each town, the number of monolingual speakers of minority languages, and the political status of each town, e.g., district center, incorporated town, or unincorporated village. Linguistic publications may reveal differences between dialects that could affect communication. The linguist himself may be aware of additional differences. He may also know which towns in the area are more important than others and what are some of the economic and political struggles between particular towns. Finally, people living in various towns may be able to contribute data about migrations from older towns. These people may also volunteer their own impressions about the relationships among the various dialects. Although such opinions are colored by each person's history, they usually are relevant to the study and therefore should be written down.

2.1.3. Selecting test points. On the basis of this background information, the survey team makes a tentative selection of reference points and subsidiary points. To begin with, the team must consider the following questions: (1) Given a language area, how many recognizable distinct dialects are there? (2) Of these dialects, which seem to be primary and which secondary? (3) Of the dialects that are probably secondary, which can be assumed to group closely with other dialects if it is necessary to omit them? (4) Which set of test tapes will be administered at each test point?

(1) Complications. There is no clear-cut procedure for selecting test points. In the first place, one must define what is meant by "dialect." Frequently the rural population lives in individual, well-defined, linguistically homogeneous villages or towns, so that one can conveniently refer to a
dialect by the name of the town where it is spoken. (In this study, therefore, when the word "town" occurs, it usually refers to a dialect spoken by the majority of people in a particular community.) In areas where social systems differ from this pattern, however, it may be necessary to define "dialect" in terms of social space.

Certain characteristics of intelligibility also complicate the selection of reference points and test points. Intelligibility can be modeled as a continuously distributed phenomenon, so that the ease with which speakers of dialect B can understand speakers of dialect A is frequently inversely proportional to the distance between A and B. A rare situation which has been reported for a set of Zapotec dialects is where three dialects A, B, and C are in a line geographically, and speakers of dialect A cannot understand speakers of dialect B, but they can understand speakers of dialect C (Morris Stubblefield, personal communication). As the distance increases, the drop in intelligibility may be gradual (though not necessarily linear), not abrupt.

In addition, intelligibility is determined by degree of linguistic similarity and by sociolinguistic factors such as the degree of contact between a pair of language groups. Because of this, three patterns of dialect groupings emerge via intelligibility testing: homogeneous languages, language chains, and language (or dialect) isolates (Naroll 1968). A homogeneous language consists of a set of dialects arranged nontransitively so that A understands B, B understands C, and C understands D in a language consisting of four dialects A, B, C, and D, but it does not necessarily follow that A understands D. Language or dialect isolates consist of sets of geographically contiguous, but mutually non-intelligible, dialects.1 These may be dialects of distinct languages or dialects of the same language.2

The complexity of the problem becomes evident when it is realized that the communication pattern within a given area may be a combination of any pair of these three basic patterns, or of all three. Because of the first two patterns, not all of the dialects will necessarily be included in the study. Because of the third pattern, the investigator must be thorough enough in his preparations that he does not overlook an important dialect.

(2) Criteria. Criteria for selecting test points include the following: the geographic size of the area, the known linguistic divergence between dialects, the populations of the various towns, the estimated linguistic homogeneity of these populations, the estimated degree of bilingualism with

1. For this monograph, the following working definitions are used to distinguish between dialect, dialectal group, and language. Dialects are spoken by distinct speech communities which can freely communicate with one another. They may be described either by an index of linguistic differences or by geographical or social space. Dialectal groups are language groups which cannot effectively communicate with one another. Each dialectal group consists of some sub-set of the dialects of the language in question and contains at least one primary dialect which is intelligible with all the others. Languages, which may consist of one or more dialectal groups, are separated by some (arbitrary) lower level of intelligibility. These definitions, then, are based on the degree of intelligibility between speech communities. For groupings wider than languages, i.e., language family, the intelligibility measure seems inappropriate. Furthermore, for studies with goals different from those of the studies under discussion here, it might be necessary to adopt some other criterion (cf. Haugen 1962.152). Naroll used the concept of intelligibility between speech communities as a means of helping define cultural units (Naroll 1968).

2. This is the case for the set of dialects covered by the Ocotlán Zapotec survey. Such a case may well have developed after the Conquest when bilingualism between dialects of indigenous languages underwent a shift to bilingualism between one dialect of the indigenous language and Spanish.
the national language, and the relative political, economic, or religious dominance of the towns. Finally, towns where some linguistic work has already been done are almost always selected.

(3) Initial hypotheses. The initial array of test points represents the survey team's judicious guess about the dialect network they are to survey, and attempts to cover all known dialect areas. Generally, if the survey team suspects that there are n centers of communication, they try to select at least \( n + 1 \) reference points. The number of subsidiary points they select depends upon the number of towns in the area, the degree of coverage desired, and the amount of time available for the study.

After selecting the reference and subsidiary points, the team specifies which tapes will be used at each test point. It is possible to use all the tapes at each test point, but this becomes unfeasible if more than five or six dialects are included in the study. In the first place, though this method would guarantee an estimate of all the network relationships, it would require too much time to carry out. In addition, the team can expect only a reasonable amount of cooperation from town authorities and subjects. Subjects tire when listening to twenty-six test tapes, especially if they cannot understand twenty-four of them. We limit the test to five or six tapes so that no subject becomes fatigued. Finally, much of the information thus obtained would be redundant. The team thus makes an additional guess about which tapes should be most useful at each test point.

These initial guesses are generally modified as the study progresses and the team learns more about the dialect relationships. The initial set of reference tapes may be increased by adding some "hometown" tapes from subsidiary points, thus converting them to reference points. For example, in the Mazatec study an initial set of four reference tapes was finally expanded to thirteen. Usually, between one-third and one-half of all the test tapes constructed are used as reference tapes in a survey that covers ten or more test points. In addition, the composition of the test set is often changed in the field. A reference tape may be eliminated when the score registered on it at some test point falls below an arbitrary threshold value of, say, 20% or 30%. The decision to modify a test set depends on the level of intelligibility that was last registered on each tape, the interests an investigator has in determining how well a tape will be understood at specific test points, and the expected discouragement a subject might feel in trying to understand too many unintelligible test tapes in a test set. Thus, adequate coverage of an area must be achieved while modifying initial guesses on the basis of trends that are observed as the study progresses.

(4) The Mixtec study. The Mixtec study by Bradley (1967) illustrates how some of these criteria may be applied. There are four characteristics of the Mixtec area. First, it is both geographically extensive and fragmented. Second, the Mixtec population is fairly large and is situated in many towns. Third, the Mixtec dialects are extremely divergent linguistically. Finally, Spanish is increasingly dominant, further complicating the linguistic picture in the Mixtec area.

Bradley divided the area into sections that contained several reference points and then selected test points for each section. Continuity in the study was achieved by using test tapes from adjacent sections in towns near the perimeter of each section. He chose twenty-six reference points for the entire area that had (a) a Mixtec-speaking population of 2,000 or more (unless two such towns were too close geographically), (b) the largest town
in a section, and (c) the political or (d) geographical center of a particular section. Subsidiary points had over 1,000 Mixtec speakers and the subordinate political status of a town (i.e., a municipio, which roughly corresponds to a county seat). A few municipios of less than 1,000 people were included in the survey. Each test point was assumed to represent a dialect that was also spoken in all other towns and ranches under its jurisdiction. Initially, about 100 of these towns were chosen as subsidiary points. However, a few were later omitted on judgments that Spanish was the dominant language there.³

(5) The Choapan study. Somewhat different criteria determined the choice of test points for the Choapan segment of the Zapotec survey (Casad 1969). The background materials included a report by Otis Leal and Dow Robinson on a week-long survey trip through the Choapan district in 1958 (Leal and Robinson 1958). This report contained a list of all the Zapotec towns, their population sizes, and a breakdown of the various linguistic groups represented in each town. It provided the framework for the network of towns set up for the Choapan experiment.

The report noted, for example, that in Santa Cecilia and Boca del Monte not only were groups of Chinantec and Mixe speakers scattered among the Choapan Zapotec, but also there were speakers of other Zapotec dialects, Betaza and Yalalag, living there. The diversity of dialects spoken in these two towns would have made it difficult to obtain samples of ten speakers of the general dialect spoken in the Choapan district. Thus it was felt that it would be uneconomical to test in these two towns. An additional factor in selecting reference and test points was that in most of the towns the

³ This was assumed if there were no children heard talking Mixtec and all the informants agreed that the younger people no longer spoke it. Obviously a more rigorous evaluation is necessary (see Chapter 8).
linguistic divergence from the Choapan dialect was minimal; the investigator was able to communicate easily in Zapotec with the inhabitants. Finally, most towns in the area were small and not widely separated. Thus, some towns were bypassed on the basis of their proximity to others or of their having a population of less than five hundred.

Five dialects were finally selected for test points: Comaltepec, at the south end of the zone, spoken by Lyman; Choapan, the district head town; Jalalhui, formed by a migration from Choapan two hundred years ago; Xochiapan, a linguistic island; and Arenal Grande, at the north end of the zone.

2.1.4. Preliminary trip to prepare test materials. After tentatively selecting test points, the survey team makes a preliminary trip through the language area. The purpose of this trip is to collect texts and to prepare test materials at the reference points. In addition, word lists and ethnographic data are collected, and the team becomes acquainted with the area. Additional information obtained on this preliminary trip is the basis for any revision of the set of reference and subsidiary points.

The work usually begins in the area of the dialect which the linguist speaks. Thus he becomes familiar with the procedures with a minimum risk of incorporating mistakes into the test because of his own lack of familiarity with the local dialect. Later, after he is sufficiently acquainted with the mechanics of the method, he can use his knowledge of one dialect for preparing the test materials in an unfamiliar dialect. In addition, he learns to work efficiently with the technician. When this trip is over, the survey team is ready to begin administering the intelligibility tests at all the test points. In each town visited during the preliminary trip, the work proceeds through three steps. First, the team visits the town authorities in order to solicit their help. Then they select an informant and prepare the test materials. Finally, they collect word lists and ethnographic data. Details, constraints, and technical suggestions on each step follow.

(1) Visit local authorities. The success of the survey is dependent upon the cooperation of the local people. The survey team generally begins by visiting the town authorities, both to gain their confidence and to find informants and subjects. In towns where there is already an investigator from outside, this step may not be particularly difficult, but where there is none it has often been extremely important. The survey team members must contact someone who is sympathetic with their requirements and who also has the influence to solicit cooperation from some of his associates. Since people in the area are often suspicious of the survey team’s intentions, their suspicion must be allayed so that the team can successfully carry out the study.

The survey team carries letters of recommendation from various government and civil agencies, addressed to the civil authorities of the towns or districts in which the study is being done. These letters contain a brief statement of the purpose of the study and some of them carry a photograph of the investigator or technician. They also request that the authorities meet the needs of the survey team.

The sources for these letters have included a department of the federal government, organizations sponsored by the federal government, the office of a state governor, and the president’s office of various district head towns.
The team presents these letters and gives a more complete explanation of the purpose, methods, and needs of the survey team. Specific needs include a place to buy meals, to work, and to sleep; informants to prepare the test tapes; and subjects to take the test. Generally they offer to pay the informants. In this way the survey team formally acknowledges the authority of these town leaders and submits to their authority as a means of getting the needed cooperation from them. Occasionally the authorities are not available, so it becomes necessary to explore other channels, such as school teachers and people to whom the team may have been referred by the linguist's friends. In general, moreover, these have been a main source of informants and subjects for the survey team, and their willingness has been prompted by the survey team's use of the letters of recommendation.

(2) Informants. The informants who prepare the test tapes are native speakers of the local dialect and are reasonably bilingual in Spanish. If possible, only those whose parents are also native speakers of that same dialect are chosen. Other restrictions on the selection of an informant are that he be cooperative, not intoxicated, and free from obvious speech defects. The team also looks for an informant whose voice quality is clear so that they can obtain a high quality in their recordings of the test materials. Characteristics that the team looks for in an informant, but sometimes cannot find, include leadership in his social group, mental alertness, and adeptness in telling stories. Most good informants have been between twenty and fifty years of age, though some younger and some older ones have proved very capable. Informants differ in their abilities, and one who can perform well on one aspect of the preparation of the tapes sometimes proves inadequate for another aspect. For example, an informant who works well in supplying a word list may not do well at translating questions from Spanish into the local dialect. Therefore it has been found helpful to have an additional informant or two available for these other tasks. To summarize: criteria for selecting informants include speech habits representative of the local dialect, status in the local groups, age, cooperativeness, mental alertness, story-telling ability, and voice quality.

(3) Elements of the test set. The elements that enter into a test are presented in the succeeding paragraphs in a logical rather than a strictly chronological order. This is done to help the reader see in detail what a test set is like and how it is constructed. The steps in test construction are the following: the translation of the introduction tape into the local dialect, the elicitation and transcription of an adequate text, the formulation and translation of ten questions on the content of the text, and the submission of the translated materials to a panel of speakers of the dialect to check for naturalness and possible translation errors.

(a) The introduction tape. An introduction tape, translated into the local dialect, is included to give a uniform and sufficient amount of instruction to all subjects tested during a survey. This tape is in two parts: a formal explanation and a sample test. The formal explanation to the subject briefly outlines the purpose and method of the study and instructs the subject as to what is expected of him. The sample test is a short stretch of speech with five questions about its content. It is designed to provide the subject with an opportunity to learn the mechanics of taking the test.

In some early surveys, the sample story was taken from an unused section of a larger text that was elicited for the test tape, or was one
elicited as a separate text. It was approximately thirty seconds in length. In later surveys, however, this sample story was contrived by the investigator himself and then elicited sentence by sentence. It contained only a few short sentences. This method, initiated by Kirk, has the advantage that it is easier to apply in the field situation.

A sample story elicited for the introduction tape is the following:

Three years ago I bought a donkey from my brother-in-law. The donkey cost me three hundred pesos. He was a young male and was very strong. One day he carried my whole family to Santo Tomás. He carried them there in seven hours. I owned him only for a little while. Someone stole him from me. I never saw him again.

A final instruction is recorded on the tape following the sample test. It instructs the subject to listen to some other tapes and to answer questions just as he did for the sample test.

The content and format for the introductory materials is illustrated below:

1. Formal introduction
   Please listen to some stories for us. We are working under the _____ government. We are studying the _____ language. We want to know how well you people understand the words of people from other towns. The machine is going to tell you stories from three or four other towns. Some of these stories you are going to understand well. Others you are not going to understand so well. During each story I am going to ask you some questions. These questions ask about what happened in the story. Listen carefully so that you can answer them. Here is one story.

2. Sample story and questions
   Three years ago I bought a donkey from my brother-in-law.
   Q. 1: When did I buy an animal?
   I bought a donkey from my brother-in-law.
   Q. 2: What kind of an animal did I buy?
   This donkey cost me three hundred pesos.
   Q. 3: How much did the donkey cost me?
   He was a young male and was very strong. One day he carried my whole family to Santo Tomás. He carried them there in seven hours.
   Q. 4: What did the donkey carry to Santo Tomás?
   I owned him only for a little while. Someone stole him from me. I never saw him again.
   Q. 5: What happened to my donkey?

There are several characteristics of the sample story to note. First, the story is segmented so that each question follows the appropriate section on the introduction tape. This format is followed in the test tapes. Second, in this example, the questions are phrased in the first person. This was done to make the elicitation easier, although the informant sometimes phrases the questions in the second person. This latter method is probably preferable because it allows the sample test to be more analogous to the test tapes where the questions are actually phrased in the third person. In addition, if the introduction tape is elicited before the other test materials are prepared, the investigator is able to train the informant somewhat before eliciting the questions based on the test tapes. Finally, the reason that stretches of speech longer than one sentence were included in the sample story preceding questions four and five was to make the sample story conform somewhat to the format of an actual test tape in which a question refers to only one sentence in a connected discourse and the subject must relate the question to the proper statement.
3. Final instructions

You have just answered some questions about a story. Now you will hear
some more stories and I am going to ask you questions about these stories,
too. Shall we begin?

(b) The text. The basis of an intelligibility test is an elicited
autobiographical text that contains sufficient information to permit the
formulation of ten questions over its content. The texts used in various
studies have generally been between two and three-and-a-half minutes
in length, though some as short as one-and-a-half minutes have been
used. Shorter texts may be insufficient in detail for test construction,
and longer texts may be filled with repetitions and irrelevant features
which bore the subject and do not contain content useful for
formulating questions.

Autobiographical material, unpredictable to the subject, is chosen as
the class of content for the texts which are used. This type probably
results in a more representative sample of the language than if
folkloristic texts were chosen, since it is not likely to contain the
specialized style and vocabulary which frequently is used in the latter.
In addition, the subject's prior acquaintance with folkloristic texts like
religious ceremonies or traditions, myths, town histories, or texts about
how to plant crops might invalidate his scores, especially if there is a
high-to-marginal degree of intelligibility between the subject's dialect
and that of the test tape. In these cases, the subject may forget the
testing situation and begin answering from his knowledge of folklore
patterns (Dundes 1963), or culture patterns, or from his own
involvement in the event related in the text. When this happens, the
subject's responses indicate little of his understanding of the other
dialect even though he answers the questions perfectly.

In addition, texts concerning political issues, sermons, or personal
affairs that might prove embarrassing to an individual still in the local
area, are excluded. A subject's reaction to offensive content in a text
can invalidate his scores on that test tape. Morris Stubblefield
demonstrated one such instance from the Zapotec survey (personal
communication). The hometown text was an account of a roadbuilding
project in the area. The scores on that text were about 25% lower than
the normal he had been registering for hometown tapes in the survey.
Later he learned that the town was bitterly divided between the
roadbuilding project and a simultaneous churchbuilding project. It also
turned out that the informant who gave the text and the subjects who
listened to it were on opposite sides of the dispute.

The content of the texts used, then, has generally been restricted to
a particular incident in the life of the informant, or of one of the
members of his family. Frequently informants are reluctant to relate a
personal experience, so that the team must accept a biographical
anecdote which it hopes is not already widely known in the area. The
team attempts to make the informant focus on one particular incident
and elaborate on all the pertinent details such as the persons involved,
the location of the incident, the time of occurrence, who did what, and
why someone did a particular thing. In this way, they hope to obtain a
text with sufficient content for formulating questions about a broad
range of details. Topics thus are things like how the informant broke
his leg, how the house burned down, how his little sister got lost, what
happened to the coffee grinder he just bought, and what happened to

4. This happened in the case of one text that was used in the Mixtec survey. The text
related an incident between the local townspeople and some outside people, in which one
of the local people was killed.
the two cows he put out to pasture. On the other hand, Stubblefield noted that a text of an informant's trip into the city may be too full of loan words to be useful (Stubblefield and Stoltzfus 1967).

The text elicited in the Mazatec town of Mazatzongo de Guerrero illustrates the kind of text that a team must often use. It relates an incident that was probably well known in the immediate area, but since it was used only as a reference point tape at test points outside of the immediate area, Kirk considered it to be an adequate text (Kirk 1967).

Three persons came from the town of Huehuetlán. They were looking for work in a place called “Cashapa.” But, unfortunately, it was raining on this particular day. When they arrived at a guila called “Gucamaya,” it was flooded where they had to cross. The front man was carrying a goatskin of liquor. When he was in the middle of the guila, he tripped and fell into the river. The tumpline which he was using to carry his cargo wound about his neck and he drowned. Then his companions returned to their home town. They did not say anything about what had happened in our town; they did not mention that their companion had been drowned in the river. They went straight to their own country.

The corpse was not found until the third day. The goatskin with liquor was still good when taken from the water. Those that went to bury the man drank this liquor. The net carrying-bag and the tumpline were picked up by a man named Márcelo. He took them because he was poor and the dead man would not need these things. Márcelo worked for a man named Gabriel. On this particular day he went to pick coffee. Shortly after he started to work, he felt a cold hand on his arm. Márcelo was scared of this person. He went out, got the things of the dead man, and put them on the man's grave.

The following list shows one set of ten questions that can be made up from the content of this text:

1. How many people came?
2. Why were they going to Cashapa?
3. What was the weather like?
4. What was it like where they had to cross?
5. What was the man carrying?
6. What happened in the middle of the guila?
7. What did the man's friends do after he drowned?
8. Who drank the liquor?
9. Why did Márcelo take the things?
10. Where did Márcelo put the things?

(c) The questions. The questions formulated from the content of a text are translated into the local dialect and embedded at appropriate places within the body of a copy of the original tape recording to form the test tape.5

These questions are directed at sampling the widest possible range of semantic systems found in the text. To ensure an adequate sampling of the grammar and lexicon within the text, a concentration of questions on any one semantic category such as color, kinship, or number is ruled out. If possible, not more than one question regarding each of these categories is contained in any one set of questions. Instead, we try to formulate questions so that they cover as many of the following categories as the content of a text allows: participant as actor, participant as goal, object as goal, specific event, ground or explanation of event, purpose, cause, result, quotative, description of

5. They were formerly given at the end. For a discussion of the shift in presentation, see Section 3.4.5.
condition, description of character, time, location, manner, instrument, and quantity.

A subject's responses, then, are elicited with information questions (as opposed to yes-no questions) that can be answered in a few words. In order to control the uncertainty regarding what he may be reacting to, these questions are made as specific as possible. In other words, rather than ask, "What did he kill?" the question is phrased, "What animal did the man kill?" Questions are avoided that would require a subject to select one item from a series in the same statement or force him to choose between several similar statements which may all be appropriate answers to the same question. If these kinds of questions must be included, potential problems are minimized by either phrasing the question very specifically or allowing leeway for ambiguities in scoring the subject's response.

The questions are formulated so that they seldom refer to statements in the text that contain words borrowed from the national language. This is because it has been felt that loan words from the national language would systematically bias scores toward higher intelligibility for subjects who are bilingual with the national language but are unable to understand other dialects of the vernacular. Since most subjects have been bilingual, it has been felt that this bias would affect the intelligibility scores considerably. Our position is ambivalent. On the one hand, loan words are often an important aspect of the linguistic system of a dialect and thus constitute a natural part of the communication that is being measured by the intelligibility test. Furthermore, it has been demonstrated that subjects are not always able to answer correctly a question that is answered by a statement in the text containing a loan word. The form of the question frequently determines the centrality of the loan word to the total amount of information which constitutes a correct answer. For example, suppose that the text contains a statement, "The man shot the rabbit," and the word "rabbit" is a loan word. If the question is phrased as, "What animal did the man shoot?" the correct answer will consist solely of a loan word. If the question is phrased as, "What happened next?" then the correct answer refers to an actor and an activity, neither of which are loan words. In this case, a subject probably would not answer the question correctly unless he understood the context the loan word occurs in. If the test were to include many questions that were completely answered by a loan word, then there would be considerable bias in the scores. If the questions referred to entire phrases, each containing only one loan, then systematic bias would probably be negligible. (In some cases there have been parallel statements in the text that did not contain a loan word and these statements were paired with the question of doubtful status.) On the other hand, in many cases of loan words there are perfectly natural equivalents in an informant's dialect. In such instances, it is fairly common that at the beginning of his work with an investigator an informant will use primarily loan words at these optional junctures and later, as his confidence with the investigator grows, he will change to using native words for the same options. The uneasiness with regard to loan words, then, seems not to be a feeling that they are an unnatural part of the linguistic system, but rather that the

6. Herman Aschmann notes a Totonac informant who gave loan words with a very high frequency at the beginning of the elicitation of a word list but began tapering off until he was giving practically none at the end of three days of work (personal communication).
frequency with which they appear in the elicited materials may be unnatural.

Questions that can be answered simply by a "yes" or "no" are also excluded. This is because a subject has a 50% chance of guessing the correct answer, whether or not he can understand the outside dialect. In addition, cases where two different questions in the same test have identical answers (e.g., the word "yellow") are avoided. In such a case, the range of semantic systems sampled is reduced from its already limited scope. Furthermore, in this case, the probability is increased that a potential subject might learn an answer merely by observing the testing and correctly associate it with a given question.

Finally, the subject is not required to make deductions from the text. Such questions would introduce the additional factors of differential I.Q. and educational background into the subjects' performance. The questions are always linked to a statement that explicitly describes an object, event, or situation. Sometimes a particular phrase implies something that is not explicitly stated in the text. If a question that refers to this kind of statement is used, it is generally phrased to refer to the information which is explicitly stated, rather than that which is implied.

It is possible for an investigator to bias the test toward high intelligibility by including questions over items in the texts that he knows will be widely understood within the area. Conversely, he can bias the test toward low intelligibility by making questions about items he knows will not be understood outside of local areas. The questions should be formulated so that items of limited, or widespread, intelligibility are neither purposely excluded from the test nor included in it. They should be formulated in an unbiased fashion rather than by the particular preference of the linguist. In this way, errors caused by the inclusion of both classes of items tend to cancel one another out in the scoring. Since the technician usually knows nothing of the language being studied, he cannot purposely frame questions toward these items. His ignorance of the language is the best assurance that the formulation of the questions will be unbiased. Therefore the technician makes up the initial set of questions for each test tape. The linguist then phrases these questions in the standard language for eliciting their equivalents at the given test points. For example, he selects a standard form in Spanish which he uses for eliciting the corresponding question in the dialects of the various test points where the tape is used. This standard form may then be adapted to regionalisms in Spanish.

Another possible source of bias in the test is controlled by translating all the questions on the test tapes into the dialect spoken at each test point. Otherwise, the subject's score would be influenced by his knowledge of whatever outside language or dialect was used for asking the information on the content of the test tapes. This constraint further enables the test to be given to subjects at the test point who are not bilingual in either Spanish or another dialect of the minority language. That is, by translating the questions into the local dialect, a larger range of potential subjects is assured for testing.

The test format used in the present procedure, i.e., questions embedded within the body of the text, is the result of changes made in earlier models during successive studies.\(^7\) Some considerations

---

7. For a discussion of earlier test formats used in Mexico, see Section 3.4.
motivating these changes were ease in teaching the subject how to take the test, adverse effects on the subject's performance due to distraction or lapse of memory, and ease in constructing the tests. Other factors were differential attention span, I.Q., or educational background of the subjects.

There are several criteria for determining what is an "appropriate" place for inserting a recorded question within the text. First, a question is located following as closely as possible the statement which contains the answer to that question; otherwise, a subject may forget just what was said as he tries to choose the correct information from among all that he heard. Second, in order to minimize the disruption of the text, natural phonological breaks in the text are utilized whenever possible. Third, the statement that is chosen must contain the word or phrase that directly answers the associated question. Fourth, if possible the selected statement should contain neither slurred nor extra rapid speech nor loan words. Finally, the additional suggestion has been made that, if possible, the questions should be evenly dispersed throughout the body of the text rather than clustered around one portion of the text.

(4) Pre-test "panel". After the team has prepared the test materials, it submits them to an informal "panel" of speakers who judge the quality of the translated portions. This panel of at least two members may consist of people who happen to be watching the team prepare the materials, or it may even consist of town authorities. Generally a town official helps the team get the panel together.

The role of the panel supplements that of the survey team. To begin with, the team attempts beforehand to secure an adequate informant, a usable text, and acceptable translations of the introduction tape and the questions for the test tapes. Then they submit these materials to the members of the panel. Whenever possible the team works individually with the panel members. This is because it frequently happens that, when the team works with all of the panel together, one member of the panel turns out to be dominant and the other members tend to concur with all of his judgments, even incorrect ones.

There are two classes of information that may be elicited from the individual members of this group. The first consists of responses to questions about the informant's speech style and text after the text has been played. The questions asked might be as follows: "Does this person speak like the rest of the people in this town?" "If he does not, then what does he say differently?" "Does he speak too rapidly for you to understand what he says?" The other information, then, consists in obtaining responses to the translated questions for each of the test tapes. For each question played, the pre-test subject might be asked, "What does this mean in Spanish?" "Which of these words do you not hear clearly?" "How could we say this better in your dialect?" With an indication of a poorly translated or poorly recorded question, a new recording of that question can be made using one of these panel members.

The panel may be used during both the preliminary and the data-collecting trip. During the former, the panel judges the content and quality of the text and questions for the reference point test tape. The procedure consists of playing the test tape copy of the hometown text on one recorder and the master copy of the questions on a second machine to three informants. The text is played in segments to each of these people.
individually and each is asked about the content and clarity of each statement that answers one of the questions in the text. He is also asked about the content and form of the corresponding questions. Finally he hears each key statement from the text with its proper question and is asked to answer the question on the basis of the statement. These responses may then be recorded. The entire procedure may require fifteen to twenty minutes for each panel member.

The same procedure may be followed with each of these speakers for any of the pertinent reference point tapes that may already have been constructed. In this way, the quality of the locally translated materials can be checked and the panel members can be used as subjects for some of the test tapes. Thus a small part of the testing can be started during the preliminary trip.

Generally, however, the pre-test panel is not used to judge the quality of the translated questions for the reference point test tapes until the data-collecting trip when the team is fairly certain which set of reference point test tapes will be used at a particular test point. It seems easier to prepare all the reference point test tape materials, check these translated materials with control speakers, and administer all the tests during one period of time. Otherwise there is the problem that some subjects will not have received the complete set of test tapes. It is usually easier to simply test ten subjects on the complete set of tapes rather than to test seven on the entire set and six additional subjects each on one-half of the set of tapes.

Finally, the extent to which a survey team uses a panel varies because of several circumstances. If the linguist is able to understand well the dialect spoken at a test point, he may feel that he can adequately control the quality of the translated materials and therefore need not rely on a panel at all. On the other hand, if he cannot understand that dialect, he will rely heavily on a panel. If the linguist cannot understand the dialect and has doubts about the informant's ability, he tends to rely on a panel more than if he is satisfied with the informant's work. Other factors that influence a team's reliance on a panel are time schedules and the amount of cooperation that seems available in a given town. In practice, then, a panel has not been used consistently as a result of field conditions and time schedules.

(5) Order of preparation. There is no strict order imposed on the preparation of the test materials. On the one hand, the investigator might begin by recording the word list. This gives him an initial acquaintance with the phonetics of the new dialect and also may alert him to grammatical changes and lexical shifts between the new dialect and the one he already speaks. This new information may in turn prove valuable for him in evaluating the content and naturalness of the questions elicited in this unfamiliar dialect. In addition, the time spent in working on the word list may accustom the informant to the use of a tape recorder so that his fear of the machine does not reflect itself in his voice quality when he records the questions and the text. On the other hand, since the word list may be obtained piecemeal from a number of different informants, it is often left until spare moments between the testing of one subject and the next or else until the testing is completed.

While several informants may participate in the recording of the word list, generally only one participates in the translation and recording of the
introduction tape and questions for the test tapes. Frequently this same informant also supplies the text for that test point and helps the investigator transcribe it.

The order for translating the questions for the reference point tapes, eliciting the test point text, and translating the introduction tape is also flexible. Some investigators have preferred to elicit the hometown text first because they consider this the most important or most difficult detail in test preparation. Others, however, have elected to begin by translating the questions on the reference point tapes. This allows them time to gain the informant’s confidence and gives him an opportunity to think about a text he could provide for the study. It usually turns out that the introduction tape is the last detail in the preparation of the tapes because it is relatively rapid and simple to elicit. It should be noted, however, that a poorly translated introduction tape has complicated the administration of the tests to subjects; therefore the introduction tape is as important to the test design as the other elements.

(a) Translating. The first problem in translating the questions for the reference point test tapes into the local dialect is to make sure that the informant clearly understands what is required of him. It sometimes proves difficult, for example, to convince an informant that he is supposed to supply a corresponding question in his dialect rather than to answer the one he heard in Spanish. Occasionally an informant may shift the utterance from the form of a question into a statement. For example, instead of offering the equivalent of “Why did these men leave the house?” the informant may say, “These men left the house.” Two other frequent problems are pronoun shifts from third person to first or second person actor and word-for-word translations from Spanish. Other problems may occur if the investigator does not know how to phrase the question properly for eliciting its equivalent.

Common strategems for detecting and meeting these problems include phrasing the question in Spanish several different ways, and using transcriptions of the same questions in other dialects as an analogy. The investigator’s familiarity with one dialect of the language often proves helpful; information obtained from the word list may provide additional clues. Furthermore, the basic form of the question used for elicitation may be determined by regionalisms in the national language, e.g., the influence of Mazatec on “campesino Spanish.” Finally, the pre-test panel should notice any errors that may have escaped the investigator’s scrutiny. If these also get by the panel, they are often detected when the testing begins. For example, if the first three subjects incorrectly answer the same question on a hometown test tape, the linguist is usually justified in assuming that something in the test needs to be changed. At this point the team stops testing until it can determine what the problem is and correct it.

(b) Recording the text. The survey team generally begins to prepare to elicit a hometown text quite a while before the recording session starts. The preparation opens with an explanation to the local authorities who may select the informant. Once the team has obtained an informant, they fully explain to him the kind of text they would like to record. They also may play him a text recorded elsewhere, or they may retell a text. At least they attempt to offer him enough ideas to stimulate his thinking. Frequently he is given some time to think about a text while he and the investigator work on something else, such as the word list. Then, when the informant is ready to give a text, he is
asked to repeat it in Spanish so that the survey team can have some idea of its content. On the basis of this knowledge, the investigator or the technician occasionally suggests additional details that could be included. Sometimes he is requested to give a different text.

After the text has been recorded, it is checked for length and content. If it is short (less than two minutes) and lacks sufficient details for the formation of at least ten questions, additional details are suggested to the informant; this additional information is then recorded directly on the end of the text that was first recorded. Sometimes, however, additional information cannot be simply appended without breaking the continuity of the text. In this case, the entire text is recorded over again in an attempt to obtain a longer, more detailed version. The original recording, though not used for testing, is preserved for eventual archiving. Occasionally, the first text is not sufficient and the only alternative is to record a completely different one. If the text is long (four minutes or more), the content is checked and only the first three minutes (or best three-minute portion) is used for the test tape.

(c) Developing a set of questions. To develop a set of questions for a test tape, the technician carefully studies the translation of the text. On the basis of the content of the translation, he draws up an initial list of twelve to fourteen questions if there is sufficient content to allow this. To decide which questions should be selected he checks the list of answers to them to see how many semantic categories (such as those mentioned on pp.12-13) are represented and how many times any one category appears. The team also listens to the tape to see which key statements are clearly enunciated and which are not. Finally, the statements are checked to see if they contain loan words. The initial list of questions is then narrowed down to ten by eliminating those whose answers (1) duplicate the semantic category represented by the answer to another question, (2) are not clearly enunciated, or (3) contain loan words. Though the technician draws up the questions, the linguist often helps him by providing him with information about the content and quality of each statement. Sometimes the team consults with the informant about particular statements in the text. Note that the amount and kind of content in a text largely determines the kinds of questions that can be asked about it; it is frequently necessary to duplicate semantic categories and to use questions requiring loan words in their answers.

(d) Copies of materials. The team keeps copies of all test materials for use in further test construction and eventual archiving. The written materials include word-for-word transcriptions of the introduction tape and the questions for the test set in the vernacular. These transcriptions are made prior to the recording of the materials. They are often used to help the informant remember what he is supposed to say into the microphone. Either the investigator can read the phrases to the informant as a reminder, or, if the informant is literate, he himself can read the utterances into the microphone.

The linguist uses the transcriptions of the questions at succeeding test points to suggest to the informant a possible way to express a question in his local dialect. The written materials also include a transcription of the text and a translation of it into the national language. In addition, the recorded materials are also kept. The original recordings of the texts are used to construct test tapes; they are also used to make additional copies in case a test tape is lost.
(6) **Technical aspects of test preparation.** Technical procedures that figure in the preparation of the test materials relate to recording these materials on tape and constructing the test tapes.

(a) **Routine procedures.** In the first place, the technician is responsible for securing the highest quality recording that both his recorders and outside circumstances allow. Routine tasks include cleaning the record and erase heads frequently and making sure that the batteries in each recorder are sufficiently charged for use when electric power is not available. In addition he must try to obtain recordings without including background noises such as donkeys braying and people talking. In some cases he can wait until the disturbance passes; in others he can shut out the noise by putting the microphone close to the informant's mouth (two or three inches), and turning the recording level down so that when the informant speaks at a normal conversational level the input signal is not over-recorded. Frequently the recording is done while there are onlookers; it is sometimes necessary to politely ask them to remain silent so that no outside noise enters into the recording.

(b) **Coordinating the participants.** To initiate the recording of each element of the test materials, the technician coordinates the linguist and the informant. In this procedure the linguist briefly instructs the informant on what he is supposed to say. The technician then starts the recorder operating and gives a pre-arranged signal to the informant, who responds according to the linguist's instructions.

For recording the questions, the linguist may request the informant to give two utterances of each question when he sees the technician's signal. When he sees the signal, he says each question twice, leaving a short pause before the second utterance. The first utterance enables the technician to judge the approximate recording level to use, the pause gives the technician time to start the recorder so that the second utterance can be recorded. A problem with this approach is that often the informant's repetition of the question is either louder or softer than his first utterance. The technician has to be alert for this in order to adjust the recording level unless the recorder he is using has an automatic volume control. In an alternate procedure, the linguist may request the informant to utter each question once; the technician then records that utterance. In either approach, the technician must be careful to have the recorder operating when the informant begins to speak. This is to make sure that each question is recorded in its entirety. He should also check the quality of each recorded question while the informant is still available so that, if necessary, he can record a question again.

When the technician records the introduction tape, he again coordinates the investigator and the informant. This time the informant is expected to utter each phrase only once. This change in procedure sometimes confuses him, especially when the recording of the introduction tape immediately follows the elicitation of the questions; it is necessary for the survey team to keep the informant aware of what they expect him to do by stating their instructions clearly to him and repeating them if he still does not seem to understand what to do. The introduction, then, is recorded phrase by phrase, in contrast to the text, which is recorded as a continuous stretch of speech.

(c) **Preparing to edit tapes.** The technician records the questions on the reference point tapes onto a working tape so that he can later insert them into the corresponding test tape. For his own convenience
in locating them, he prefaces each set of questions on the tape with a statement such as "Questions from ______ dialect on the test tape from ______ dialect." He also might preface each question in the set by its number, e.g., "Question no. 1," etc. He should use the national language when recording these prefatory statements, especially if there are bystanders who may be curious about what he is saying into the machine.

(d) **Constructing test tapes.** To construct a test tape the technician first copies a particular text, or a three-minute portion of that text, from a master recording. This is done with two tape recorders connected by a dubbing cord. One, the output recorder, takes the master reel and is set for playback; the other, the input recorder, takes a reel of empty tape and is set for recording. Some machines must be set differently for input from another recorder than for input from a microphone. It is also necessary to adjust the recording level on the input recorder. The way in which this is done depends on the recorders being used: in some cases the strength of the input signal can be controlled by adjusting the volume control on the output recorder and the recording level control on the input recorder; in others, the volume control on the output recorder is independent of its output channel so that the recording level control on the input recorder is the technician's only means of regulating the strength of the input signal. The technician must regulate the input signal so that on playback the copy is easily heard and the quality of the copy does not contain distortion because the input signal was too strong, or motor noise and static from the amplification system because the input signal was too weak. (These conditions assume that all the connections between the dubbing cord and the recorders are properly made.)

When the recorders are properly set up, then, the technician starts the input recorder running a second or two before he turns on the output recorder. The speed used for recording the copy (and usually the master text) is three and three-fourths inches per second (ips). During the copying process, he stops both recorders several times; he stops the output recorder so that he does not lose his place and omit or duplicate a section of the text when he resumes dubbing, and he stops the input recorder so that he can leave space on the tape for subsequent insertion of the questions at the proper places within the copy of the text. The spaces are from two to eight seconds in length, depending upon the method he uses to insert the questions.

The technician uses one of two methods to insert the questions into the test tape. On the one hand, he may dub the individual questions from the working tape onto the appropriate stretches of blank tape within the test copy. These stretches of blank tape, left when the copy was made, are about eight seconds in length—sufficient to allow for the question, one second of silence preceding it on the tape, and three to five seconds of silence following it on the tape. The stretch following the question is left to allow the technician to insert slightly different versions of that question into the test tape without erasing the part of the text that follows as he modifies that tape for use at additional test points. This method has the advantage that it allows one to retain copies of all the sets of questions used in the study; these may be archived for later reference. It has the disadvantage that, since the borders of the segments of tape containing the questions are not clearly marked, the technician has no rapid way to locate each question and no exact way to observe how large his margin of safety is during subsequent modifications of the tape. It should be noted that
these problems can be adequately handled simply by marking each end of the stretch of blank tape with a piece of splicing tape placed across the glossy side, not the side that faces the recording head. It is also feasible to write a number on the splicing tape to indicate which question is to be dubbed into that space.

The other way to insert questions into the test tape is to cut apart the working copy and splice each one in at the proper place. In this method, then, the technician leaves large enough stretches of blank tape within the text to allow for (1) some loss of tape cut off in making the two splices, (2) the inclusion of a short stretch of blank tape preceding the question on the tape to orient the subject, and (3) a stretch of blank tape following the question to serve as a "safety zone" during editing for use at subsequent test points. This method has the disadvantage that most of the recorded sets of questions will be erased during the course of the survey and thus be unavailable for archiving. (However, word-for-word transcriptions of these will be kept.) In addition, it takes more time to construct test tapes this way. The splices do provide a natural and an exact measure of the amount of tape which can be safely used to record a new set of questions when the tapes are being modified at succeeding test points. They also aid in locating each question, but if the technician uses dummy pieces of splicing tape with the other method, he achieves all this with less time spent in editing the tapes.

Regardless of the method used, the linguist usually shows the technician exactly where to segment the text to insert each question. In both methods, as already noted, the technician leaves stretches of blank tape both preceding and following each question. The section preceding the question is designed to orient the subject; he periodically hears a brief pause of about one second which alerts him to the fact that he will be asked a question about the text. If the pause is too short, the subject may be unable to shift his focus soon enough to answer and it may be necessary to repeat the question. If the pause is too long, he may forget what the text said. This could require a replay of both the question and the relevant part of the text. The blank section following the question is needed because either the form of that question or the informants' rates of speech may vary significantly as it is retranslated at succeeding test points. If the length of the pause varies too much, a following part of the text might be erased.

(7) Work pattern of the team. The linguist and the technician must work together flexibly. One reason for beginning field work in the dialect most familiar to the linguist is to allow him and the technician to organize themselves into a fairly natural routine at a minimum expense of time and quality in the test results. Generally, the linguist is responsible for working with informants and subjects and the technician is responsible for working with the tape recorders. Since there is no strict division of labor (especially when both team members can speak a dialect of the language being surveyed or can adequately communicate with informants and subjects in the national language) the actual work pattern is arranged by common consent between the two members of the team.

For certain tasks the team must decide upon an efficient routine which may ignore the usual distinction between the linguist's work and the technician's. For example, when the text is being transcribed, the technician may play it in short stretches for the investigator and informant, replaying certain portions when necessary. On the other hand, the investigator may work more efficiently operating the recorder himself, since he is probably a
better judge of how long a stretch of speech to play and how far to rewind
the machine in order to replay a particular section. As to the word list, the
technician may act as coordinator as he does in recording questions for the
test tapes. It is also possible that he could record the word list himself
while the linguist is working on something else, for instance the elicitation
of the questions. The linguist may, on the other hand, record the word list
while the technician is gathering ethnographic data. The particular routine
selected is merely one that seems most convenient to the team members.

(6) Linguistic and ethnographic data. Linguistic and ethnographic data
comprise the second and third aspects of the methodology originally
outlined by Crawford (see Section 3.3). The linguistic data may be used for
partially validating the intelligibility results by revealing linguistic similarities
that contribute to high intelligibility between a pair of dialects or by
pointing to specific linguistic changes that hamper intelligibility between
those dialects. In addition, these data may suggest linguistic features that
are common to several dialects and therefore can be incorporated into
written materials or adapted for use in mass communication. (For an
example of how some linguistic data may be incorporated into a test for
estimating the range within which distinct dialects can use a common
orthography, see the discussion on p. 87.) The ethnographic data may
point to limits on the extendability by revealing situations and attitudes that
would favor or hinder the use of a particular dialect in a program.

Word lists of three to five hundred items are collected at each of the
test points. These are generally written down and recorded on tape. Partial
verb paradigms and other grammatical data are often elicited with the word
list. Sufficient data are elicited to provide the investigator with a sampling
of equivalent grammatical and semantic subsystems from the dialects
included in the study. These materials may form the basis for comparative
studies.

An ethnographic questionnaire adapted from one prepared by Laurie
McIntosh is being used in the surveys (see Appendix E). Information
obtained by the questionnaire includes geographical notes, the political
structure and status of each town, features of the local economy, language
use, and contacts with outside towns. General information is noted on one
sheet, and specific data are noted on a separate sheet for each person
interviewed. An investigator should be discrete both in asking for certain
kinds of information such as jealousies between towns and in writing these
data down. He may decide to commit the information to memory (if he is
even able to obtain it), so that he can record it on the questionnaire later
on, especially if he finds that people are reluctant to answer his questions.
In other areas he may find that he can easily obtain a sample of opinions
and statements.

To summarize: the first stage of a dialect survey includes a preliminary
trip through the language area defined for study. The purpose of the trip is
to collect texts and prepare reference point test materials. Linguistic and
ethnographic data may also be collected. The work accomplished on this
trip prepares the survey team for the second trip, during which the actual
intelligibility tests are administered to a sample of the inhabitants of each
test point.

2.2. How the Data are Collected. The second stage of a dialect survey
involves the collection of intelligibility data at all reference points and
subsidiary points. At each test point, the following steps are carried out: (1)
the test set is adapted to the local dialect, (2) the subjects are selected for testing, and (3) the intelligibility tests are administered to these subjects.

2.2.1. Adapting the test set. To adapt the test set for use at each test point, it is necessary to make a final decision as to which reference tapes to use. The questions for these test tapes are then translated into the local dialect and recorded in the right place on each tape. An introduction tape is also constructed. In the case of test points not visited previously by the survey team, it is necessary also to elicit a text and to construct a hometown test tape from it.

(1) Final specification of test set. It was noted earlier (in Section 2.1.3) that the survey team makes an initial specification of the test set for each test point based on preliminary data and that this guess is usually modified as the team collects more data during the preliminary trip. The final specification of the test set is made at each point before the tests are prepared. It is based on trends observed as the tests are administered.

A pretest has sometimes been used to ensure an adequate selection of test tapes. If the team does not know whether to include a particular reference tape in the test set, they can play it to three persons, score their responses, and then either include or exclude it for use at that test point on the basis of the average score registered by these subjects. The upper limit for excluding the tape is 95% or above. The lower limit might be 30% or lower. The purpose for specifying these limits is to avoid overloading the test set with tapes of either very high or very low intelligibility. In the pretest, the linguist may ask the questions in the national language rather than in the local dialect; he stops the recorder at various places in the text and asks the subject what the text said. He may ask ten or more questions on the content of this text. The method is merely a rule-of-thumb for deciding how to make up a test set.

(2) Technical modifications. The reference point tapes are modified at a test point by translating the questions for these tapes into the dialect of that test point. The questions are then inserted into the respective test tapes. This can be done by first recording all of the sets of questions on a master tape and later dubbing them onto the test tapes (as described on p. 20) or by having an informant record the questions directly onto the tape. This alternative is a somewhat faster method, but it results in some sets of questions being erased and made unavailable for archiving.

The introduction tape is also translated and recorded in the local dialect. As noted previously, this adapts the method so that monolinguals can be tested. In addition, a subject's scores are not biased as they might be if he failed to understand the instructions and questions because they were given in a dialect other than his own.

At the subsidiary points previously not visited by the survey team a text is elicited so that a hometown test tape can be constructed. A detailed description of how to elicit a text and construct a test tape has been given earlier (see Section 2.1.4).

Constructing a hometown test tape and modifying the reference point tapes requires three to four hours for an experienced team. The team is then ready to select subjects and begin administering the intelligibility tests.

2.2.2. Administering the tests.

(1) Overview of the administration of the tests. The test set consists of the following tapes: first, an introduction tape; next, a test tape from the
subject's own dialect; and finally, four or five reference point test tapes. About forty-five minutes are required for administering the test set to a subject.

The administration of the test is completely in the local dialect. No subject is required to respond in another language, though some prefer to do so. To control distractions the tests are administered through earphones. This also prevents bystanders who are potential subjects from becoming acquainted with the test materials beforehand. The reference tapes are rotated among the successive subjects so that each tape is presented in one ordinal position in the sequence just as many times as it is presented in each of the other positions from first to last. This is done to increase the probability that factors of learning and fatigue will have a comparable effect on the mean scores registered for all of the test tapes and decrease the likelihood that the mean score on one test tape will be greatly affected simply because it always happened to occur in the same ordinal position in the test set.

This rotation of test tapes also helps to confound bystanders who try to remember the subject's responses when their turn comes. The subject's responses are evaluated by the tester and the evaluations are entered on score sheets. It is sometimes necessary to use an interpreter to judge a subject's responses. Bilingual subjects frequently are their own interpreters, however.

(2) Introduction tape. The introduction tape, as noted above, explains to the subject the nature and purpose of the study and provides him with a sample story. The sample story with its questions is designed to provide all subjects with an adequate and reasonably uniform amount of training before they are given tests on which their performance is considered to be a suitable estimate of how well they can actually understand another dialect. The sample test may be graded in the sense that a subject's responses may be recorded; his performance is used only to determine whether he really knows how to take the test. Any part of this tape, or all of it, is sometimes repeated for a subject.

The linguist usually does not proceed with the testing until he is satisfied with a subject's performance on the introduction tape. If he decides to reject a subject, he usually terminates the testing after playing the hometown test tape. In addition, he thanks the subject for cooperating and tips him (if the team has decided to tip subjects). Whatever he does, the linguist tries to make the rejection as inconspicuous as possible.

(3) Hometown test tape. The second tape that is presented to a subject is the test tape for his own dialect. This tape provides the basis for (a) further training the subject, (b) comparing his scores with those of the other subjects of that test point, (c) comparing the sample mean score of subjects from that test point with sample mean scores from other test points on a particular reference point test tape, and (d) adjusting scores, either individually or as sample means (town averages).

The linguist has sometimes used the subject's performance on the hometown test as an additional criterion for deciding whether to continue or terminate the testing. Subjects usually score nearly 100% on the hometown test with the present test (the main difference between subjects is that of the time required to administer the test to them, not their hometown test score). Linguists administering the test have noticed that most subjects who scored less than 80% on the hometown test were not cooperating fully.
(They usually seemed antagonistic, intoxicated, or incapable of learning how to take the test.) Thus, a linguist generally expects a subject to score at least 80% on his hometown test, though he does not automatically reject any who score less than that.

Training beyond that provided by the introduction tape is sometimes required for the test tapes because the subject must select one statement from a body of text to answer a particular question. (The content of the sample story is insufficient in quantity to make this a problem for him when he takes the sample test.) Rather than answer from a specific statement in the text, a subject sometimes tries to remember and repeat everything he has heard. To teach him what to focus on, it is sometimes necessary to repeat both the part of the text which answers the question and the question itself. Sometimes it is even necessary to ask the subject directly what the statement said.

After the linguist is certain that the subject has learned how to take the test, he repeats only a portion of the text or a question when the subject asks him to, because he did not hear it, or when the linguist is sure that the subject did not understand a particular question. Most subjects are adequately trained by the time they have finished the hometown test tape.

Occasionally a linguist gives hints to the subject if he does not answer quickly. Such hints are designed to lead the subject to an incorrect answer if he tries to guess by repeating what the linguist has suggested. Since the responses would be scored incorrect if no hint were given in the first place, it seems reasonable to put this condition on the giving of hints. On the other hand, if the investigator gives a hint which implies something other than what the text said, but the subject responds with the correct answer, the linguist can be reasonably confident that the subject answered correctly because he understood that portion of the text. It has frequently been found that hesitation on the part of the subject, or an insistence on repeating the question rather than answering it for the hometown test tape may mean that the subject has not yet learned how to take the test. For succeeding tests, however, the same behavior often indicates that the subject does not know the answer to the question because he cannot understand what the text says and furthermore he is embarrassed to admit it. This is seen, for example, when a subject answers the questions to one reference test tape easily, but repeats the questions or hesitates to answer them on the next tape. It usually happens that he scores high on the first tape and quite low on the second, a pattern followed by other subjects at that test point. For succeeding test tapes, therefore, the administrator may choose to score the subject incorrect for such a response and continue the testing without attempting to focus the subject's attention on the information sought by a particular question.

The linguist may forego using the tape and personally question the subject over the content of the statement in the text. Occasionally a linguist has refrained from playing any of the questions to a subject for a test tape, or even for the entire set of test tapes. This is because a few subjects have seemed confused by the recorded questions but have responded to the administrator's questioning. This practice is debatable since it introduces some lack of uniformity into the testing situation. However, it seems that the lack of uniformity in the ability of subjects to catch on to taking the test may be based on personal or educational differences between subjects. It may be necessary to vary one's approach so that he can even test subjects
such as these. The scores made by subjects tested this way are equivalent to those made by subjects who do catch on to the test format.

4) **Reference point test tapes.** After a subject responds to the hometown test tape, he listens to the reference point test tapes. These tapes are ordered so that the ones likely to be more difficult are bracketed by others that should be easier to understand. This ordering is determined by the investigator's best guess and may be revised after two or three subjects have been tested. If possible, the ordering is fixed so that no subject is presented with two difficult tapes in sequence (50% or lower intelligibility). If the tests are too difficult for a subject he loses interest in the testing; some subjects have even refused to finish. In addition, bystanders who might have cooperated with the team have refused to do so after learning that the test tapes were difficult to understand.

Since it is rarely possible to test subjects in isolation, the test tapes are played through earphones. In this way external distractions are kept at a minimum, and observers who may be tested later do not become acquainted beforehand with the test materials. The bystander can hear the subject's responses to the questions; therefore, it is necessary to ensure that they do not associate these responses with a particular test tape. This is adequately accomplished by rotating the order of presentation of the reference point tapes from subject to subject.

It was mentioned earlier that the reference point tapes are presented to successive subjects in an order that systematically changes (see Section 2.2.2). In earlier studies, however, the tapes were presented in a fixed order, with the hometown test first. The survey team observed irregularities in the subjects' performance; scores were frequently higher for a reference point tape than they were for the hometown tape. In addition, subjects often did poorly on the last tape in the sequence. The first effect was attributed to incomplete learning; the second to fatigue. Since these seemed to operate systematically it was decided to rotate the tapes between successive subjects so that each test tape would occupy each ordinal position in the sequence of tapes in a systematic way, thus distributing the disturbing effects equally over all the test tapes.

This systematic rotation, then, proceeds as follows: after a test tape has been played, it is rewound in preparation for the next subject, generally by the technician with a second machine, and placed "face down" on the table. This is repeated for all of the reference point tapes until the entire set is in this "face down" pile. The pile of test tapes is then inverted, and the top tape is removed and placed on the bottom of the pile. In this way, an ABCDE order becomes BCDEA and so on. Thus five different orders can be distributed among ten subjects. An alternative rotation includes the hometown test tape. This further randomizes the testing by distributing the effect of incomplete learning by the subject over all of the test tapes instead of allowing the cumulative effect to be on the score of the hometown text. Furthermore, this reduces the possibility that someone will learn all the answers to the hometown text and associate these answers with that particular test tape.

A third system of rotation pairs the easy and difficult test tapes (including the hometown tape) and then rotates by pairs so that only the easy test tapes appear first in the ordering. However, this method is somewhat more complicated for devising easily-selected orderings. All three systems have been used; the first two seem generally preferable.
(5) Scoring. Subject responses are evaluated as correct, incorrect, and occasionally, half-correct. The usual correct response consists of the repetition of a word or phrase from the appropriate sentence in the text. Sometimes, however, the subject responds with a synonym or paraphrase based on some association with the words found in the text. For example, the answer to one question for a Mixtec text was a term meaning ‘one who walks around in the mountain’, i.e., a goatherd tending his animals. Several subjects responded with the answer “young man.” These responses were written down and checked with the informant who provided the text. It was learned that in the text, and generally for the region, a goatherd is a young man or boy. The assumption was made, then, that the subject’s responses were made on the basis of this extralinguistic knowledge about the phrases in question. Thus, these responses were evaluated as being correct. The point is that the answer which is not a word-for-word repetition of the phrase found in the text is not necessarily an incorrect response. It is necessary for the tester, when he evaluates a response, to consider possible features of the text and possible associations with these features that might influence a subject’s response.

The linguist’s evaluations of those responses which he is sure are correct or incorrect are written down, but their content is not. The information about correct responses is at least partially retrievable from the transcriptions and translations of the test materials. To explain why subjects missed certain questions we might do well to write down in full all incorrect responses.

On the other hand, doubtful responses are written down in full on the score sheet with the linguist’s evaluations of them. This has been helpful especially in cases where a doubtful response seems plausible. With this information recorded it has been possible to revise the evaluation and to correct occasional inconsistencies in scoring. Evaluations of doubtful responses are made, as much as is possible, in terms of “correct” or “incorrect.” However, the linguist sometimes feels that he has no alternative but to evaluate a response as half-correct. For example there are instances when a subject obviously understood a portion of the text, but that particular portion is not related to a particular question. Furthermore, the subject’s thinking becomes fixed on this irrelevant portion and the linguist is unable to bring the subject back to the material related to the question. In such a case, the evaluation might be half-correct, rather than correct or incorrect. We have been somewhat reluctant to allow a linguist to use evaluations of “half-correct,” but this reluctance seems no longer justified. It turns out that, is we carefully record the content and evaluations of responses which the linguist was obligated to call correct or incorrect, and later if we are able to determine exactly what the correct answer is, we usually find that the linguist has inconsistently evaluated the response. Furthermore, the inconsistencies tend to cancel one another out so as to have the effect of turning a linguist’s evaluation of a doubtful response into an automatic evaluation as half-correct. Thus it seems plausible to allow such evaluations since their use is convenient for the linguist and does not systematically bias the scores.

Finally, a few alert subjects have attempted to mimic the phrase from a reference point text. A simple translation test into his own dialect, or into Spanish, has detected this and enabled the investigator to score that particular response. For example, if an investigator suspects that the subject is mimicking a particular response, he may ask him, “What does ______ mean in your dialect?” or “What does ______ mean in Spanish?” If the subject is unable to answer those questions, then his response may be evaluated as incorrect.
subject is unable to answer those questions, then his response may be evaluated as incorrect.

(6) **Flexible aspects of the testing.** The roles of the linguist and the technician are fixed for some aspects of the test administration and fairly flexible for others. The investigator instructs the subjects, administers the test tapes to them, and evaluates the responses. He may also enter the evaluations on the individual score sheets, or leave this to the technician. If he does the latter, then the team members improvise a system for communicating whereby the technician knows what the investigator's evaluation is with a minimum amount of verbal communication. For example, in the Mazatec study, Kirk evaluated the responses and said "okay" in Spanish to each subject's response, whether correct or incorrect; he added, in English, "auht" to each incorrect response. In this way the technician was able to write down the evaluation for each question. In addition, he wrote down the difficult-to-evaluate answers to various questions as Kirk quietly mentioned them.

One advantage of leaving the paper work to the technician is that it allows the linguist to concentrate more closely upon the subject. It also allows the writing of scores to be a little less conspicuous to the subject. On the other hand, the investigator may feel that he could record the evaluations more efficiently himself.

Technical aspects of the testing are also flexible. It was pointed out earlier that the survey team uses two recorders, one for recording the test materials and the other for playing the materials when the tests are constructed (see Section 2.1.4). Although the input recorder we use has higher recording capabilities than the output recorder, it gives less playback time per set of batteries than the output recorder does. Therefore it is not used for playing the test tapes. However, it is used during the testing to rewind the test tapes for presentation to the following subject. (It is possible to use the playback recorder to do this rewinding, but that would increase the time for administering the tests. In addition, the batteries in the input recorder are usually so discharged from use in test preparation that they would not be used at a subsequent test point. Rather than throw away half-used batteries, the technician usually decides to use them up completely in rewinding the test tapes.)

Two or three sets of earphones are connected to the output recorder, depending upon how the survey team decides to arrange the work pattern. If the linguist operates the output recorder, the technician is left to rewind the used test tapes with the other recorder and to record the investigator's evaluations on the score sheets. Only two sets of earphones are needed if the team follows this pattern. On the other hand, the technician may operate the playback recorder, leaving the investigator to listen through a third set of earphones and to record his evaluations of the subject's responses. This method is more complicated technically, and it is possible that the first method in which the investigator operates the test machine is a little more efficient, especially if it is necessary to stop the machine and to repeat part of the text. Both methods, however, have been used.

The technician, then, rewinds the test tapes. When he has rewound the entire set, he rotates them for the following subject in the manner described earlier. If he also records the evaluations of the subjects' responses, he may compute the individual scores during the testing. However, since there are frequently observers close by, and since a few of
these may know how to read, he may postpone his calculations until no one is around to watch.

To summarize: stage two of a dialect survey is carried out on a second trip through the language area during which the test sets are modified for each test point, and ten subjects are selected there. Generally the sample consists of the first ten speakers of each dialect who are cooperative. These subjects are engaged either through the help of a local person or through the efforts of the survey team members. Each subject, then, is tested on the set of tapes. The mean score for the entire sample of subjects on each test tape constitutes a point estimate of the level at which speakers of one dialect understand speakers of another dialect.

2.3. How the Data Are Processed. As the third stage of a dialect survey, the materials collected are archived, and the intelligibility results are tabulated and interpreted in such a way that their results can be used in decision making.

2.3.1. Archiving the materials. The data for archiving include both tape recorded and written materials. The taped materials consist of the original recordings of all text materials, the introduction tapes, the hometown test tapes, and the reference point test tapes. The written materials consist of transcriptions of the texts used in the test tapes, translations of these texts into the national language, transcriptions and translations of the questions used for the test tapes, transcriptions of the introduction tapes, word lists, ethnographic questionnaires, individual score sheets, and preliminary reports.

The original recordings of the texts are kept on five-inch master reels of tape. Each text, whether used in the study or not, is prefixed on the master tape by a recorded phrase such as "Mixtec of Santiago Nuyoo." The texts which were actually used in the study are also identified on the box in which the master tape is kept, and the boxes themselves are labelled "Survey Texts—Master Copies" and they may be numbered sequentially, e.g., "Totonac one," "Totonac two," etc.

The introduction tapes and test tapes are generally kept on the three-inch reels which were used during the testing. To each of these tapes, a section of leader tape has been spliced on which the designation of that particular tape is written; e.g., "Introduction Tape: San Jerónimo"; "Test Tape: Huehuetlán." Each of these tapes is transported in a labelled box or envelope during the survey; these boxes may be retained for archiving the test tapes. In an alternative method, the tapes are spliced together and wound onto a five-inch storage reel; the identifying leader tapes are retained for each test tape, and the contents of each storage reel are indicated on the outside of the box. Generally the introduction tapes are stored on one five-inch reel and the test tapes are stored on another. Each of these boxes containing storage reels may then be identified and numbered, e.g., "Choapan Zapotec Storage Reel No. 2." The advantage of the first method is that it leaves the individual test tapes readily available for future use if they are ever wanted. The second method allows a more economical use of three-inch spools during the field procedures; that is, as the testing is completed at a given test point, the technician is able to wind the introduction tape (and any other test tape that is not going to be used again in the study) onto one of the storage reels. Then he is able to use at another test point the three-inch spools on which these test tapes were wound.
The written materials are filed in manila folders according to one of two systems. In the first, all of the written materials for a test point are filed in the same folder. The file itself is labelled "Dialect Survey Written Materials"; the dividers are headed "Mazatec Study," "Zoque Study," etc.; and the individual folders within a section of the file are designated by the names of the particular test points. In the second classification, the written materials are classified into test materials, word lists, and ethnographic data. A separate folder is used for each of these categories for all the test points of a particular study. The system can be illustrated as follows:

File: Dialect Survey Written Materials  
Divider: Mazatec Study  
Folders:  
Mazatlán—Test Materials  
Mazatlán—Word List  
Mazatlán—Ethnographic Data

The second classification is preferable because it allows quicker reference to a particular class of data than the first one.

Data on the informants are entered with the materials that were obtained from them, on the standardized form for the word list or on the scoring form for the tests. It is convenient to collate all the data on informants for each study and to file them in a separate folder. A standard form for these data includes at least the following: name, age, date, sex, language and dialect (test point), educational background, residence (local or outside), and type of material elicited, e.g., text, questions, word list, etc.

The transcriptions and translations of the texts, questions, and introductions used in the tests are typed in an interlinear format. Individual score sheets are kept in separate notebooks or binders for each survey. The score sheets for each test point are ordered sequentially for the subjects tested. An abbreviation for the test point and the number of the subject tested are noted in the upper right-hand corner of each sheet. For example, the score sheet for the sixth subject tested in San Jerónimo Técoatl is labelled Te-6. A table of contents is provided for each notebook. This facilitates the locating of a section of the score sheets when one wants to refer to it. Each section of score sheets for a particular test point is preceded by a page with a summary table of the scores for these subjects on the test tapes. At the bottom of each column of this table the sample mean score of a test tape is indicated; for convenience a summary table of sample mean scores for the entire study is placed at the front of the binder. Examples of the score sheet and summary charts are provided as Appendix C and Appendix D. It is useful to have standard forms for these printed beforehand.

2.3.2. Displaying the data in matrices. The intelligibility scores are displayed in three kinds of tables. The data thus arranged are converted into a form useful for decision making. The first table presents the raw scores that the sample of subjects from a particular test point registered on a set of test tapes. In this table the subjects are enumerated down the left side; the test tapes used are listed along the top by abbreviations for the names of the corresponding reference points. Thus, subjects correspond to rows in the table and test tapes to columns. The score a particular subject received on a test tape is entered in the appropriate cell of the column corresponding to that test tape. The entry in the table may be either the
number of correct responses out of a possible ten that a subject scored, or the number of correct responses multiplied by ten, which gives a point estimate of the percentage of intelligibility at which subject X understands the test tape Z. The way that these scores are indicated in the table is purely a matter of preference. There will be one of these tables for each of the test points.

The second table is a summary of the sample mean scores obtained from the tables of individual scores mentioned above. The scores in the cells of each column of the first set of tables are summed and divided by the number of subjects in the sample (occasionally less than ten). The quotient is multiplied by ten to yield a point estimate of the level at which the sample of subjects understood a particular test tape. It should be noted that if the entries in the tables of individual scores were previously converted into percentages, and were not simply left as the number of correct responses a subject made to a test tape, then the percentage for entry into the summary table can be directly obtained by summing the scores in the cells of a column and dividing the sum by the number of subjects in the sample. Both methods are algebraically equivalent; their use depends solely upon how the data were entered in the individual raw score table. In this summary table the reference tapes are indicated down the left side; the test points are indicated along the top.

The third kind of table retains the format of the summary table of sample mean scores; the only difference between the tables is that the raw sample mean scores have been adjusted by means of a proportion. The adjustment begins by normalizing all hometown test scores to 100, i.e., perfect intelligibility, which it is assumed holds for all native speakers of a dialect. The scores for all reference point tests are then proportionately adjusted to this base of 100 by the formula

\[
\frac{100}{hs} = \frac{X}{rs}
\]

where \(hs\) is the raw hometown score, \(rs\) is the raw score on a reference point test tape, and \(X\) is the adjusted score. The general effect of this adjustment is to preserve the relative magnitudes of intelligibility registered between a set of test tapes by one sample of subjects. However, the absolute value which is added to a particular test tape is determined by the raw score on the hometown test tape and the raw intelligibility score on that reference point tape. In general, the lower the raw score for a reference point test tape, the smaller the value that is added to that raw score by the proportion.

Adjustments have also been made on the basis of an additive factor rather than a multiplicative one. In this case, the raw hometown score is subtracted from 100, and the resultant difference is added to each of the raw sample scores on the reference point tapes. The same absolute value is thus added to all the reference point test scores. The formula is

\[
as = rs + (100 - hs),
\]

where \(as\) stands for adjusted score, \(rs\) is the raw reference point test score, and \(hs\) is the raw hometown score. The principal difference between the effect of this adjustment and the proportional one is seen when both a hometown raw score and a reference point test score are relatively low, say
92% and 43% respectively. In this case the additive adjustment is the higher, 51% and 47%.

These adjustments were made on the basis of earlier studies in which hometown scores were generally low and quite variable. It was assumed that whatever was affecting the hometown score was also affecting reference tape scores. Thus it was felt that two sample mean scores of 65% on a reference tape did not necessarily indicate the same degree of understanding as, for example, when one sample registered an 87% hometown score, and the other a hometown score of 68%. If the “constant noise” assumption is correct, an adjustment based on the proportion gives the first sample a score of 78% on the reference tape and the second one a score of 91%. The interpretation of the mean scores, then, differs, depending upon whether it is based on raw scores or adjusted scores. (The validity of the adjustment must be determined by additional data.) Since later modifications of the test have yielded increasingly reliable results, there has been an increased tendency to simply raise the hometown scores to 100 and leave the other scores unadjusted in readying the scores for computation.

Finally, when an adjustment is made, it is on the level of the sample mean rather than the individual mean score. This is because a sample mean score is more reliable than an individual mean score. In addition, there is often a temptation to significantly alter or reject a highly anomalous individual score on a particular test tape. For example, in the individual raw score table for Mazatzongo (Table 1), notice the score which the seventh subject recorded on the Huautla test tape. It would be tempting to adjust that score to seven, or to reject it completely on the strength of a comparison of that subject’s other scores with the scores of the other subjects. Either choice is highly suspect and the discrepancy should be left to random error.

Examples of these three tables, individual raw scores, raw sample mean scores, and adjusted sample mean scores are presented on the following pages. The data are from the Mazatec study. Starting at the left, the columns for the raw data matrices are arranged from highest to lowest intelligibility, based on the sample mean score on each test tape. The first column at the left of each table presents the scores on the hometown test tape. Since the columns are in descending order, it can be seen that there is a general profile of scores common to the majority of subjects of each sample. Some obvious aberrations are also noticeable. In explaining this, it should be noted that where the sample means for a pair of test tapes are quite close, there is not much agreement among the subject responses as to which of the two test tapes consistently scores lower. This phenomenon reveals that the test is not sensitive to relatively minor intelligibility differences, say 5-10%. Other irregularities can be attributed to incomplete learning by the subject of how to take the tests and to the relative order in which the test tapes were presented.

The second table presents the raw mean sample scores for the Mazatec study. The scores along the diagonal are the raw hometown sample means. These means lie between 89% and 100%. An examination of rows 1, 9, 15, 21, 22, 23 and the corresponding columns reveals the six major dialectal areas of Mazatec as determined by the 80% threshold value. All twenty-three towns, then, fall within the dialectal areas of Chiquihuitlán, Huautla, Ixcatlán, Jalapa, San Jerónimo, and Soyaltepec. Regarding scores
**Overview of a Dialect Survey**

### Reference Points

#### San Lorenzo
- Lo: San Lorenzo
- Te: San Jerónimo
- Hu: Huautla
- Mz: Mazatlán
- Ja: Jalapa

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Lo</th>
<th>Te</th>
<th>Hu</th>
<th>Mz</th>
<th>Ja</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>10</td>
<td>6.5</td>
<td>6</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>8.5</td>
<td>6.5</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3.</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4.</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>5.</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>6.</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7.</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>6.5</td>
<td>2</td>
</tr>
<tr>
<td>8.</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>9.</td>
<td>9.5</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>10.</td>
<td>10</td>
<td>8.5</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

#### Mazatzongo
- Ms: Mazatzongo
- Hn: Huehuehtlán

<table>
<thead>
<tr>
<th>Ms</th>
<th>Lo</th>
<th>Te</th>
<th>Hn</th>
<th>Hu</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>2.</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>3.</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>4.</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>5.</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>6.</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>7.</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>8.</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>9.</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>10.</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

#### San Mateo
- Mt: San Mateo

<table>
<thead>
<tr>
<th>Mt</th>
<th>Hu</th>
<th>Te</th>
<th>Lo</th>
<th>Hn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2.</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>3.</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>4.</td>
<td>10</td>
<td>7.5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>5.</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>6.</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>7.</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>8.</td>
<td>9</td>
<td>8</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>9.</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10.</td>
<td>8</td>
<td>7.5</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

Subject no. 1 was the informant who supplied the hometown test tape. He was not actually tested on the hometown test tape.

Table 1. Raw Scores for Three Samples of Ten: Mazatec Study
Table 2. Raw Sample Mean Scores: Mazatec Study

Tests from towns on the vertical axis were given to subjects in towns on the horizontal axis.
### Table 3. Adjusted Sample Mean Scores: Mazatec Study

Tests from towns on the vertical axis were given to subjects in towns on the horizontal axis.
computed on samples of less than ten, two reference points, San Pedro (11) and Zacatepec (17) yielded samples of nine and six, respectively.

Finally, the summary table of adjusted mean scores (Table 3) can be used to define the major dialectal areas by applying the 80% threshold value. In this case there are four clearly defined areas: Chiquihuitlán, Huautla, Jalapa, and San Jerónimo. There are four additional areas of marginal intelligibility with either the Huautla or San Jerónimo dialectal areas: Ayautla, Ixcatlán, Mazatzongo, and Soyaltepec.

2.3.3. An optimization model. Grimes has recently attempted to interpret intelligibility data in terms of an optimization model (Grimes, in press). He views the determination of minimum communication needs as a problem in resource allocation. (He shows that the model can be applied to a wide range of measures of dialect relationships. For this study, however, we confine our explanations strictly to its application to intelligibility measures.) He begins with an analogy to the Hitchcock-Koopmans transportation problem in which, for example, a fixed quantity of goods is produced at M factories and is to be delivered to N shops. Each source (factory) has a certain portion of the total goods to produce, and each destination (shop) requires a certain amount of these goods. The total amount of goods produced by all the sources is equal to the total amount of goods required by all the destinations. In addition, the costs for shipping a unit of goods from any one of the sources to any one of the destinations are known. The problem is to specify at what level each source factory must produce goods and where each must ship them so that the cost of shipping will be minimal for the entire network.

The procedure that allows one to solve the transportation problem also fits the survey problem. In order to maintain the formal restriction that the total amount of goods produced must equal the total amount of goods received, it is necessary to add a dummy destination for unused source quantities; this keeps the calculations mathematically consistent. Grimes also adds a fixed cost to the known shipping costs. It is added for each source that is used to ship any goods at all to some destination. If a source is not used, then the fixed cost for it is zero.

(1) The dialect survey formulation. Allocations made in a solution to the transport problem translate easily into the features of communication networks. Goods actually shipped from a source to a destination correspond to adequate communication from a given primary (source) center to elsewhere in the language area (destination). The destinations to which goods are delivered correspond to secondary points in the survey, each of which is set by the solution procedure in adequate communication with one primary center. This constraint is expressed in the transport model as a destination requirement of 1; when it is not met between a primary center and a secondary point, the lack of communication is given the value of 0. Moreover, communication from the primary centers must be distributed among all of the secondary points for the language area.

Shipping costs are conceptually parallel to difficulty in communication from a primary center to a particular secondary point. This measure of difficulty is derived from the intelligibility data. The fixed cost corresponds to a threshold value for grouping destinations around a source. It has the effect of minimizing the number of sources brought into play at any threshold level. A grouping of destinations around a source corresponds to a grouping of secondary points around a primary dialect center, separated
from other groupings by low intelligibility. The problem, then, is to
determine for a multilingual area the smallest possible set of primary
centers capable of providing adequate communication to the entire area.

(2) How the networks are computed. To compute an optimal network,
intelligibility scores are first converted into cost figures. This is done by
subtracting each score from 100, which represents perfect intelligibility. High
intelligibility scores thus yield low cost figures and vice versa. Whereas high
intelligibility scores indicate ease of communication, high cost figures signal
difficulty.

The cost figures are inserted into a cost matrix in which each column
represents a real test point and each row represents a reference point. The
figures entered at the intersection of a row $i$ and a column $j$ thus indicate
how difficult it is for speakers of dialect $j$ to understand speakers from
dialect $i$. The cost matrix also includes a dummy test point (destination)
column in which all costs are zero.

Before any calculations can be carried out, the cost matrix must be put
into an optimization table along with several other elements. The first of
these elements is an allocation matrix which has the same number of rows
and columns as a cost matrix. The linguistic networks calculated by the
method are represented by this allocation matrix. The second element is a
requirement vector consisting of one component for each real test point;
each component has the value of 1. The sum of the real test point
requirements thus equals the number of real test points included in the
table. In this way the solution procedure guarantees that every test point in
the table will come out in adequate communication with some reference
point.

In addition, there is a source quantities vector each of whose
components corresponds to one of the reference points in the table. Each
component is equal to the sum of the real test point requirements. Thus the
procedure leaves room for the possibility that any reference point could be
the center for all of the test points in the network, if communication is
adequate between it and all the test points.

The table also includes a fixed cost vector whose components
correspond to the reference points. This vector minimizes the total number
of centers of communication that will be selected with reference to a
threshold value that is equal for each component. The threshold value limits
the degree of difficulty under which adequate communication can still be
made from reference point $i$ to test point $j$. Finally there is a resultant cost
vector whose components also correspond to the reference points. The
values entered in this column are based on calculations that are described
in a later paragraph.

Table 4 shows how these elements are arranged in an optimization table
in its zero state, i.e., before computation begins. The cost and allocation
matrices are separated by a double line from the other elements in the
above table. The values assigned to the elements and the relationships
among them are also shown. For example, in the table, each cell of the
cost matrix $A$ contains a cost figure which is based on data collected
during the Choapan Zapotec survey. Since reference points are always test
points, cost figures for communicating from each reference point to itself
are included in the matrix. They are indicated along the diagonal. (This
overlap of reference and test points results from applying the survey
procedures; it is not a necessary part of the model.) Note that these cost
Table 4. Choapan Optimization Table: Zero State

Elements of allocation matrix B are written here as superscripts to the corresponding elements of the cost matrix A. For example, \(2^0\) indicates an allocation of 0 to a cell whose cost is 2.

<table>
<thead>
<tr>
<th>(i = 1)</th>
<th>1</th>
<th>2 (= 3)</th>
<th>4</th>
<th>5</th>
<th>d</th>
<th>c</th>
<th>x</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>10(^0)</td>
<td>0(^0)</td>
<td>15(^0)</td>
<td>30(^0)</td>
<td>24(^0)</td>
<td>0(^5)</td>
<td>15</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>20(^0)</td>
<td>0(^0)</td>
<td>0(^0)</td>
<td>4(^0)</td>
<td>2(^0)</td>
<td>0(^5)</td>
<td>15</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>20(^0)</td>
<td>8(^0)</td>
<td>0(^0)</td>
<td>2(^0)</td>
<td>2(^0)</td>
<td>0(^5)</td>
<td>15</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>25(^0)</td>
<td>(20(^0))</td>
<td>(15(^0))</td>
<td>0(^0)</td>
<td>12(^0)</td>
<td>0(^5)</td>
<td>15</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Key

- \(A\): Cost Matrix
- \(B\): Allocation Matrix
- \(j\): Test Points
- \(i\): Reference Points
- \(r\): Requirement Vector
- \(d\): Dummy Destination
- \(c\): Fixed Cost Vector
- \(x\): Source Quantity Vector
- \(\xi\): Resultant Cost Vector

Figures are normalized to 0. This is based on the fiction that perfect communication holds among all people who speak the same dialect. (This assumption seems to be useful for interpreting intelligibility results, but it cannot be justified for other kinds of tests.)

Each cell of the allocation matrix \(B\) corresponds to a cell of \(A\). The quantities entered into the cell of \(B\) are indicated as superscripts to the cost figures of \(A\). These quantities, 1 and 0, denote the presence or absence of adequate communication between reference point \(i\) and test point \(j\). The cost value in the dummy column \(A\) is 0; there is no cost for communicating to it from one of the reference points. The allocation entered initially in the cells of this column for \(B\) equals the source quantity, i.e., the number of real test points. The communication requirement for each real test point is indicated by entering a 1 in each of the first five cells of the requirement vector \(r\). The value in the sixth cell of \(r\) equals the sum of the components in the source quantity vector. The threshold value, set at 15, serves as each component of the fixed cost vector \(c\). The quantity in each component of the source quantity vector \(x\) is five since there are five real test points. The value of the cells of the resultant cost vector is zero since no allocations of resources have been made and no sources are in use.

The calculations involved are (a) an initial allocation of resources to the columns of matrix \(B\), (b) the computation of resultant row cost values, and (c) the minimization of the total network cost by reallocating resources between pairs of rows in the matrix.

Table 5 shows the initial allocation of resources subject to the test point requirements. This allocation is made either from the source vector \(x\) or
Table 5. Initial Allocation of Resources from x

<table>
<thead>
<tr>
<th>i</th>
<th>Reference Points</th>
<th>j</th>
<th>Test Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Comaltepec</td>
<td>1</td>
<td>Comaltepec</td>
</tr>
<tr>
<td>2</td>
<td>Choapan</td>
<td>2</td>
<td>Choapan</td>
</tr>
<tr>
<td>3</td>
<td>Jalahui</td>
<td>3</td>
<td>Jalahui</td>
</tr>
<tr>
<td>4</td>
<td>Xochiapan</td>
<td>4</td>
<td>Xochiapan</td>
</tr>
<tr>
<td>5</td>
<td>Arenal Grande</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

from the vector d (it turns out the same either way) to the cell in each real test point column that has the lowest cost figure. If there are two cells in a column with the same minimal cost figure, then the resource is arbitrarily assigned to the cell closest to the top of the column, as in Column 3 of Table 5. For each row that contains a cell to which a quantity is allocated, there is a decrement in the corresponding component of the dummy column in B. The amount of this decrement equals the sum of the quantities allocated to cells in the row it corresponds to. To see how this decrement is made, compare the dummy column with Rows 1, 2, and 4 in Tables 4 and 5. The requirement vector r relates to the allocation of resources by showing a constraint: each real test point must receive adequate communication from a reference point. The presence of 1's in the allocation matrix shows that the initial allocation has been made.

During all calculations two relationships between destination (test point) requirements and source quantities are held constant. First, the sum of the source quantities in a row that represents a reference point and the quantities that remain in the decremented component of d that the row corresponds to equals the source quantity in the cell of x. Thus, the sum of the source quantities in each row of the matrix in Table 5 (the first six columns of the table) is 5, which equals the source quantity in x. Second, the sum of the components of d plus the sum of the components of r equals the sum of each component in x minus the sum of the real test point requirements. In this case, for example, the sum of x1 is 20 and the sum of the requirements r1 is 5; therefore the sum of the d1 components equals 15. This figure, which Grimes calls the n + 1th component of r, is entered into the cell where the vectors d and r intersect. These relationships are maintained to keep the calculations correct; they do not correspond directly to any element of this survey formulation.
A resultant cost figure $C_i$ is derived by selecting each row in Matrix A, multiplying the cost figures in each of its cells by the quantity in the corresponding cell of B, adding up the results for these cells, and adding to this the fixed cost, if any. Thus, for the second row of the matrix, in Table 5, the calculation is as follows:

$$C_2 = (10 \times 0) + (0 \times 1) + (0 \times 1) + (4 \times 0) + (2 \times 1) + (0 \times 2) + 15 \quad = 17$$

For the third row, however, the cost is 0. This is because the procedure specifies that the fixed cost is added only to the rows to which an allocation has been made outside the dummy column. Since no allocations have been made to this row, the fixed cost factor is not activated. Finally, the sum of the resultant costs for the first, third, and fourth rows is 30. This cost will eventually be reduced.

The total network cost is derived by summing the resultant row cost values for all rows in the matrix. In Table 5 this network cost is 47. The cost developed within the matrix itself is only 2, but since allocations have been made to three different rows, the fixed costs add up to 45. In the initial allocation stage, the matrix cost alone is minimal. Note, however, that allocations have been made to several separate centers. Since the fixed costs are added for each allocation, the total network cost is high.

The network cost is reduced by reallocating quantities from one row in B to another. In this process, the threshold value, which is assigned as the fixed cost, comes into play to limit the destinations to which a reallocation can be made from a particular reference point. To begin with, a pair of rows is selected; for example, Rows 2 and 4 in Table 5. Then the resultant cost for each is computed and the two costs are summed. Note that the resultant cost for Row 2 is 17; for Row 4 it is 15. Their total cost is 32. To reduce the total cost for a pair of rows, the cell to which a source quantity has been allocated is compared with the corresponding cell in the other row. If switching the allocation to the second row would result in a higher cost to one of its cells, but would eliminate the fixed cost element of the other row in such a way as to bring the total cost of the two rows down, the switch is made. This is indicated in the matrix by interchanging the allocation quantities between the two cells.

To illustrate the reallocation, compare the fourth cells of Rows 2 and 4. The cell of 4 has a cost figure of 0 and an allocation of 1. The corresponding cell of 2 has a cost figure of 4 and an allocation of 0. Since the cost figure in 2 is less than the threshold value of 15 (in the fixed cost vector), the source quantity in 4 can be reallocated to 2. This increases the resultant cost of Row 2 from 17 to 21, but it eliminates a fixed cost component from Row 4, since there are now no source quantities allocated to it and the fixed cost is not activated. Thus, the total cost for the two rows is reduced from 32 to 21.

Each reallocation produces a new state of the matrix. This new state, in turn, is the base for making further reallocations. (After the initial allocation, all new allocations continue to meet the demands of the requirement vector.) Since the cost increment in the matrix is smaller than the threshold value, the total network cost is reduced when an allocation is made. (A possible exception to this is when an outlier is joined to a reference point at the threshold value. In this case the network cost remains unchanged.)
Table 6. Cost Reduction by Quantity Reallocation

Table 6 shows the cost reduction after two source quantities have been reallocated. The reallocation involved the first, third, and fourth rows, whose total cost shown in Table 5 is 30. By reallocating the quantities from Rows 1 and 4 to Row 3, the cost is reduced to 19, i.e.,

\[ C_3 = (2 \times 1) + (8 \times 0) + (0 \times 0) + (2 \times 1) + (2 \times 0) + (3 \times 0) + 15 = 19. \]

For the network, the total cost is reduced from 47 to 36.

After one pair of rows is compared the process continues, a pair of rows at a time, until all the pairs of rows in the matrix have been compared and the most economical reallocations have been made. In some cases it will be necessary to make successive reallocations within the same pair of rows on separate passes as will be done, for example, with the second and third rows in Table 6. Additional passes through the pairs of rows are made until subsequent reallocations do not further reduce the total network. At that point, then, the network is in the optimal state that can be allowed under the stated threshold value. Table 7 presents the optimal network for the Choapan data with a threshold of 15. The total network cost has been reduced from 47 to 29.

Grimes points out that, although the computations can be done by computer, they can be as readily performed by hand. For hand computation, the table can be simplified by dropping the dummy test point and source quantity columns, thus bypassing the management of source quantity totals in the rows and columns. (All this can be retained to help keep the calculations from going astray, however.) Table 7 is presented in this abbreviated form.

Grimes also suggests three refinements in the allocation process. In the first place, the individual cost figures represent point estimates of the average difficulty of communicating from a reference point to a test point. It would represent the data better to use range estimates of the difficulty of communication. These are derived from the data by standard procedures and are stated in terms of a range of values within which the true level of communication is presumed to lie at a given probability level (see Appendix H, p. 171); e.g., it might be better to say that test point \( j \) understands

\[^{8}\text{For the algorithm for deciding whether to reallocate or not, see Bellman and Dreyfus 1962, 70-96.}\]
reference point at a level between 83% and 93% instead of at an absolute level of 88%. The certainty of this range estimate is 95%, assuming randomness in the sampling procedure. (We cannot meet the randomness assumption with our present procedures, however.)

In addition, the decision to reallocate quantities in Tables 4 to 7 is based on a simple quantitative difference between cost figures in the corresponding cells of a pair of rows. It would be better to test the statistical significance of the difference between these figures and then reallocate the quantity only if the figure in one cell is significantly lower than the one in the other cell. (If range estimates are used, tests of significance would still have to be carried out since it does not follow that two ranges are distinct simply because they are disjoint; neither are they equivalent merely because they overlap.) For Choapan, for example, the minimum cost of 29 is for Row 3; the cost for making all the allocations to Row 2, on the other hand, would be only 31. Since this cost difference is derived from scores that differ by only 2% (see cells two and three in the fourth column) and this difference is probably not significant, the network minimum cost for either reference point is probably not distinct, if quantity reallocation is based on statistically significant differences between cost figures.

Furthermore, because of the error probably incorporated into the intelligibility figures (and therefore inherent in the cost figures) and because of the difficulties in stating an exact threshold value as the limit on effective communication, Grimes suggests that a minimal network should be calculated for a series of threshold values. The degree of consistency through such a series suggests what the real groupings of dialects are more definitely than do results that are based on only one threshold value.

3 Contour maps of networks. A network can be represented graphically by a kind of contour map in which each contour corresponds to one threshold value in the series. Contours are drawn around each center and its subsidiaries, beginning with the lowest threshold value and ending with the first one that encircles all the dialects in the network. Usually each higher-valued contour will be more inclusive than the preceding one. A direct consequence of the requirements vector composition, which allows each test point to be associated with only one reference point, is that the contours are drawn around clusters of dialects (each cluster consisting of a
separate center and its subsidiaries) in such a way that a particular point is associated with only one center even though subjects at that test point may have registered high intelligibility with two or more reference point tapes. The number of contours that separate two points is inversely proportional to the strength of association between these points.

![Diagram of the Choapan Zapotec Network](image)

**Figure 2. The Choapan Zapotec Network**

**Key**
- Co: Comaltepec
- Ch: Choapan
- Ja: Jalahui
- Ar: Arenal Grande
- Xo: Xochiapan

The Choapan network, which has only one center at a threshold of 15, is represented by such a contour map in Figure 2. The primary center chosen is Choapan (Row 2 in the tables) rather than Jalahui (Row 3). As noted earlier, the network cost is equivalent for either point. The cost figures in the cells of Row 2 are selected as threshold values, i.e., 0, 2, 4, and 10. The separate dialects appear as peaks, all of which rise to the same height; the distinctness of these dialects is shown by the depth of the valleys that separate them. (This is determined by counting the contours.) The arrows represent the relation "subsidiary of" and are directed toward the dialect which functions as the primary center for a grouping. In terms of strength of association with Choapan, the dialects are in the following order, from strongest to weakest: (1) Jalahui, (2) Arenal Grande, (3) Xochiapan, and (4) Comaltepec. The primary separation occurs between

---

9. Grimes suggests that the requirements vector composition might be increased to two (or more) so that the individual test point requirement could be divided and allocated to two or more reference points for these cases of dialect overlapping (Grimes, in press).
Comaltepec and the rest of the dialects. A second one occurs between Xochiapan and the three most related dialects: Ch, Ja, and Ar. These relationships coincide with linguistic, historical, and sociological facts of the area: Comaltepec and Xochiapan are somewhat divergent from Choapan linguistically. The towns of Jalahuí and Arenal Grande were formed by migrations from Choapan.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Ċ</th>
<th>Ċ</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td></td>
<td>0₁</td>
<td>26</td>
<td>3₁</td>
<td>2₁</td>
<td>67</td>
<td>64</td>
<td>72</td>
<td></td>
<td>20</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>2</td>
<td>33</td>
<td>0</td>
<td>4</td>
<td>35</td>
<td></td>
<td>72</td>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>3</td>
<td>23</td>
<td>13</td>
<td>0</td>
<td>84</td>
<td></td>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>4</td>
<td>6</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>5</td>
<td>39</td>
<td>60</td>
<td>32</td>
<td>0₁</td>
<td>15₁</td>
<td></td>
<td>20</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>6</td>
<td>8₁</td>
<td>10</td>
<td></td>
<td>0₁</td>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>7</td>
<td>26</td>
<td>66</td>
<td></td>
<td></td>
<td>0₁</td>
<td>2₁</td>
<td>5₁</td>
<td></td>
<td>20</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>115</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Optimal Allocation for Ocotlán Zapoteca Area

Key

i Reference Points
1 Ocotlán
2 San Miguel
3 Santa Inez
4 San Pedro
5 Ayoquezco
6 Santo Domingo
7 San Pablo

j Test Points
1 Ocotlán
2 San Miguel
3 Santa Inez
4 San Pedro
5 Ayoquezco
6 Santo Domingo
7 San Pablo
8 Guelavía
9 Quiavín

The final table and contour map in this section are given to show what the results look like for a complex dialect network, that of Ocotlán Zapotec. The study covered ten test points that fall into four dialectal areas. The largest is San Antonio Ocotlán with its outliers, San Pedro Mártil and Santa Inez Yatzche. The dialectal center that includes San Miguel Tilquipán and Santo Domingo Jalicea is itself a weak outlier to Ocotlán. The linguist, who speaks the Ocotlán dialect, was able to converse in Zapotec with speakers from these subsidiary dialects. The optimal network shows other centers at San Pablo Gúila and Ayoquezco de Aldama. The threshold value is set at 20.

10. This survey was carried out during December of 1967 and January of 1968 by the author and Donald Olson of the Summer Institute of Linguistics. Additional tests were made by Olson and Don Nellis in December 1972. Further testing is necessary.
Figure 3. The Ocotlán Zapotec Network

Key

An  San Andrés  In  Santa Inez
Ay  Ayoquezco  Do  Santo Domingo
Gu  Guelavia  Oc  Ocotlán
Qu  Quiaviní  Ti  San Miguel

The Ocotlán network is represented in Figure 3 as one flat-topped peak and two twinned ones, all rising to a height of 0 from a plain at a contour level of 100. This level represents the separation between the Ocotlán area and surrounding areas such as San Pedro el Alto and Asunción Mixtepec. The three peaks are separated from each other at a level of 55. Ayoquezco and its outlier San Andrés Zabache are separated at a level of 15 while the Ocotlán dialects are separated from the Santo Domingo dialects at a level of 26. In addition, San Pablo Güíla and its outliers San Lucas Quiavini and San Juan Guelavia are grouped together at a threshold of 8, as are the outliers to Ocotlán and that to Santo Domingo. Finally, the threshold levels are 8, 15, 26, 40, 55, and 70. These values were selected for conveniently displaying the dialect groupings. In general, a lower threshold was added only if a broader grouping emerged somewhere in the pattern. The thresholds at 40 and 100 are the sole exceptions; they were included to show how deep the cleavage really is among the dialect areas.
2.4. How the Conclusions Are Used. The reason for making surveys is that administrators can use the conclusions reached. There are several questions that they want answered before they can make decisions: (1) How many separate dialect areas are indicated by the intelligibility results? (2) How feasible is it to select a particular dialect for official or literary use in each dialect area? (3) What priorities suggested by the survey results should be established for allocating available personnel? The answers are based on the interpretation of intelligibility results, also taking into account linguistic and sociological data. The interpretation of these data is crucial; the administrator wants to make only as many allocations as are necessary. He needs to know that these decisions are correct.

2.4.1. Intelligibility tests distinguish dialect groupings. The number of dialects in an area is specified from the results of the intelligibility tests. The criterion for deciding what the distinct dialect areas are is a theoretical threshold that corresponds to the lower level of intelligibility at which communication between a pair of dialects can still be termed adequate. This is the level at which two speakers of separate dialects, both bilingual in a lingua franca like Spanish, would switch to that lingua franca rather than try to communicate in their own dialects. Since intelligibility is non-reciprocal (Point \( j \) understands Point \( i \) better than Point \( i \) understands Point \( j \)), it often turns out that there are several places that easily understand Point \( i \) even though Point \( i \) may not understand any of them very well. In addition, Point \( i \) may not understand any other reference point well. The threshold value, then, groups the set of subsidiary points with one or more reference points and separates each grouping around a reference point from other groups. The groups formed at this threshold are the distinct dialectal areas.

For previous studies, this threshold has been estimated at about 80% intelligibility (or as a communication cost of 20). Since the test scores reflect a certain amount of measurement error and it is probable that minimum threshold values vary from test to test and from place to place, depending upon linguistic and sociological factors that hamper intelligibility, a single value such as 80% is likely to be inadequate as an estimate of this limit. It seems more plausible to express this threshold as a series of values corresponding to a range of intelligibility levels. Preliminary tests suggest that it lies between 75% and 85% (see Section 4.3.3). The threshold might then be calculated at 75%, 80%, and 85% (corresponding to fixed costs of 25, 20, and 15). Either higher or lower thresholds would be plausible for other kinds of tests.

There are several reasons for emphasizing this communication threshold. On the one hand, it is preferable to carry out activities such as educational programs and mass communications in the language each person knows best, as UNESCO, for example, concluded in 1953. To the extent that it is feasible, these activities should be carried out in each dialect. On the other
hand, it is often infeasible to do this because there are insufficient personnel. National interests may also rule against it (Bull 1964). It is therefore necessary to group dialects together in the way that guarantees that all groups will receive communication in a speech form that is acceptable to them. The communication threshold allows one to interpret the data so that both conditions are met. The index of relative dialect extendability is thus a function of this threshold.

Figure 4. The Mazatec Network

2.4.2. A display of dialect groups. The contour maps of Section 2.3.3. display the dialect groupings that intelligibility results imply. This display makes it easy to identify dialect centers and determine the extent of the dialect area associated with each. Each contour level groups a set of dialects at a threshold value that is derived from an intelligibility score. The

11 Bull discusses the need for reducing the total number of vernaculars. However, he seems to advocate imposing an outside language on a group, in spite of group feelings. This seems undesirable to me. In addition, he gives no suggestions about how to decide which languages should be used and which bypassed.
set of possible thresholds ranges between 0 for perfect communication and 100 for no communication at all. The critical threshold range is between 15 and 25. Finally, a group that is separated from other groups by a series of several contours is more likely to be a valid grouping than one that is isolated by only one contour. Dialect distinctness, then, is a function of the range of contour levels that separate groupings. All dialect areas separated by a series of contours that include at least the threshold range probably need to have separate field teams allocated to them. The farther beyond the threshold that the separation goes, the stronger the case for distinct allocations.

Figure 4 is included to illustrate how tentative decisions can be reached from the study of a contour map. It displays the groupings of Mazatec dialects implied by the raw intelligibility scores of the Mazatec survey. (A similar map based on adjusted scores appears in Grimes, 1970.) The most obvious groupings are around the centers of Chiquihuitlán (Cq), Huautla de Jiménez (Hu), Jalapa de Díaz (Ja), and San Jerónimo Técoatl (Te). These are all separated by at least the threshold, i.e., by contours with values of 25 or greater. (Jalapa is a little closer to Huautla than the others are.) The assignment of three other dialects, however, is dubious; they are grouped by contours just within the critical range. Soyaltepec (So) is a weak outlier to Te, Ixcatlán (Ix) is a weak outlier to Hu, and Ayautla (Ay) is a somewhat more definite outlier to Hu. The rest of the dialects are grouped with either Hu or Te at contour levels easily within the critical range: San Miguel (Mg), Tenango (Tg), Independencia (In), San Mateo (Mt), Santa María Asunción (As), and Mazatlán (Mz) with Hu; and San Pedro (Oc), Santa Cruz (Ac), San Lucas (Zq), Chicholutla (Ch), Mazatzongo (Ms), San Lorenzo (Lo), San Antonio (An), Santa Ana (St), Zacatepec (Zc), and Huehuetlán (Hn) with Te.

For the Mazatec area, then, the map indicates at least four clearly marked dialect areas: Chiquihuitlán, Huautla, Jalapa, and San Jerónimo. The distinctness of the Chiquihuitlán and Jalapa dialects is based on uniformly low intelligibility scores. The distinctness of the Huautla and San Jerónimo groups, however, is based on low intelligibility between the outliers of the two groups, rather than on low intelligibility between the centers themselves. Surprisingly, intelligibility between these centers is high: San Jerónimo subjects scored 76 on the Huautla tape; Huautla subjects scored 89 on the San Jerónimo tape. Yet most of the outliers of San Jerónimo scored significantly low enough on the Huautla test tape that their effect en masse separates the groups profoundly.

12. The raw scores that subjects from Chiquihuitlán obtained on various reference point tapes were as follows: Huautla 38, Ayautla 37, Soyaltepec 29, Ixcatlán 20, and Jalapa de Díaz 10. Jalapa speakers obtained the following scores: Huautla 73, Ixcatlán 62, Soyaltepec 51, San Jerónimo 46, and Mazatlán 35. Other scores show that none of these reference points can be considered outliers of Jalapa: Ixcatlán subjects scored 64 on the Jalapa tape; Soyaltepec subjects scored 43, and Huautla subjects scored 35. Although the tape from Chiquihuitlán was not tested elsewhere, it is highly unlikely that any other Mazatec dialect could be considered an outlier to it. In the first place, at least for the dialects of Huautla and Ayautla, nonreciprocity in intelligibility is unlikely to favor Chiquihuitlán; i.e., Huautla subjects would probably score lower on the Chiquihuitlán tape than Chiquihuitlán speakers scored on the Huautla tape. (This would fit the pattern observed consistently during the study.) Furthermore, for any case in which nonreciprocity might favor Chiquihuitlán, it is unlikely that the degree of it will be great enough to allow another dialect to become an outlier to Chiquihuitlán under the stated threshold conditions. Finally, the distinctness of the Chiquihuitlán and Jalapa dialects is based on one more assumption: that the choice of reference point test tapes was exhaustive, i.e., no potential center was overlooked from which communications could be made to them.
It seems necessary to recognize at least one more dialect center for the Mazatec area. Soyaltepec and Ixcatlán are distinct enough from both San Jerónimo and Huautla that they probably fall below the threshold of effective communication from either. Ixcatlán speakers do not understand the Soyaltepec dialect very well (they scored 70 on the tape from Soyaltepec); but it is not certain whether Soyaltepec speakers understand the Ixcatlán dialect since the Ixcatlán tape was not tested there. The two dialects are close enough geographically that it would seem that there should be high intelligibility in at least one direction, so that Soyaltepec might understand Ixcatlán better than 80%. On the other hand, the difference between the tone systems may be sufficient to cause low intelligibility in both directions (Eunice Pike, personal communication). In that case, personnel should be allocated to both dialects. In short, we know which test is missing to make our picture of the Mazatec area complete.

2.4.3. Feasibility. Although intelligibility results show the number of distinct dialect centers in an area, it does not necessarily follow that these results directly indicate allocation needs. Even if there are sufficient personnel to make an allocation to each dialect center, it may be infeasible to do so for other reasons.

In the first place, just because a person can understand another dialect does not guarantee that he will accept materials written in that dialect. It probably does not hold even when oral communication is involved, as in radio broadcasting, but it might be especially true when the two dialects cannot be written the same way. There are also instances in which speakers of one dialect readily understand oral materials from another but refuse for purely nonlinguistic reasons to accept written materials. In these cases it is necessary to add another allocation to the pattern initially suggested by the intelligibility results. For example, the Tzotzil data indicated that the six towns studied would fall into three dialectal areas: Huixtán, Carranza, and the area encompassing Chamula, Zinacantan, Chenalhó, and San Juan del Bosque. However, the hostility between Chamula and the other three towns is such that it was decided to treat Chamula as a fourth Tzotzil center.

On the other hand, even though a dialect area may be clearly distinct from the other dialect areas, it may not be worth the trouble to establish an official or literary use for that dialect because of extensive bilingualism with the national language, or because of a demonstrably rapid shift in language use toward bilingualism.

Some groups feel so inferior about their language that they rapidly shift toward the use of a national language. Parents may be reluctant to teach their children to speak the vernacular. People may even deny that the vernacular is spoken at all. If these behavior patterns actually reflect the attitude of the entire group, or of a large enough proportion of it, then it will probably be infeasible to do more than carry out terminal linguistic studies. On the other hand, if only a few people hold these attitudes, even though they may be politically powerful, it will probably be necessary to allocate there.

To determine the feasibility of allocating field personnel to a dialect area, it is therefore necessary to supplement intelligibility data with sociological and linguistic observations. Linguistic data, for example, allow one to estimate how many of the subsidiary dialects can use the orthography devised for the primary dialect without making any changes in it (see
Section 4.3.4). In addition, they allow one to see how much the lexical items and syntactic patterns of each subsidiary point differ from those of the primary dialect and how much they differ among themselves. If they differ sufficiently, it may be necessary later on to adapt the written materials for some of the subsidiary dialects. This, in turn, may require another field allocation. (Appendix G illustrates a case of a pair of dialects that cannot use the same written materials, partly because of phonological differences.)

In some cases it may be more desirable to stimulate greater acceptance of the primary dialect than to adapt written materials to subsidiary dialects. It will therefore be necessary to estimate whether people at the subsidiary points will accept the speech of the dialect center more readily if they hear it more often. To estimate this, sociological data must be considered, such as the speakers' awareness of differences between their dialect and others, their attitudes regarding these dialects and the people who speak them, and their political and economic ties with other dialect areas. These data are obtained by making observations and using questionnaires (see Section 8.3.3 and Appendix E). Enough data should be collected to determine whether attitudes expressed or behavior patterns observed are characteristic of an entire group or merely of certain individuals.

Finally, for each dialect area to which field personnel might be allocated, it will be necessary to evaluate the degree and stability of bilingualism with the national language, as outlined in Chapter 8.

2.4.4. Priorities. Relative priorities must be assigned to dialect centers in the event that there are not enough personnel to allocate a field team to each one. In assigning priorities, all the dialect centers are ranked, regardless of how many field teams are available. There priorities are assigned with reference to definite purposes for making allocations. For example, a dialect which is low in priority for initiating a literacy program because all of its speakers are bilingual with the national language might be high in priority for making linguistic studies because it is rapidly dying out.

Additional criteria for revising the intelligibility ranking must be selected. Most of them will be sociological and demographic. Possible criteria include the following: (1) the degree to which speakers of each dialect are bilingual (determined by age groups and by sex), (2) the occasions on which people in each dialect use the national language and their own language, and the apparent stability of this relationship, (3) the relative distinctness of each dialect from others in terms of intelligibility, (4) patterns of economic and political dependencies among the dialects, (5) the relative size of the dialects, and (6) local circumstances indicating that one dialect is more favorable to an immediate allocation than another. An additional criterion, though of a different sort, is whether a program was once started in a dialect area but subsequently left unfinished. Still other factors will undoubtedly have to be taken into account in each situation.

In each situation, priorities are assigned after all the dialect areas have been compared in terms of the entire set of criteria selected. It is likely that no single criterion by itself is adequate. For example, it is probably unreasonable to consistently assign high priorities to large dialectal areas and low priorities to small ones. It does seem reasonable, however, to assign higher priority to the larger of two areas if both are equivalent in terms of other criteria such as bilingualism, dialect distinctness, and local
circumstances. To cite another possibility, a low incidence of monolingualism may mean vastly different things for two groups. For a small group the proportion of speakers who are monolingual may be confined exclusively to those above forty years of age. On the other hand, the same proportion of monolinguals may be distributed among all age groups in a large dialect. In either case, the attitude of monolinguals and bilinguals toward the use of the vernacular language will be relevant. Thus, several criteria applied together allow one to compare dialect areas so that a reasonable order of priorities can be associated with them.

For the Mazatec dialects mentioned in Section 2.4.2, an order of priority can be stated on the basis of three characteristics: (1) the relative degree of bilingualism that each dialectal group was presumed to have with Spanish; (2) the relative degree of intelligibility between each dialectal group and Huautla and San Jerónimo, the principal dialects; and (3) the size, or range, of each dialectal area. The survey results showed that the Huautla dialect was most widely understood, followed closely by the San Jerónimo dialect. In addition, the Jalapa and Soyaltepec dialects were almost equal in size, but Soyaltepec was somewhat more closely associated with San Jerónimo than Jalapa was with Huautla. The intelligibility scores showed that Chiquihuitlán was clearly distinct from the other dialectal areas. However, it is small (one town with about three thousand speakers) as opposed to the twelve towns in the San Jerónimo area (about 15,000 speakers). Furthermore, superficial observations seemed to indicate a high degree of bilingualism for the Chiquihuitlán community, while bilingualism in both San Jerónimo and Jalapa de Díaz was clearly known to be low. Finally, bilingualism in the Jalapa area seems to be less than that in Soyaltepec and is probably more stable. From these observations, priorities can be assigned in descending order as follows: (1) Huautla, (2) San Jerónimo, (3) Jalapa de Díaz, (4) Chiquihuitlán, and (5) Soyaltepec.

Additional information about any of the factors mentioned above or even the inclusion of additional factors, could force changes in this ordering. For example, additional knowledge about the degree of bilingualism in Chiquihuitlán substantially alters the initial conclusions from the study. Allan Jamieson, who is now studying the Chiquihuitlán dialect, observes that people in the area are not as proficient in Spanish as we had previously thought. He has also learned that a significant number of Chiquihuitlán speakers have formed colonies in one or two outside communities and that the emigrants are still speaking Mazatec (personal communication).

13. Gudschinsky suggests that timing could be a critical factor in assigning priorities. For example, it would have been better to have prepared literacy materials for the Soyaltepec dialectal area before the Miguel Alemán dam was built, because afterwards the town was flooded and lowland Mazatecs from various dialectal areas were mixed together in the relocation program (personal communication).
3 HISTORICAL BACKGROUND

3.0. In this chapter we consider some of the main studies of intelligibility testing that have been carried out since C. F. Voegelin and Zellig S. Harris first suggested the method. These are cited to show how the results have been obtained and how they have been interpreted. The discussion is organized as follows: (1) the Voegelin-Harris approach, (2) Wolff's critique of their approach, (3) Crawford's adaptation of it for use by the Summer Institute of Linguistics, and (4) subsequent developments.

3.1. Voegelin and Harris. Voegelin and Harris provided the impetus for collecting intelligibility data when they suggested four ways to distinguish between language and dialect (1951). Two involved the assessment of intelligibility. The first method consisted of asking an informant what he perceived the dialect relationships to be; this allowed one to make a preliminary grouping of languages and dialects. The second tested a subject's comprehension of oral materials. The procedure that they suggested required that a subject translate portions of narrative speech into a language spoken by the investigator. The subject's translation was evaluated by comparing it with a translation already prepared by members of the speech community where the sample was taken. The evaluation of the translation was first a measure of how well the speech samples were understood. Secondly, it revealed factors which caused difficulty in a subject's understanding of a particular text.

Regarding intelligibility per se, Voegelin and Harris pointed out that it could be non-reciprocal as well as reciprocal. Although they did not elaborate on this, they did link non-reciprocity to extralinguistic factors (1951.325). Furthermore, they distinguished between neighbor intelligibility and mutual intelligibility. Apparently they felt that neighbor intelligibility was due to extensive contact between two language groups; mutual intelligibility resulted from a close genetic relationship. Finally, even though they did suggest that the results of testing an informant's comprehension from another dialect could provide an index of dialect distance, they felt that the primary use of this kind of data would be to see which dialect features
would be the same for various groups of speech communities, a kind of dialect geography (1951.329).

3.1.1. Olmsted. Olmsted used a questionnaire to ascertain intelligibility in a case of reported non-reciprocal intelligibility between two California Indian languages, Achumawi and Atsugewi (1954). The principal difficulty in obtaining this information was that there were few living speakers of either language. Olmsted’s criteria for selecting informants were that each be a native speaker of the language and that he use it with at least the same degree of fluency and frequency as he uses English. He chose these criteria because few informants younger than fifty years of age were even functionally bilingual with the Indian language. The questions Olmsted asked related to both an informant’s personal contact with the language of the other group and his knowledge of old-timers who might have spoken the other language.

The reported state of intelligibility was that the Atsuwegi understood Achumawi but not vice versa. Of the ten Achumawi speakers interviewed, only one claimed to have understood any Atsugewi and he attributed this to having worked with the Atsugewi on a construction project. Furthermore, only four Achumawi reported knowledge of old-timers who had known any Atsugewi, and those were said to have lived with the Atsugewi. On the other hand, three of the five Atsugewi claimed personal knowledge of Achumawi, and all of them knew old-timers who spoke the other language. The Atsugewi informants attributed this knowledge of Achumawi to intermarriage. Olmsted, then, concluded from the study that the sociocultural relationship between the two groups was the basis for non-reciprocal intelligibility between the two languages, rather than strict linguistic properties of either one. He added that, in other cases, however, non-reciprocal intelligibility might be due solely to linguistic factors.

3.1.2. Hickerson, Turner, and Hickerson. Hickerson, Turner, and Hickerson applied the second method suggested by Voegelin and Harris to determine relationships among seven Iroquois dialects (1952). They used two tests. one for measuring global understanding of discourse, the other for measuring detailed understanding of shorter stretches of speech. The first test required a translation of an uninterrupted two-minute text. Responses were scored on the basis of their correspondence with ideas contained in the text. The second test consisted of a thirty-second stretch of speech played in short sections. A translation of each of these sections was then required of the subject. His responses were scored as incorrect or one-third, two-thirds, or fully correct. A comparison of scores for the two tests showed that results were more nearly alike when the level of intelligibility was fairly low. When intelligibility was shown to be high, the scores were most different. Generally, the scores from the second test were higher. The scores were used as criteria for distinguishing between dialect and language. Language boundaries were set at 25% intelligibility and dialect boundaries at 75%. The decision was considered to be arbitrary.

Three controls used in the study deserve mention. First, the investigators attempted to select for testing subjects who had not learned another dialect than their own. Second, they became aware in the testing that a subject’s ability to translate a text about a myth or ritual could be explained more readily in terms of his knowledge of folklore than his ability to understand
the dialect where the text was collected. Therefore, they eliminated this class of text from their test materials. Finally, the subjects' translations of the test materials were recorded mechanically. Thus the investigators were able to study the translations and to evaluate them accurately after the field work was completed. The significance of these controls for our present methodology should be apparent.

3.1.3. Pierce. A second study using intelligibility data was made by Pierce among speakers of a group of Algonquian languages (1952). He attempted to objectify the measure of dialect distance based on the percentage of information transfer (intelligibility). He also sought measures of the degree of relationship among four speech communities.

Pierce followed Hickerson and Turner in using only non-folkloristic, spontaneous narratives for testing. He used only one test, which roughly corresponded to the Hickerson-Turner test II. This test consisted of a one-and-a-half minute stretch of speech taken from one of the narrative texts. The abstracted portion was in turn divided into shorter segments which were separated by natural pauses. The test consisted of from fifteen to nineteen of these short sections. The subject was asked to translate each of these sections. The investigator would repeat any of these sections if the subject requested it.

For scoring responses, Pierce constructed a "standard grading translation." For each narrative text he obtained translations into English from both the informant who gave him the text and from two other speakers of the same community. He then divided the translations into sentences and analyzed each sentence into units. From the variants of the corresponding sentences of the three translations he determined a common denominator sentence for each sentence of the narrative. For example, the translations of a sentence might be as follows:

"They used to get me up early."
"They wake me up."
"They wake me up early in the morning." (1952.206).

The common denominator sentence with its units would be (1) they (2) awaken (3) me (4) early. Pierce could then grade a subject as correct, incorrect, or half-correct for all units in each sentence of the text.

In addition, Pierce noticed an uncontrolled factor in the scoring of responses; i.e., a subject's raw scores could be greatly affected by his ability to understand English. An adjustment was made by dividing the subject's score on materials from an external dialect by his score on materials from his own dialect. In other words,

\[
\text{adjusted score } Y = \frac{\text{score on other dialect}}{\text{score on own dialect}}
\]

He felt that this adjustment would adequately cover the effects of subjects' differential ability in English on the scores, though he admitted that the

1. This seems especially true when intelligibility is high enough for a subject to recognize the basic content of the text even though he may not be able to understand many of the particular features it contains.

2. Pierce considered that the error due to a subject's deficiency in English would be a constant value in hampering scores on all of the tests, and therefore, relative intelligibility levels for all tests would be valid. The adjustment was intended to be a first approximation to the actual level of intelligibility. The assumption that the error value is
resultant score might still be considerably below the actual level of intelligibility between fluent speakers of the dialect in question.

Finally, in order to obtain an objective measure of dialect distance, he computed a percentage of mutual intelligibility between any two dialects. The computation is as follows: people from dialect A score 85% on a text from dialect B; people from dialect B score 75% on a text from dialect A. The arithmetic mean, or average, of the two scores is 80%. This value of 80% is considered the percentage of mutual intelligibility between dialect A and dialect B. It is also assumed to be an objective measure of the linguistic difference between the two dialects. Apparently Pierce felt that both intelligibility and linguistic difference were symmetrical relations between pairs of dialects.

In a later pilot study, Pierce (1954) investigated the overlap between intelligibility results and glottochronology for estimating both dialect distance and time-depth for Crow and Hidatsa. He noted that intelligibility testing and glottochronology have largely separate domains in the estimation of dialect distance. Intelligibility testing finds its province with those groups which retain mutual intelligibility in spite of linguistic divergence, whereas glottochronology is useful where the time-depth between dialects is more than 500 years. He felt, however, that there was an area in which the two approaches overlapped and the results could be used to supplement one another.

The test was composed of three items: a 200-word Swadesh list of non-cultural vocabulary, a 50-word list of nouns, and two five-minute texts. For the word lists, the test was scored on the correspondence between a subject's translation into English of a particular item and the original translation of that word into English by the informant who gave Pierce the lists. If the translations were identical, the response was considered correct. If the translations were not identical, the response was considered incorrect.

The results for the pilot study (which used only one Crow informant and one Hidatsa informant) showed an intelligibility level of 36% from Crow to Hidatsa and 26% from Hidatsa to Crow for the 200-word Swadesh list. The percentage of mutual intelligibility for the list was rated at 31%, the average of the two scores. For the 50-word list, the Hidatsa subject scored 24% on the list of Crow nouns, and the Crow subject scored 20% on the list of Hidatsa nouns. The calculated level of mutual intelligibility was therefore 22%. Thus there is a considerable difference between scores on the 200-word list and scores on the 50-noun list. The intelligibility between the two dialects as measured by the texts was negligible. Both texts were approximately five minutes in length, and each test subject was required to translate the text of the language he did not speak. Since the Hidatsa subject correctly translated only eleven words of the Crow text into English and the Crow subject correctly translated only one word of the Hidatsa text, Pierce rated the probable degree of mutual intelligibility between the dialects at less than one percent. A comparison of the intelligibility results showed 31% mutual intelligibility on a word list with 59% cognates as constant for each subject is a useful one, but it is likely to be unrealistic; other variables probably interact with the subject's deficiency in the language spoken by the investigator (and thus used to evaluate the responses).

3. Note that the term "dialect distance" assumes that intelligibility is the same in both directions. For evidence that this assumption is misleading, see Sections 4.3.1 and 4.3.2.
opposed to less than 1% intelligibility on a text. From this, Pierce suggested that Swadesh's prediction about greater retention rates for certain items may be correct.

He also computed time-depths for the 200-word list. For one computation he used the percentage of cognates found among words correctly understood by the two subjects. This computation yielded a time-depth of 1,967 years. For the other computation he used the percentage of cognates determined by linguistic criteria. This computation yielded a time-depth of 1,182 years. Thus, the intelligibility-based computation exceeded the linguistically-based one by approximately 66%.

3.1.4. Biggs. A further study of intelligibility was carried out among speakers of Yuman languages by Bruce Biggs (1957). In this study he compared the method with others in order to test the validity of intelligibility testing as a measure of dialect distance. Though the method was essentially that of earlier studies, some of his procedures and assumptions are interesting.

Biggs constructed three sub-tests for each language in order to increase the lexical ranges covered by the test materials. For one language these sub-tests consisted of a moralistic text of a pow-wow, an autobiographical text, and a set of isolated sentences translated into the language. The first text was eventually disregarded because of difficulties encountered in scoring it.

The method of test administration he used was quite flexible. Several of the tests were given to groups of subjects. Usually one of the subjects acted as spokesman for the group. Help from the other members was accepted as long as it was not given by a person bilingual in the test language. Each subject was first given an oral introduction and was then tested on his native language. The rest of the tests were not given to him in a fixed order, but rather in an order alternating between easy tests and difficult ones. Repetitions of a phrase to the subject were allowed, and subject responses in English were written down instead of being recorded on tape. Each subject was asked to identify the language of each test. The scores of those who could do so, and were judged to have considerable prior knowledge of some test language, were discarded for that particular test. During the testing, Biggs made no indication to the subject of an incorrect response, but occasionally he acknowledged a correct response to encourage a subject, especially if the test language was difficult for him to understand.

In regard to scoring responses, Biggs notes that there are actually two problems of possible language deficiency in obtaining a master translation which is the basis for evaluating a subject's translation of a text. The first is the investigator's own lack of knowledge of the test languages. As already noted, the other is the subject's level of ability in English. To aid the subject, Biggs decided to ignore grammatical categories such as person, number, and mode. He restricted his evaluation to content words in the subject's translation.

The index of intelligibility was taken to be the highest score obtained from a sample of subject responses to a particular test. Other studies used the mean of the subject responses as the index. The assumption that Biggs used was that purely linguistic structures would set an upper limit on intelligibility between pairs of languages. Thus the highest score would most
closely approximate this limit, and all lower scores could be attributable to psychological factors such as intelligence, motivation, or translation ability.\footnote{The validity of this assumption is dubious. First, it is not standard statistical practice to use the extreme value of a sample of measures as the representative figure for that sample. Second, the upper limit of intelligibility is not necessarily set by linguistic factors. It may be set by purely sociological considerations in some cases. See the following section on Wolff's criticisms, Section 4.3.7, and Appendix J.}

In a table comparing cognate percentages with intelligibility scores, Biggs shows that intelligibility was 70\% or higher when the cognate percentage was above 85\%, but that it dropped rapidly as the percentage of cognates dropped below that point. Between a pair of languages with 57\% shared cognates, the index of mutual intelligibility was only 12.5\%. The relative dialect distances indicated for pairs of languages by the two methods were basically in agreement, however.\footnote{More specifically, a scattergram of the data shows two widely separated clusters of pairs of scores: cognate measures in the 57-63\% range are paired with intelligibility scores ranging between 10.5\% and 19\%, while for those between 85\% and 95\% the intelligibility scores ranged between 72\% and 95\%. In addition, an isolated intelligibility score of 49\% would probably have been paired with a cognate count of about 75% had Biggs obtained it. All this suggests a linear relationship between intelligibility and shared vocabulary, as has since been substantiated by Ladesfod (1971) and by Bender and Cooper (1971; see Section 4.3.7).} Biggs also noted general agreement between the intelligibility results and the method which counts structural similarities between languages. In comparing the test that used an elicited text with that which consisted of elicited sentences, Biggs felt that the latter would often be easier to construct and would probably be more sensitive to differences between closely related dialects. Finally, he comments that intelligibility scores could serve to distinguish between separate languages and dialects of the same language, but that there is a point at which the decision is arbitrary.

3.2. Wolff's Criticisms of the Early Studies. The validity of intelligibility testing as a measure of linguistic proximity and the implications of interlingual communications were discussed by Hans Wolff (1964). His treatment of uncontrolled factors in the testing and other inadequacies of the method can be summarized as follows: (1) The test seems primarily to measure a subject's ability to translate from one language to another, but translation ability involves more than intelligibility and thus is not a fair measure of dialect distance. (2) The language in which the subjects are required to make their responses is a third language, usually the investigator's; therefore the subject's proficiency in that language is a factor in his response. (3) The subject being tested may be reacting against the investigators, their methods, or the use of their language. Thus, if the method is to yield valid results for measuring linguistic affinities, not only must the subject have no knowledge of the test language prior to the testing situation, he must also be proficient in the use of the target language. He further must have no adverse reactions to either the process of translation or to the investigators, their methods and equipment, and the use of their language. Finally, Wolff notes that for areas with a high degree of bilingualism, the test cannot distinguish between intelligibility due to learning of another dialect and intelligibility due to language similarity. He also states that the test does not explain the significance of non-reciprocal intelligibility.

Wolff concluded that intelligibility would not serve as a valid measure of linguistic proximity. However, he considered that data on intelligibility would
be useful for orthography work, which requires a degree of language standardization. He found two phenomena that caused him to explain intelligibility in terms of sociological factors rather than linguistic ones. The first was a very low correlation between other measures of linguistic similarity and the degree of intelligibility claimed for them. The other was the consistent measurement of non-reciprocal intelligibility for these closely related dialects. The measurements were made both by asking and by testing informants whenever possible.

In considering some cases of non-reciprocal intelligibility, Wolff illustrated economic and political conditions and intergroup attitudes that influenced communication. In one case, the larger and more prosperous group claimed no intelligibility with a smaller group which was quite similar linguistically. On the other hand, speakers of the smaller group acknowledged intelligibility with the language of their prosperous neighbors. Wolff noted that the smaller group had been ruined economically by a shift in the course of a nearby river which cut off their access to a major harbor. In addition, the smaller group had lost their political power during a rebellion several decades previously. He suggested that the intelligibility data reflect the political and economic conditions of the two groups.

Wolff also mentioned the case of languages with high functional values. In this case the person who speaks such a language has certain advantages over one who does not. These languages are most widely understood throughout a multilingual area regardless of the genetic relationships among all the languages.

From the above two cases and several others, he suggested that non-reciprocal intelligibility exists in a bilingual area because sociological factors permit it. Conversely, he suggested that the lack of non-reciprocal intelligibility indicates that there are elements in the sociological pattern that preclude intergroup contact. In either instance, linguistic proximity plays a minor role. Finally, in the case of non-reciprocal intelligibility the language of the dominant group, or the language with the higher functional value, will be the most widely understood of the pair.

3.3. Crawford's Adaptation.

3.3.1. In relation to literacy. Crawford agreed with Wolff's criticisms about the interpretation of intelligibility data and the primacy of non-linguistic factors in interlingual communication (1967). Since intelligibility can be more directly determined by language dominance, interdialectal learning, and other sociological considerations rather than by linguistic features which indicate genetic relationships, he expected the results of intelligibility testing to be different from those obtained by comparative linguistic methods. Rather than reject the method of intelligibility testing, he felt that it could be quite useful if interpreted differently from the Voegelin-Harris viewpoint. Specifically, sociological factors that tended to invalidate the method as a measure of dialect distance were particularly relevant to the literacy problem, which consisted of (1) estimating the expected useful range of materials prepared at some point within a given language area, and (2) comparing a set of points within that area as potential centers for

6. Cf. the comparisons made by Pierce (1954) and Biggs (1957) with regard to intelligibility scores and glottochronology.
distribution of literacy materials. Therefore, he interpreted intelligibility data as an index of the extendability of dialects in a communications network.

3.3.2. How an index of extendability is derived. Crawford felt that an index of dialect extendability could be derived from the measure of intelligibility. Intelligibility results give only a first approximation to the communication pattern in the sense that they show the maximum geographic domain of each of a small number of centers within the language area. The first interpretation of the results is that, if there is no adequate communication between a pair of dialects and if a third dialect cannot serve as a common source for guaranteeing adequate communication to both of them, then they will be assigned to distinct centers of communication, or dialectal centers. Adequate communication is defined in terms of a threshold value (or range of threshold values) of minimum intelligibility. This range of threshold values is empirically determined for each study and each kind of testing device. Particular sociological factors, moreover, might lower the limit of maximum extendability of a dialect without being directly revealed by the test results. The results merely indicate the presence or absence of intelligibility between dialects. Finally, some factors such as the acceptability of written materials from one dialect by speakers of another are outside the scope of the test since the basis of the test is the comprehension of oral, rather than written, materials. Since dialect extendability may be more or less influenced by variables that do not play a role in determining intelligibility, and since intelligibility results indicate the maximum domain of extendability, the true domain is likely to be smaller than that indicated by intelligibility results per se. Additional kinds of data therefore are taken into account in order to make approximations converging on the true domain of a dialectal center.

The approach to the initial questions for literacy programs, as Crawford saw it, thus required that three kinds of data be integrated to derive the index of extendability: (1) Intelligibility tests would be administered at various test points within a language area. The results of this testing would provide indices of the maximum extendability to other dialects from centers in that area. (2) Linguistic data would be collected to correlate with the intelligibility results. These data could validate the test results by revealing linguistic phenomena that might have caused a breakdown in comprehension. These data might also provide an inventory of linguistic features suitable for subsequent literacy purposes. (3) Ethnographic information would be collected for correlation with the first two sets of data. This information would be used to validate intelligibility results by noting, in the case of non-reciprocal intelligibility, specific indicators that a particular dialect is a prestige dialect, or one with a higher functional value than others in the area. This information might indicate other factors that would inhibit communication between a pair of dialects, or it might modify the domain of a dialectal center implied by the test results by demonstrating that speakers of a sub-dialect would refuse written materials from the primary dialect even though they understand the primary dialect quite well.

3.3.3. Expected field conditions. Both the character of the populations from which data would be obtained and the availability of investigators put severe limitations on the method to be employed. First, to be useful the test would have to be easy to construct and administer. Next, it would have to
be applicable to all members of the population. Since many are illiterate, monolingual, or both, the test would have to be given orally and in the local dialect of each test point. Since many of the people who might be tested have never taken a test in their lives, it would be necessary to train each subject to take the test. It would have to be simple enough so that neither the subject's lack of an academic background nor his uneasiness in a novel situation (including earphones and talking machines) would invalidate his performance. Finally, the local people would have to be induced to cooperate, both town officials and potential subjects, even in towns where many would be suspicious of outsiders with tape recorders.

3.3.4. The first test: the Mixe study. Crawford developed the first intelligibility test of the type used in Mexico. He began in the Mixe area where he was personally familiar with several of the dialects. He selected ten points in the area as reference points. At each he elicited a text of unspecified content for the preparation of three sub-tests: a content repeat, a sentence repeat, and an evaluation of relative order of difficulty of the test tapes heard. In the content repeat test, the subject was asked to recapitulate the first two minutes of a text in his own words. In the sentence repeat test, every third sentence of the rest of the test was played to him and he was asked to repeat one sentence at a time. Thus Crawford utilized tests based both on an elicited, free narrative text and on individual sentences. As the finale, the subject was asked to rank each of the four test tapes he had heard in their order of relative difficulty for him. The test was administered to ten subjects in each of nineteen towns throughout the Mixe region. Each subject was instructed orally on how to take the test. He listened to the test tapes through earphones, and Crawford evaluated his responses on a five-point scale.

There were three features peculiar to the Mixe testing. First, two equivalent test tapes were constructed at each of the ten reference points. Thus Crawford prepared two sets of reference point test tapes to check internal consistency. He chose to make the test by alternating the sets of test tapes at each town along his route. The result was that each pair of geographically closest towns was matched with the two sets of reference point tapes; only one set of tapes was used at any test point. Then, by comparing the scores obtained on the test tapes at each pair of test points, Crawford was able to estimate the reliability of the two sets of measurements. The second feature was that Crawford attempted to measure the intelligibility between each test point and all the reference points by playing the entire set of reference tapes to subjects at the test points. He limited the number of test tapes that were to be given to any subject to four of the ten, but distributed sets of four tapes among successive subjects. Apparently the tapes were numbered from one to ten; tapes numbered 1-4 were presented to the first subject, tapes 5-8 to the second subject, tapes 9-10 and 1-2 to the third subject, and so on. Thus responses were obtained from four of the ten subjects on each test tape. The third feature was that Crawford scored responses in terms of how well a subject

7. The validity of Crawford's method rests on the assumption that both populations are identical with respect to how well they understand outside dialects. This assumption is of doubtful validity. While Crawford's choice of matching the towns geographically increases the likelihood that both populations are identical in this respect, it would have been better to have used both sets of reference tapes at each test point.
repeated what the text or sentence said, rather than in terms of whether he understood the content of the test materials or not.

Regarding the comparison of the sub-tests, Crawford found that the content-repeat test usually gave lower scores than the sentence-repeat. However, for highly intelligible dialects, the sentence repeat became so easy that a subject’s response seemed more like mimicry than a test of intelligibility. Crawford decided that the rank-ordering sub-test was unsatisfactory because he had to force responses from subjects unaccustomed to questions of that kind. The results that he did get showed little correlation with the scores of the content-repeat test.

3.4. Subsequent Developments. Later studies followed a design similar to that applied in the Mixe study, but in each innovations and refinements were made in the method. In the following sections, several studies subsequent to the Mixe are treated in roughly chronological order.

3.4.1. The Mixtec Study. Several major changes in the method were adopted for the Mixtec study carried out by Bradley (1967). These changes resulted in a measuring instrument which is actually the basic one that has been employed in all subsequent studies. In the first place, after making some preliminary tests, Bradley decided to use only tests that were based on recordings of free narrative texts. He excluded tests based on individual sentences. In addition, rather than test the entire set of reference tapes at a test point by dividing them into small sets, each of which was to be administered to only a few of the subjects. Bradley tested each subject on the same set of reference tapes. A third change concerned the subject responses: subjects were no longer required to repeat the content of a text. Instead, a subject's understanding of a particular text was estimated by the number of correct answers he gave to ten questions based on specific statements in that text.

The test battery employed in the Mixtec study consisted of an introduction tape, a hometown test tape, and four or five reference point test tapes, as in present studies. The introduction tape contained only a formal explanation to the subject. It did not provide him with a sample test for instruction. Some control over the content of the texts was maintained. Autobiographical materials were usually chosen. A few multiple-choice and yes-no questions were incorporated into the tests. Most, however, were short-answer, content questions. All of the questions for a set of test tapes were translated into the local dialect at each test point. Each test tape was constructed by copying the first two-minute portion of the original recording of a text. The questions over the copied portion of that text were then recorded on tape immediately following the text. Thus the subject would hear the continuous text, and he would be required to remember what he had heard in order to answer the questions referring to that text.

The tests were administered through earphones to the subject. The test tapes considered to be most difficult were bracketed by those estimated to be easier for the subject to understand. The hometown test tape was always the first test tape presented to the subject; for all subjects in a sample of ten, the order in which the reference point tapes were presented was identical. Scores for subjects who failed to score at least 50% on the hometown test tape, who came from the same house as another subject, or who failed to complete the set of test tapes were disregarded for the study. All questions missed by more than half of the subjects on their own
hometown test tape were disregarded in computing scores. This adjustment was carried out for all test points where that test tape was played. Bradley estimates that 15% of the questions for all test tapes in the study were disallowed in this way (1967:11).

Because of the particular test format employed in the Mixtec study, two or three factors greatly influenced the scores obtained. The first was the lack of previous schooling for many of the subjects. They had to learn how to take the test at the same time that they were being evaluated on their performance. This showed up in scores on hometown test tapes, since these tapes were invariably the first ones presented to the subject. Thus scores on hometown test tapes were occasionally lower than scores on some reference point test tapes. The mean scores for hometown test tapes ranged from 65% to 92%, with a mean of 84%. Bradley adjusted these scores by adding 15% to those which were between 60% and 69%, 10% to those between 70% and 79%, and 5% to those between 80% and 89%. The range of adjusted mean scores was then 80% to 94%, with a mean of 89%. The other factors influencing the scores included lapse of memory and varying lengths of attention span. These factors were especially noticeable for test tapes that were only marginally intelligible with the subject's own dialect.

While the Mixtec study was in certain respects a pilot study, the version of the measuring instrument was applied over a wide area geographically and to a rather large number of subjects (about seven hundred for the entire study). Its results were treated and interpreted in essentially the same manner as the results obtained by later and more refined versions. For interpreting the scores, Bradley estimated the threshold for communication between dialects to be 70% (a point estimate based on the mean score for a sample of ten subjects). He considered the range between 70% and 80% as marginal for intelligibility, and above 80% as signalling adequate communication.

3.4.2. The Tzotzil study. The Tzotzil study was carried out by Crawford and Stoltzfus. Some modifications were introduced into the Mixtec test design (Stoltzfus 1967). These included the addition of a sample test to the introduction tape, prefaces to the questions for each test tape, and the rotation of the order in which tapes were presented to succeeding subjects. All of these modifications were introduced in order to control the lack of question-answering experience on the part of many of the subjects.

The sample story was about thirty seconds long. It was dubbed from a text that was elicited specifically for constructing a sample test tape. Three questions were formulated over the content of this sample story. The sample test was included to provide the subject with an initial, ungraded experience in taking an intelligibility test. It was designed to minimize the error due to learning that would be shown in the score of the first test tape the subject received, generally the hometown test tape. This innovation was made during the latter half of the study.

The preface to the questions merely reminded the subject that he had heard a story and was now going to be asked to give certain information regarding what he had heard. This was designed to gain his attention and to provide him with an additional orientation for taking the test. "Yes-no" questions were allowed, but the number was held constant for all test tapes.
The rotation of the order in which tapes were presented between subjects was introduced to control the disconcerting phenomenon of a subject's registering higher intelligibility scores on reference point test tapes than he did on his own hometown test tape. Since the reason seemed to be that the subject was still learning how to take the test, and since his performance improved with each test tape he completed, it was decided to rotate the order in which succeeding subjects received test tapes. In this way, the cumulative effect of learning by the subjects would be spread equally among the scores for all of the test tapes used with a sample of subjects.

A final contribution of the Tzotzil study was the use of a test based on isolated sentences along with a test based on an elicited text. The sentence test consisted of seven isolated statements recorded from a text. Each of these was prefaced by an introductory phrase in the local dialect and was followed by a question in the same dialect. Crawford and Stoltzfus wanted to measure intelligibility with reference to connected discourse as well as that registered for short stretches of speech. They wanted to know how scores obtained with sentence tests compared with those obtained through text tests. They expected a difference between the scores to reflect the effect of a wider linguistic context on intelligibility. (If context were the only variable involved, the sentence test should tend to score lower than the text test since the primary effect of context is to resolve ambiguities, thus heightening understanding.) Since the comparison was made only for a small sample, the results were highly tentative. They showed no essential difference between the two methods.

The Tzotzil study also provided the suggestion for a modification that was incorporated into later test formats. This involved dubbing from each text the statements that carried the answers to each question. Each of these statements was then paired with the corresponding question on a segment of tape which was spliced with the portion of text used for the test tape. The pattern of statements and questions on the tape, then, can be diagrammed as follows: S1-Q1; S2-Q2; etc. This modification was designed to reduce the effects of differential I.Q., memory ability, and attention span on subjects' scores by reducing the time between a subject's hearing a particular statement in the text and his being asked a question about it. In addition, juxtaposing statements and their corresponding questions made it possible for an investigator to locate a statement on tape rapidly if he felt it was necessary to repeat it to a subject.

Finally, Crawford and Stoltzfus estimated the threshold of minimum effective communication as 70%. The study covered six dialect areas.

3.4.3. The Chol study. In the Chol study the innovations incorporated in, or suggested by, the Tzotzil study were first applied consistently. It was made by Stoltzfus and Wilbur Aulie, who obtained intelligibility data for seven dialect areas.

In the test format, the sample story was incorporated into the introduction tape, the statements dubbed from the texts were paired with the corresponding questions for the test tapes, and the order in which the test tapes were presented was varied between the subjects. (The method used to rotate the order of the tapes is discussed in Section 2.2.2.)

The effect of these modifications on the hometown raw test scores was considerable in relation to the scores recorded for earlier studies. The range of raw mean scores for the seven hometown test tapes was from
84% to 99%, with a mean of 93%. For the Mixtec study, for example, the mean unadjusted hometown score was 84%, the mean adjusted score was 89% (see p. 62). In only one instance was a mean score for a reference point test tape higher than the mean score on the hometown test tape. The difference was less than 3%.

The number of questions on each test tape was increased to fifteen. The purpose was to compare mean scores of a ten-question test with one of fifteen questions. The mean variation between the two for twenty-five cases was 2.1%. The amount by which mean scores of tests based on the fifteen questions varied above or below the mean for tests with ten questions was less than 10%, with one extreme case in which the difference was 12.1%.

The introduction of dubbed statements preceding each question of the test tape had an unexpected effect. Instead of answering the question on the basis of the statement that preceded it, many subjects tended to repeat the question. This meant that the subject still did not understand what was required of him and entailed further instruction on the part of the tester. Thus, while the modification significantly altered the scores recorded, it also tended to increase the difficulty in preparing a subject adequately for the test. The gain in reliability, however, justified the complication.

Finally, Stoltzfus estimated that the level of intelligibility at which materials produced at one point would serve at another point in the Chol area would be 90% or above. Below this level, literacy materials would need adapting for the sub-dialect. This estimate was based on adjusted test scores.

3.4.4. The Mazatec study. The Mazatec study was carried out by Paul Kirk in twenty-three Mazatec towns (Kirk 1970). The only change from the format employed in the Chol study consisted of a standardized sample story for the introduction test tape. Kirk elicited this story phrase by phrase at each test point, rather than dubbing it from an additional elicited narrative. The story was recorded and then divided into three phrases which were either dubbed from the original copy or recorded again by the informant. These phrases were paired with their respective questions in the same way as the test tapes in order to form the instruction tape.

The system of rotating test tapes for successive subjects was varied somewhat depending upon the estimated difficulty that a subject would experience in understanding the test tapes. If all the test tapes seemed fairly easy, the hometown test tape was rotated in with the rest of the tapes. If none of the test tapes seemed fairly easy, the hometown test tape was administered first to the entire sample, and the rest of the tapes were rotated in order of presentation to each subject. In a few cases, difficult tapes were paired with easy ones and the tapes were rotated by pairs. Regardless of the system of rotation used, the attempt was made to ensure that no subject was presented a difficult test tape as his first one. In the same spirit, an attempt was made to bracket all difficult test tapes by easier ones so that a subject would not be presented in sequence with two test tapes he could not understand.

The tendency for subjects to repeat questions rather than answer them, first noticed during the Chol study, was again observed. Kirk agreed with Stoltzfus in: this indicated that some subjects had not learned adequately
how to take the test. He felt that for others, however, this was an attempt to avoid admitting that they did not understand the text and thus could not answer the question. These observations emphasized the necessity for a tester to note how subjects tended to behave with respect to the various test tapes (see p. 25).

The minimum threshold of effective communication was estimated by Kirk to be about 80%. One contribution to the method was the attempt by Kirk to validate this estimate by correlating it with subjects' evaluations as to whether other dialects were easy or difficult to understand. His tentative conclusion was that dialects rated as easy registered above 80% for the sample mean and dialects rated as hard registered below 80% (see Section 4.3.3).

Another pilot study was designed to measure the extent to which sample mean scores differ when several tests from the same reference point were constructed by using texts that differed in content and narrator. There was no significant difference recorded for sample mean scores on four texts from the Huautla dialect (see Section 6.3.3 and Appendix H). The final study compared mean scores for samples of three and six subjects with the mean scores of samples of ten subjects for the entire Mazatec study (see Section 6.3.3).

The format of the Chol and Mazatec studies was also used in the Chinantec, Mesilla Totonac, and much of the Zapotec studies. A feature peculiar to the Chinantec study was that the statements dubbed from the text were recorded on the test tape following the corresponding questions, rather than preceding them. This modification was introduced because of pronoun shifts to second person that were encountered consistently during elicitation. It had the effect of turning the test into a dialogue between the subject and the tape recorder, however. Putting the cues after the question increased the possibility of mimicry by the subject. In addition, some subjects who reacted as if in a dialogue with the tape recorder sometimes thought they were being asked about their own personal activities. Some of these may have forgotten the text entirely. Unfortunately we do not have any way to estimate how these consequences of the modification affected the test scores. Because it was felt that this modification could affect test results significantly, it was not adopted for any subsequent study.

3.4.5. The Zoque study. The latest modification in the test format was initiated by Stoltzfus for Zoque. It consisted of leaving stretches of blank tape between segments of the text as it was dubbed onto the test tape. The questions for that test tape were then recorded on these stretches of blank tape. Thus the questions on the test tape were embedded within the body of the text itself. The test was interrupted as soon as possible after the end of the statement in the text that answered a particular question. This format can be diagrammed as follows, with \( T_i \) representing a segment of the text and \( Q_i \) the corresponding question: \( T_1 - Q_1 \); \( T_2 - Q_2 \); \( T_3 - Q_3 \); etc.

There were two purposes for the modification. The first was to make the test procedure easier for the subject to learn, and thus simplify the administering of the test. During early studies, subjects who readily answered the questions to the sample story did not necessarily know how to respond to the initial questions for the hometown test tape. One way in which this was manifested was by an insistence on repeating the question rather than answering it. Many subjects seemed to forget that the questions related to the content of the story they had just heard. All this seemed to indicate that the subjects were still not trained to respond to the questions.
Because the sample test seemed to be inadequate, the investigator had to use the hometown test tape to finish training the subject. This situation lengthened the time for test administration and sometimes generated tension between the tester and the subject. Occasionally, potential subjects who were observing the testing decided on the basis of another subject's difficulties that they themselves could not answer the questions either. Rather than risk public embarrassment, they either left the testing area or refused to cooperate. It was hoped that by keeping the subject within the context of the story the testing procedure would be easier to learn.

The second purpose was to simplify the technician's work in constructing the test tapes. The steps of dubbing each statement from the text, pairing these statements with the corresponding questions, and splicing the sets of statements and questions with each of the test tapes, are eliminated. All that is needed is to leave spaces of blank tape at appropriate places in the copy of the text which is used for the test. The questions are then recorded onto these stretches of blank tape. To modify the test tape at a following test point it is necessary only to record the newly translated questions over the ones used at the previous test point.

This format has subsequently been used in the Huave, Choapan Zapotec, Pame, Otomi, and Aztec studies. It gives results equivalent to those obtained in the Chol and Mazatec studies. Hometown scores approximate 100%, and the threshold range for this design is estimated at between 75% and 85%. Test administration time may be slightly reduced for each subject from that of the Mazatec study. It generally takes forty to forty-five minutes to administer a set of six test tapes to a particular subject.
4 VALIDITY

4.0. Questions of validity arise because the method consists of estimating a particular criterion property through the assessment of a particular trait; the maximum geographical area within which a mode of speech can effectively serve for communication is estimated from intelligibility test scores. Any such measurement process incorporates a certain amount of random and systematic error. This error must be taken into account since its magnitude directly influences one's ability to make inferences from the results. The concepts of reliability and validity are used to estimate the kind and degree of this measurement error (Blalock 1968a:14).

4.1. "Reliability" and "Validity". It is easy to confuse reliability with one of the senses strictly applied to validity. Reliability is the degree of consistency between identical, repeated measurements of the same thing (Siegel and Hodge 1968:56). It is estimated by coefficients of correlation between tests, obtained by one of three basic approaches (Upshaw 1968:65). In the first the same test is administered twice to the same set of subjects with a time interval between testings. The second requires the administration of two forms of the same test to one set of subjects. Both are administered close together in time; often the second form immediately follows the first. The third method separates the individual test items into two groups, such as odd-even, and then scores the subjects on each group of items. A correlation coefficient is then computed between the two sets of scores to yield a measure of internal consistency. Since correlation coefficients are affected by the variability of the sample to which a test is administered, and since this range of variability differs from sample to sample, the reliability coefficient of a test (which is a correlation coefficient) will be different for successive testings that involve different samples (Downie and Heath 1965:221). Thus, the reliability coefficient will not be a single, invariant value. Finally, the more items that a test includes, the higher its reliability coefficient will be (Downie and Heath 1965:218).

Reliability, then, is concerned with repeated measurements of an instrument; i.e., would the same scores be obtained if the test were given
again (Siegel and Hodge 1968.56). For this study, therefore, the following questions are central in establishing reliability: (1) If the same study were carried out by different investigators using the same tests and samples of subjects, how well would the results of the two series of tests correlate? (2) How closely would sample mean scores correlate if different test tapes from the same reference point were administered to the same sample of subjects? (3) How closely would sample mean scores correlate if the same set of test tapes were administered to separate samples of subjects at a given test point? (4) How would the intelligibility scores correlate if the length of the test were expanded? (5) How well would intelligibility scores based on samples of different sizes correlate?

Validity, on the other hand, is used in at least two senses. The first, often called logical validity, or “face validity,” refers to the degree to which a test actually measures what it claims to measure. Logical validity cannot be determined operationally; it rests on the assumption that the measuring instrument defines something that corresponds to the right theoretical construct. This untestable assumption is used to link theoretical concepts and operationally defined concepts. The degree of logical validity which is then attributed to an index depends solely upon a consensus by experts as to the plausibility of that index or of the individual items which make up the index (Blalock 1968a.13). For the present method, for example, it is necessary to consider the plausibility of the assumption that intelligibility is the right indicator of how acceptable it would be to use a primary dialect to communicate with speakers of a secondary one. In addition, do the tests we use really measure intelligibility?

In its second sense, called empirical validity, the concept refers to the degree of consistency among alternative measures of the same theoretical construct. These alternative measures must be equally plausible, but they must involve disjoint operations (Siegel and Hodge 1968.56). For example, a questionnaire and an intelligibility test might be proper alternative measures for establishing validity while two forms of an intelligibility test would not.

Measuring instruments fall into two classes: assessment and prediction (Upshaw 1968.67). Assessment instruments assign numbers to observations of a property to indicate how much of the property exists for each observation. For example, students’ knowledge of chemistry might be measured by a test consisting of thirty-five items. The set of observations would consist of the set of scores that the students obtained. The property “knowledge of chemistry” would be measured directly for each student by the number of correct responses he made. Prediction instruments, on the other hand, are used to predict to some outside criterion that is not directly observed. Such an instrument may involve the assessment of another property as an intermediate step (Upshaw 1968.66). For example, intelligibility tests assess the degree of intelligibility among dialects. The measure of intelligibility, in turn, predicts values for an outside criterion, the maximum extendability of dialects, under the condition of a threshold range of adequate communication. In both cases, however, the empirical validity of an instrument is demonstrated by the degree to which it accurately predicts to the criterion measure. This is generally expressed by a correlation coefficient which relates the original instrument to some other measuring device.

In order to determine the empirical validity of any one of a group of instruments, it is necessary to use multiple instruments to measure the
same variable. An additional advantage of using several instruments, at least in the initial stages of research, is that if two or more of them give equivalent results, then the investigator can choose the one that is more economical to apply and discard the other (Blalock 1968a.13). Different instruments, however, can vary in intent so that either of two instruments might give results equivalent to the original instrument only with regard to the criterion measure while differing greatly in terms of their correlation with the assessment of the property that the original instrument directly measures (Upshaw 1968.67). For example, a test based on subjects' opinions about the difficulty they have in understanding other dialects might assess the same trait as an intelligibility test and thus highly correlate with it, while one that is based on a subject's attitudes toward people of nearby towns would assess a different trait and not correlate with intelligibility at all.

In summary, then, reliability refers to the consistency between values obtained by repeating measurements with the same instrument. Validity, in its logical sense, refers to the plausibility that an instrument measures what it is supposed to measure. In its empirical sense, it seeks to determine whether there is a correlation between measurements obtained by using independent instruments, and if there is, it asks how strong this correlation is. In practice, both senses of validity are implied in one's attempt to assess the validity of an instrument (Siegel and Hodge 1968.56). Finally, the relationship between reliability and validity is expressed by the oft-quoted dictum that an instrument can be reliable without being valid, but it cannot be valid without having some degree of reliability.

4.2. Validation Must be Indirect. Assessing the validity of an index is a problem in fitting theoretical (postulated) concepts to intuitional (measured) ones (Jackson and Curtis 1968.112). The method used must take into account the following aspects of the problem: (1) the nature of the relationship between theory and research, (2) the kind of variable(s) being measured, (3) the components and meaning of an index, and (4) the possibilities for replicating an experiment (Blalock 1968a.14, 17). In addition, conceptual schemes may have to be designed to relate already collected empirical data, or rules may have to be specified for collecting the data which are to be explained in terms of an already formulated theory (Jackson and Curtis 1968.112). Because the problem is frequently so complex, the method used to assess validity must often be an indirect one.

4.2.1. Theory and research. In the first place, theory and research are related only indirectly; the kinds of concepts associated with each are distinct. Concepts by postulation are part of a deductive theory, and are wholly or partly defined by the postulates of the theory. Concepts by intuition, on the other hand, are obtained by observing data and classifying and analyzing them (Blalock 1968a.10). Blalock points out that all of the concepts in a deductive theory are concepts by postulation. Since there are no concepts by intuition in a deductive theory, neither the theory nor any of the propositions implied by the theory are directly testable (1968a.11). To link the two kinds of

1. Furthermore, it is not always possible to distinguish between validity and reliability.
2. In another sense, however, theory and research are intimately related; it is frequently difficult, if not impossible, to label one aspect of a problem "practical" and another "theoretical."
concepts, then, one assumes that each concept by intuition corresponds to
some postulated concept (1968a.7). A further complication is that while
some of the postulated concepts embodied in a theory may be associated
with concepts by intuition, others may not be; those not associated with
intuitional (operational) concepts must be left out of propositions that are
intended to be testable hypotheses, since tests of hypotheses are carried
out in terms of operations (1968a.11).

4.2.2. The variables being measured. The nature of the variables being
measured will determine the complexity of the linkage between theoretical
and operational concepts. For some variables such as age, sex, and place
of birth, the linkage is likely to be simple enough that they can be
measured directly. For such cases, then, the concept of validity may be
useful; the validity of the instrument that measures them may be easily
assessed (Blalock 1968a.14). For other variables, however, the linkage is not
direct; they are abstractions whose causal relations with other variables are
assumed. These variables must be measured in terms of their hypothetical
effects. Blalock discusses three such variables, i.e., the physicist's concept,
mass; power as used by both physicists and social scientists; and
discrimination as defined by the sociologist. Intelligibility is another variable
that is measureable only indirectly. (In these more complex situations,
Blalock suggests that the notion of validity may be so simplistic as to be
misleading; there may be no single indicator of validity [1968a.19].)

4.2.3. The components and meaning of an index. A single indicator will
probably be inadequate to validate an index that is composed of several
items. It is necessary to establish the plausibility of each of these
components and to correlate them with external criteria. To validate the
index of dialect extendability it is necessary to determine whether
intelligibility is a plausible indicator. It is also necessary to determine what
causes variation in the dependent variable, intelligibility. By specifying these
causes it will then be possible to demonstrate the validity of particular test
scores, some of which may have seemed intuitively to be too high or too
low. Furthermore, the limit on the index of extendability is determined by a
threshold range of intelligibility. Since this range will vary depending upon
the type of instrument used, characteristics peculiar to each language
studied, and sociological features of different areas, it will have to be
validated for each study.

Finally, the investigator's own understanding of the concept "maximum
(or optimal) extendability of dialects" will affect the degree of plausibility in
determining the number of potential centers of communication on the basis
of intelligibility data. If, for example, one assumes that all that is of interest
is the geographical area within which a native speaker of a given dialect
can make himself understood, then the use of data on oral intelligibility
seems quite plausible. On the other hand, if one's interest is to determine
the maximum area within which a set of printed materials will serve
potential readers, then the plausibility (or external validity) of the oral
intelligibility index is decreased because a subject's reaction to oral
materials is used to predict his reaction to written materials (Siegel and
Hodge 1968.55). To increase the plausibility of the index in this case, then,
one has to include additional items, and the validity of each of these must
also be assessed.
4.2.4. Replication. Another aspect of the problem of testing validity relates to whether or not an investigator can replicate an experiment. Replication consists of repeating an experiment, in many cases making systematic changes in outside variables for successive repetitions of the experiment. Regarding the method of intelligibility testing, some outside variables are the following: (1) the investigator evaluating subject responses; (2) the category and content of the text; (3) the speech quality and style of the informant; (4) the sample of subjects selected at a given test point; and (5) characteristics of the individual subjects, e.g., age, sex, education, language ability, and attitude toward the investigator. Replication, then, has at least two related purposes: to determine the effect of such outside variables on the measure of the dependent variable, and to allow a more accurate measurement of it (Blalock 1968a.15).

If the investigator is able to replicate his experiment, when he formulates a theory he can indicate the effect of these outside variables and specify their relationship to both theoretical and intuitive concepts. If he cannot replicate, then he must make untestable assumptions regarding the behavior of these outside variables. He must include these assumptions along with others that he has already made regarding the linkage between theoretically and operationally defined variables in a theory which he then submits either in part or as a whole to empirical evaluation. Replication, however, permits one to formulate these theories with a smaller set of untestable assumptions (Blalock 1968a.16). The point is that, for the most part, replication of the experiments carried out in Mexico has consisted of using the same instrument for successive studies while modifying it to control certain problems observed during the testing. In this way, the original instrument has been developed to yield a more reliable measure of intelligibility. Because of the ways in which the original instrument has been modified, we feel justified in assuming that variables such as lapse of memory, external distractions, differential I.Q., and differences in the educational background of the subjects do not materially affect the test scores obtained by using the latest versions of the instrument. On the other hand, replication in a broader sense has been restricted because of limitations on time and available personnel. Therefore, the effect on the test results of other kinds of disturbing influences such as those mentioned in the previous paragraph is still not well known. (Some data are available, however; they are discussed in Section 6.3.3 and in Appendix 1.) For these variables, then, we may be forced to assume that their quantitative effects are negligible (which is not the case), if we want to formulate a testable theory.

4.3. Toward a Validation Procedure. An approach to fitting a set of postulated concepts to a set of measured variables has been suggested by Blalock. It is specifically designed for situations like the survey problem, in which a large number of variables are involved, and controls for measurement error are difficult or impossible to use (1968b.158). Briefly, he requires that a main theory be linked to an auxiliary theory in a mathematical model which corresponds to an oversimplified picture of reality (1968b.159). The main theory consists of an inventory of postulated concepts that have specific causal relationships to one another (1969.29). The auxiliary theory is formulated to adapt the general theory to a particular population, a given set of measuring instruments, and a specific research design. It specifies an inventory of measured variables that are linked in an explicit way to some of the variables in the main theory. In addition, it adds
to the model a set of variables that cause measurement errors. It relates them to certain measured variables and makes explicit the extent of random error and specific kinds of nonrandom errors that they cause (1968a:25). Both the main and auxiliary theories must be based on empirical data. They must also be explicitly stated in such a way that the propositions that are derivable from them can be tested with reference to further empirical data (1969:8). The theory will be confirmed, modified, or rejected, depending on the results of these empirical tests. Validation will thus proceed by eliminating unsatisfactory theories rather than by definitely proving any one theory (1969:152).

Bialock's approach is sophisticated. It presupposes that we know what the relevant variables are and that we have measured them. In addition, it assumes that we have measured the correlations between these variables. Finally, it assumes that we have reasonable estimates about the size of error terms that must be included in a set of simultaneous equations which form the framework for stating testable propositions. Unfortunately, we cannot meet these presuppositions at the present time. However, we do have enough data from various studies that we can begin to state a verbal theory which can be formalized in a rough way. This semiformal account can tell us what kind of additional data we need to collect in order to more fully take advantage of Bialock's approach.

In the following section, then, we present these data to provide an empirical base for applying this indirect procedure. (A fuller and somewhat more technical discussion of Bialock's method is given in Appendix J.) We also present these data to show why we feel that the method of intelligibility testing is valid. The data relate to the following questions: (1) How plausible is the measure of intelligibility as an indicator of dialect extendability? (2) What variables cause intelligibility to vary? (3) What is the threshold range of intelligibility that corresponds to the minimum level of effective communication between a pair of dialects? (4) What variables determine the acceptability of written materials?

4.3.1. The plausibility of the index. To determine the plausibility of using intelligibility results as an index of dialect extendability we must take into account the intended use of the index. The practical problem is that it is impossible to carry out programs based on oral communication for every dialect that can be identified, much less prepare written materials for each of them. To meet the communication need in a feasible way, it is necessary to stimulate speakers of some dialects to accept a slightly different dialect for a particular function and thus modify their own habits of language use. Therefore the index will be used to help answer the following question: Which dialects (or speech forms) are likely to be acceptable to speakers of other dialects?

This can be answered only in terms of extralinguistic variables such as group attitudes, perceptions, values, and behavior. An accurate evaluation of these properties allows one to predict whether or not one group will accept the form of speech another group uses. Since intelligibility is dependent upon at least some of these extralinguistic variables (Wolff 1964), we have used the measure of it as the central element in an index of dialect extendability.

The conclusion that intelligibility is determined by extralinguistic variables as well as by linguistic ones is based on two observations. In the first place, it is characteristically nonreciprocal, or asymmetric. For most cases,
speakers from dialect A can understand a test tape from dialect B better than speakers from dialect B understand the test tape from A. For example, for half of the pairs of scores that Biggs obtained during his study of Yuman languages, the degree of nonreciprocity is between 5% and 15%. For the Mazatec area, San Jerónimo subjects scored 76% on the Huautla test tape while Huautla subjects scored 89% on the San Jerónimo test tape; Huautla scored 60% on Mazatlán, and Mazatlán scored 89% on Huautla; Huautla scored 35% on Jalapa, and Jalapa scored 73% on Huautla. These are just a few of the examples that can be cited; they show both that intelligibility is nonreciprocal and that the degree of nonreciprocity may be considerable.

A related observation is that when there are several dialects A that understand a dialect B better than B understands any of them, it usually turns out that B is an economic and political center on which the A dialect areas are dependent. As a result, speakers of the A dialects learn to speak the B dialect, but speakers of B do not learn to speak any of the A dialects. Olmsted, for example, concluded from his study of Achumawi and Atsugewi that intelligibility was nonreciprocal because Atsugewi speakers learned Achumawi but Achumawi speakers did not learn Atsugewi. The explanation was sociological, not linguistic (see Section 3.1.1). Olmsted’s conclusion is reinforced by data from the Tarija study carried out by Casad and Hollenbach.Briefly, speakers of the San Martín dialect understood both of the other two Tarija dialects, Chichuaxtla and Copala, much better than speakers of the other dialects understood San Martín. This can be accounted for in terms of economic and political dependencies. (For a more complete treatment of these data, see Sections 4.3.2, 4.3.4, and Appendix J.)

While nonreciprocity usually results from interdialectal learning, there may also be cases in which it can be adequately explained by linguistic phenomena. For example, since Portuguese has undergone a consonant deletion rule that Spanish has not, the surface phonological forms of Spanish correspond more closely to underlying proto-forms than the surface forms of Portuguese do. One might therefore predict that Portuguese speakers can understand Spanish better than Spanish speakers can understand Portuguese.3

Most previous researchers have treated the intelligibility relation between a pair of dialects as though it were symmetric, or at least have overlooked the significance of the nonreciprocity indicated by their data.4 As a result they have talked about intelligibility, measured it, and interpreted it in a misleading way. For example, the term “mutual intelligibility” implies a reciprocal, or symmetric, relationship. It may be useful for cases in which intelligibility is very high or very low in both directions, but in most cases it would be more accurate to simply say that intelligibility exists among a set of dialects or languages in varying degrees. We might do better to use the term “cross-language communication” (Weinreich 1957; Ladefoged 1968).

A measure of “mutual intelligibility,” such as the one Pierce derived (1952), will often be misleading and theoretically uninteresting for the sociolinguist. For example, the pair of scores for Huautla and Mazatlán,

3. This was suggested to me by Peter Landerman (personal communication). Also, note Voegelin and Harris (1951:326), who mention symmetrical and asymmetrical contrasts as kinds of linguistic differences.

4. For example, see the studies discussed in Section 3.1, and Yamagiwa’s study treated in Section 4.3.2.
89% and 60%, would yield an index of mutual intelligibility of 74.5%. But the figure 74.5 tells us very little. On the other hand, from the pair of scores we can predict that when a Huautla speaker and a Mazatlán speaker talk together in Mazatec, the Mazatlán speaker must adapt to the Huautla dialect if communication is to be successful, and, in fact, he will adapt to it. We make this prediction because the 89% mean score that Mazatlán subjects made on the Huautla tape is well within the range of adequate communication, while the 60% mean score that Huautla subjects scored on the Mazatlán tape is demonstrably below the threshold of adequate communication (see Section 4.3.3).

In earlier studies, the measure of intelligibility was interpreted as an index of genetic relationships among dialects. (Genetic relationships, unlike intelligibility, are reciprocal.) Hans Wolff pointed out that since intelligibility is determined by extralinguistic factors it is not a valid indicator of such relationships (1964). He measured dialect relationships by means of both a standard lexicostatistic method and intelligibility testing; the results of the two methods correlated only slightly (1964,441). Wolff concluded, then, that the two methods were not comparable; even though sometimes the two might give equivalent results, this did not happen often enough that one could adequately predict intelligibility solely from a measure of linguistic similarity.

Wolff’s conclusions are exactly what one would expect since the sets of variables that underlie the two kinds of measures are partly disjoint. Linguistic similarity is based on phonological, morphological, syntactic, and lexical variables; intelligibility is based on all these plus interdialectal learning and convergence due to borrowings (Appendix J). These facts are in themselves sufficient to account for the differences between the grouping of Mazatec dialects as given in Figure 4 and the grouping given by Gudschinsky (1958). Two sets of measuring instruments based on different assumptions, designed for different purposes, and performing operations on partly disjoint sets of variables should yield different results.

Any index of dialect extendability that is based solely on the measure of linguistic similarity will probably be inadequate, even though the procedure allows loan words to figure in the calculations. The presence of loan words may reflect former communication patterns that may differ greatly from current ones (cf. Wolff 1967). Thus a high incidence of loan words in a pair of word lists does not necessarily mean that high intelligibility currently exists between a pair of dialects.

Recent studies by Ladefoged and others (1972), and by Bender and Cooper (1971), have substantiated the linear relationship for closely related languages that data from Biggs’s study suggested. In brief, Ladefoged used a revised lexicostatistic word list which allowed him to determine the proportion of words common to the various Ugandan languages being

5. Wolff shows that in one case both communication patterns and language use patterns changed, but that change in use of loan words lagged behind the change in communication patterns for a considerable time (see Section 8.3.3).

A measure of similarity, to be maximally useful, will probably have to evaluate both the percentage of likeness as well as the various kinds of differences among dialects (cf. Bender and Cooper 1971,50). It is also likely that at least some kinds of linguistic differences will have to be examined for their effects on nonreciprocal scores (cf. p. 73).

Finally, if the investigator does not want to set up intelligibility tests, he will at least have to supplement his questionnaire data about informant attitudes with observations about societal relationships (Section 7.1).
studied. From these lists he also derived a measure of phonetic similarity. He then showed that both of these measures correlated very significantly with the degree of comprehension between those pairs of the languages he tested (1972.77).

Ladefoged points out that linguistic similarity may not always be an adequate predictor of intelligibility. In particular he cites one case in which speakers of one language understood a second one much better than the investigators would have predicted from the proportion of shared vocabulary. Although these subjects denied any knowledge of the other language, they really did understand it well, partly because of its use in radio programs (1972.76-77). The investigators therefore concluded that "within this group of languages, knowledge of the percentage of words in common allows us to predict the degree of comprehension, except when questions of prestige are involved" (1972.77: italics mine). It is significant that this conclusion was reached even though the investigators discarded test scores of anyone whose family spoke a second Ugandan language or who had lived for a significantly long time in an area where a second language was predominant (1972.65).

Bender and Cooper followed Ladefoged's study with one they did on the Sidamo subgroup of East Cushitic. (Their study was part of the Uganda Language Survey and included the Alaba, Burji, Derasa, Hadiyya, Kembatta, and Sidamo languages.) They reported results roughly equivalent to Ladefoged's, citing highly significant correlations between intelligibility and each of the four measures of linguistic similarity: proportion of shared vocabulary, of shared root morphemes, of shared grammatical morphemes, and percentage of shared total combination of root and grammatical morphemes (1971.42). In addition they found that the measure of geographical proximity to major population centers was significantly related to intelligibility (1971.48).

Nevertheless, they were well aware of the substantial degree to which intelligibility did not correlate with the measures of linguistic similarity (1971.50). This was in spite of the fact that they had restricted their sample as much as possible to subjects who had had no previous exposure to the other languages. In order to estimate how much the subjects' scores might possibly be influenced by prior knowledge of the other languages, they had the respondents rate their own ability on each of the test languages. These self-ratings were then averaged to yield a group self-rating. In almost half of the cases the group self-rating was zero. The average value of eleven other cases was very near zero. In only two cases was the group average rating significantly high: both Hadiyya speakers and Alaba speakers gave themselves high ratings on their ability to understand Kembatta (1971.40).

It is interesting that the difference between the classification of the Sidamo languages by intelligibility and the usual genetic classification involves just those two language groups for which Bender and Cooper report that the self-ratings indicated a possibly significant amount of prior

---

6. This procedure was followed because they were interested in describing just the linguistic features that "enable speakers of one language to understand a second, related language upon hearing it for the first time" (1971.50-51).

There are other aspects of their study that deserve comment, including their finding of no statistical significance to the observed nonreciprocity in the test scores they obtained. It is sufficient to point out that differences between at least some of their conclusions and ours is the result of the different methods used in setting up the tests and selecting subjects as well as in collecting and analyzing the data.
knowledge of another language. In a scale where zero meant "no ability at all," and 1 meant "able to understand a little of an ordinary conversation," Hadiyya speakers had a mean score of 1.5 on Kembatta, while the mean for Alaba on Kembatta was 1.8. Assuming that the category labels of their scale correspond to those of the one proposed by Ladefoged, 2 on the scale would have been "the main points of a conversation" (see Section 7.3). In short, intelligibility lumps Kembatta, Alaba, and Hadiyya together, while the genetic classification pairs Kembatta-Alaba with Sidamo-Deresa, all four of which are then paired with Hidayya. Intelligibility generally relates Hadiyya more closely to the rest of the languages than the strictly linguistic classification does (1971.45).

Bender and Cooper suggest causes such as heavy borrowings from a closely related language or a group effort to preserve archaic forms as possible explanations for the differences between the two classifications. On the other hand they also suggest that the anomaly may be due to a lack of sufficiently detailed data on the genetic relationships in the Sidamo subgroup (1971.50). From our viewpoint it would be worth while to investigate the sociocultural relationships among the groups to see whether or not the Hadiyya and Alaba groups are socially and economically dependent on Kembatta. It would not be at all surprising to find such an explanation to be the most satisfactory one.

In conclusion, intelligibility seems to be a more plausible indicator of dialect extendability than linguistic similarity simply because intelligibility is partly based on interdialectal learning; if people adapt to a different dialect for one purpose it is likely that they can be encouraged to adapt to it for an additional one. It is unlikely that this kind of behavior can be estimated entirely from a measure of linguistic similarity. It certainly cannot be estimated from a measure of genetic relationships.

4.3.2. The variables that determine intelligibility. In an attempt to explain which variables determine the degree of intelligibility among dialects, data are cited from Mazatec, Trique, and the study of Japanese dialects carried out by Yamagiwa. Some of the variables that contributed to measurement error are also discussed.

7. This does not mean that a measure of linguistic similarity is irrelevant. Various indicators are necessary to establish validity, and linguistic similarity is one of them. This is the import of the studies by Ladefoged and by Bendor and Cooper. Linguistic similarity is not the sole predictor of intelligibility, however, even when only highly related languages are involved. Bendor and Cooper themselves were aware of this (1971.50), and that is precisely what Wolff's conclusions and the data we present here show.

At this point, the notion of "mutual" intelligibility is useful in order to begin sorting out all the kinds of things that contribute to intelligibility and to evaluate the relative importance each one enjoys. At least it is simpler in the long run to use a single value of intelligibility for each pair of dialects in order to arrive at the correlation between intelligibility and various measures of linguistic similarity.

Once the correlations between these two variables are established (and possibly, standardized), it will be necessary to identify and determine the correlations between various nonlinguistic variables and intelligibility. The next step would be to determine and assign relative weights to both linguistic and nonlinguistic predictors. The weighted coefficients of correlation could then be included in a general prediction equation (or a series of prediction equations) which would allow us to reasonably predict to the kind of nonreciprocal relationships we consider in this chapter. In Appendix J I suggest what the result of carrying out these procedures might look like in relation to a particular field problem.
(1) Mazatec. Several characteristics of the Mazatec area probably determined the intelligibility scores that were obtained. Two of these are the overwhelming dominance of Huautla and the obvious linguistic differences between the Huautla dialect and the dialects spoken in a set of towns that are geographically isolated from Huautla. Huautla is the largest town in the Mazatec zone. It is also the economic, political, and religious center for much of the highland area. There is a great deal of contact with the Huautla dialect by the people of surrounding towns, and these people have learned to understand the Huautla dialect. On the other hand, Huautla speakers have never accommodated themselves to any other Mazatec dialect. This unequal state is reflected by two sets of scores that tend to be non-reciprocal in favor of Huautla, i.e., all the A dialects understand the Huautla dialect better than Huautla understands any of them. In the first place, for two of the three pairs of scores between Huautla and another dialect, the other dialect scored significantly higher on Huautla than Huautla scored on it (see pp. 72-73). In addition, relatively high scores were obtained on the Huautla test tape at all of the test points where it was used (the only exception was Chiquihuitlán), while very low test scores were obtained whenever test tapes from dialect areas outside of the Huautla area were used at outliers to Huautla. For example, San Lorenzo, Mazatzongo, and Huehuetlán (outliers of San Jerónimo) scored 82%, 76%, and 69%, respectively, on the Huautla tape. On the other hand, San Mateo, Santa María, and Tenango (all of which had registered nearly 100% intelligibility with the test tape from Huautla) scored only 41%, 25%, and 24%, respectively, on the San Lorenzo test tape. Subjects from San Mateo and Santa María scored only 28% and 14% on the tape from Huehuetlán.

In considering the above scores, we can say that the interdialectal learning through prolonged contact results in high intelligibility. Furthermore, interdialectal learning is unidirectional, favoring the dialect which is the economic, political, and religious center. Finally, since there is practically no contact between speakers of the San Lorenzo-Mazatzongo dialect and those from outliers to Huautla, and since there are numerous linguistic differences between these two dialect areas, we would expect that Huautla speakers could barely understand San Lorenzo-Mazatzongo speakers, if at all. (Some of the phonological differences between these dialects are discussed in Appendix G.)

That San Lorenzo-Mazatzongo speakers do understand the Huautla dialect somewhat, correlates with the fact that many of them travel to the Huautla dialect area for commercial reasons. However, the San Lorenzo-Mazatzongo area is primarily associated with San Jerónimo rather than with Huautla. San Jerónimo is the administrative and political head for most of the area. Most people of these towns who enter or leave the Mazatec area pass through San Jerónimo but not through Huautla. Subjects in these towns scored better on San Jerónimo than they did on Huautla.

8. The exception was San Jerónimo; Huautla subjects scored an average of 13% higher on the San Jerónimo tape than those from San Jerónimo scored on the Huautla tape. Several possibilities suggest themselves: experimenter maturation—the testing began in San Jerónimo; subject variation—the Huautla subjects seemed generally better educated and more cooperative than those in San Jerónimo (SIL already had established friendships in Huautla but not in San Jerónimo); linguistic differences—the linguistic differences between the two dialects may tend to cause intelligibility to be higher on San Jerónimo. Note that the 13% difference is much less than the differences for the other pairs of scores (26% and 39%).
An auxiliary theory that tests the validity of the Mazatec intelligibility results must also pinpoint major sources of error. First, there was an obvious change in accuracy of scoring responses due to experimenter maturation. That is, the longer Kirk used the method, the more efficient he became in teaching subjects to take the test and in scoring their responses. This is especially true since the testing began in the section of the Mazatec area least familiar to Kirk. Second, negative attitudes on the part of the subjects in some cases also adversely affected the test scores. The only estimate available of the effects of these two factors must be derived from a comparison with the hometown test scores. The twenty-two hometown scores are bimodally distributed with modal scores at 91% and 94%. The median score is 94%, and the mean 93.4%. Of the hometown scores less than 93%, all were from towns in the Huautla-San Jerónimo dialect area. (In five cases it is fair to say that difficulties were encountered in securing adequate informants or subjects.) The mean hometown test score for the first five samples of subjects tested during the study was 92.4%; for the last five samples it was 94.6%. Thus, experimenter maturation resulted in a change of less than 5% in the hometown test scores, and it seems reasonable to assume that the effect of the same factor on non-hometown test scores was comparable. In the auxiliary theory, then, errors due to experimenter maturation will be considered systematic but negligible, whereas those due to negative attitudes on the part of the subject will be considered nonsystematic or random.

An unknown amount of systematic error was introduced into the intelligibility scores because of the sampling method. With only one or two exceptions, the samples consisted exclusively of male subjects. Since Mazatec men travel to other dialect areas much more than Mazatec women, it is reasonable to suspect that men, on the average, understand the other dialects better than women, and that the true population value for intelligibility is lower than the sample mean scores obtained during the study. Since we are not justified in generalizing to population values on the basis of these mean scores, this systematic sampling error is brought into the auxiliary theory as an unmeasured variable (Blalock 1968b:195). (If this error is reasonably constant, however, we can plausibly assume that the dialect groupings implied by the data are essentially the same as those we would find if we had the true population values. By using multiple thresholds to determine groupings, rather than a particular level of intelligibility, we are able to make generalizations in spite of measurement error.)

(2) Trique (Casad 1970). The Trique language area consists of three principal dialects, San Andrés Chicahuaxtla, San Juan Copala, and San Martín Itunyoso. Each of the head towns of the dialect areas falls into a separate administrative district, Putla, Tlaxiaco, and Juxtlahuaca, Oaxaca, respectively. The Chicahuaxtla and Copala dialects are roughly equal in size; each is composed of approximately 6,000 speakers. The San Martín dialect is much smaller, about 1,500 speakers.

9. An alternative to this would be to redefine the population as consisting entirely of men. This would seem to be a vacuous move, however, and misses the point that what we really need is for the sample of subjects to adequately represent the entire population with regard to its overall ability to understand another dialect as the result of normal intragroup contact.
There is marketing of goods and an exchange of labor between some of the barrios of Chicahuaxtla and Copala. In addition, there are weekly markets in these two towns. However, there is no market in San Martín, so that people of this dialect are rather heavily dependent on the resources of other Trique dialect areas and of Spanish-speaking communities. The pattern of migration and intergroup contact, then, is as follows: many people from San Martín attend the weekly market in Chicahuaxtla; a few from San Martín occasionally attend the market in Copala. Many people from Chicahuaxtla have done seasonal work in the lowland Trique coffee fields of Tilapa, which is a barrio of Copala. In addition, people from some of the Copala barrios frequently attend the market in Chicahuaxtla, rather than the one in San Juan Copala, partly as a result of intergroup jealousies, and partly because there is a road and public transportation to Chicahuaxtla. Furthermore, few people from either the Chicahuaxtla or Copala dialect areas ever go to San Martín. Finally, women travel almost as often as men within the Trique dialect areas.

During the study, both a standard 250-word list and a 100-word lexicostatistic list were collected in each of the three dialects. Barbara E. Hollenbach computed the percentage of cognates between each pair of dialects for both lists. The 250-word list, adapted especially for the Trique study, showed a consistently lower percentage of cognates than the 100-word lexicostatistic list. However, the relative relationships of the three pairs of dialects are the same for both counts. Chicahuaxtla and Itunyoso are most closely related; Copala and Itunyoso are next closest; and Copala and Chicahuaxtla are least related. (The count for the 100-word list is actually based on 94 words. Hollenbach collected the 100-word list for San Martín from a pottery vendor who had come to the marketplace in Copala. The informant seemed more interested in selling pots than giving out words, so that Hollenbach was unable to check six words that she questioned, therefore she omitted them from the computations [personal communication].) The results of the lexicostatistic count appear in the following table.

<table>
<thead>
<tr>
<th></th>
<th>250-word list</th>
<th>100-word list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copala-Chicahuaxtla</td>
<td>65%</td>
<td>78%</td>
</tr>
<tr>
<td>Copala-Itunyoso</td>
<td>70%</td>
<td>84%</td>
</tr>
<tr>
<td>Itunyoso-Chicahuaxtla</td>
<td>77%</td>
<td>87%</td>
</tr>
</tbody>
</table>

*Based on 94 words

Table 9. Percentages of Cognates Between Trique Dialects

Intelligibility tests were administered at each of the following five test points: San Martín Itunyoso, San Andrés Chicahuaxtla, La Laguna Chicahuaxtla (a barrio of San Andrés), Sabana (a barrio of San Juan Copala), and San Miguel Copala (another barrio of San Juan). The test results are presented below in Table 10. The reference tapes are indicated down the left hand side of the table; the test points are indicated across the top. Several pairs of test scores for the three dialects show the non-reciprocity of intelligibility: San Martín subjects scored 98% on the Chicahuaxtla test tape; Chicahuaxtla subjects scored 83% on the San Martín test tape. San Martín subjects also scored 92% on the Sabana
(Copala) test tape, while Sabana subjects scored 64% on the San Martín test tape. Finally, Chicahuaxtla subjects scored 74% on the Sabana test tape, but Sabana subjects scored only 57% on the Chicahuaxtla test.

<table>
<thead>
<tr>
<th>Reference Tape</th>
<th>Test Point</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>It</td>
</tr>
<tr>
<td>Itunyoso</td>
<td>99</td>
</tr>
<tr>
<td>Laguna</td>
<td>98</td>
</tr>
<tr>
<td>Sabana</td>
<td>92</td>
</tr>
</tbody>
</table>

Table 10. Intelligibility Scores: Trique Study

- Juxtlahuaca
- Sabana
- Copala
  - San Martín
  - Tlaxiaco
- Laguna
  - San Andrés
- San Miguel
- Putla

Figure 5. The Trique Area

The sociological and linguistic facts cited above seem to fully account for both the degree of intelligibility between the pairs of dialects and the non-reciprocity of the scores. The dependency of San Martín on the other two dialects is paralleled by the high intelligibility that subjects there registered on the test tapes from Chicahuaxtla and Copala. San Martín subjects scored higher on the Chicahuaxtla tape than on the Copala tape; San Martín people go to Chicahuaxtla more often than they go to Copala. Chicahuaxtla-Laguna subjects understood the San Martín tape better than the Sabana-San Miguel people did; this also probably reflects the greater social interaction between Chicahuaxtla and San Martín. Laguna people understood the Copala tape better than Chicahuaxtla people did. Both from the subjects' own admission and an investigator's personal acquaintance with some of the subjects, we learned that several of the Laguna subjects had worked in Copala whereas the Chicahuaxtla subjects had not. (Many Chicahuaxtla people go down to work in lowland areas south of San Andrés.) Presumably, had testing been done in the barrios of Tierra Blanca and Tilapa (outliers of San Juan Copala), the scores of Copala subjects for the Chicahuaxtla test tape would have been higher.
Several assumptions regarding measurement error also need to be stated. Errors due to the investigator’s lack of facility in the local dialect can safely be assumed to be negligible. In the San Martín-Chicahuaxtla area, the primary tester, Bruce Hollenbach, who speaks the Copala dialect fluently, was assisted by friends who have learned to speak the Chicahuaxtla dialect.\textsuperscript{10} In addition, the degree of dialect divergence encountered was appreciably less than that of the Mazatec dialects. Almost all the subjects tested were cooperative; error due to negative attitudes can therefore be assumed to be negligible. As in the Mazatec study, the subjects were men, except for one woman who was tested in Copala. Since Trique women travel to other Trique dialect areas almost as often as men, the error due to a restricted sample is probably considerably less than that for the Mazatec study, though it still would have showed up had we included several women in each sample. Finally, a small amount of systematic error can be attributed to subject maturation during the testing. In some cases subjects performed significantly better on the final third of a test tape than they did on the first third (see pp. 113-115).

(3) Yamagiwa’s study. In a paper published in 1967, partly as an answer to Hans Wolff’s critique of Voegelin and Harris (Section 3.2), Yamagiwa presented the results of a study of intelligibility among Japanese dialects. In addition to answering Wolff, he sought to determine the relative distances between pairs of dialects, the linguistic phenomena that impair a person’s understanding of a non-native dialect, and the sociological factors that underlie intelligibility (1967.8, 14, 16).

Several characteristics of the experiment are especially interesting. Yamagiwa restricted his sample to a set of sixty-five university students and graduates, forty-two from the Tokyo area and twenty-three from other areas. Because these subjects were academically sophisticated, Yamagiwa was able to evaluate responses in a way that was quite different from earlier intelligibility studies. The students were asked to make a written translation of short portions of speech recorded in ten different dialect areas. They heard three repetitions of each recorded speech sample. The written translations were then compared with translations into Standard Japanese that had been checked by two scholars at the Tokyo University of Foreign Studies. Scoring was based on the number of items left untranslated or mistranslated and, apparently, on the number of characters in the student’s translation that agreed with the characters in the Standard Japanese translation (1967.4). Finally, Yamagiwa presented the students with a questionnaire to obtain data with which to test the validity of the intelligibility scores. Each student was asked to state which dialect he found most difficult to understand, which he found easiest to understand, and why he considered a particular dialect to be easily understood.

A summary of the intelligibility scores shows that the Kyoto dialect was overwhelmingly the most widely understood. In addition, Yamagiwa points out that the students generally best understood the dialect that was spoken in the areas where they had resided for the longest time. A comparison of the mean scores of Tokyo area students with the mean scores of students from other areas showed that both groups were essentially agreed as to the relative order of difficulty they experienced in understanding the various test

\textsuperscript{10} Claude Good and Lester Blank, Mennonite missionaries who reside in San Andrés and Laguna, respectively.
tapes. The most understood dialects were from the center of the country; the least understood dialects were in pockets along the periphery (1967.4-6). Yamagishiwa showed that the questionnaire data consistently agreed with the intelligibility data regarding the dominance of the Kyoto dialect and in regard to the minor dialects least understood. The students' explanations for their understanding of a particular dialect included the following: (1) they had heard it on radio or television, (2) they had lived in that area for a long time, (3) they knew many people who came from that area, (4) their parents were born in that area, and (5) they had travelled to that dialect area (1967.8). Finally, he points out that the area of the well-understood dialects is also the area where the political, economic, and major educational activities are centered.

While Yamagishiwa was justified in claiming a degree of validity for his study, it is fair to question certain aspects of both the internal and external validity of the experiment. In the first place, it gives no way of estimating the error that is due to the particular criteria that were used in scoring the responses. Had he used a hometown test tape for each student regardless of whether or not the student was a native speaker of a local dialect as distinct from modern standard Japanese, and had he taken equivalent samples of students from each dialect area, he would have been able to compare the performance of each sample for the study.11 As it was, he exhibited only two sets of scores that can be considered "hometown" scores in any sense of the term, i.e., students from Hyogo and Osaka on the Osaka test tape (39%), and students from Aichi and Gifu on the Aichi test tape (51%). In addition, it could have been very enlightening to have used a second method of scoring intelligibility on these texts; that is, he could have asked questions about the content of selected portions of each test tape and then estimated the degree of intelligibility from student responses to these content questions. The two sets of results, then, could have been correlated. Another point is that all of the students tested were at that time attending Keio University (1967.5). One wonders how much less agreement there would have been between scores of Tokyo area students and non-Tokyo area students had Yamagishiwa tested people in the areas where they originally resided. The scores probably represent fairly well the relative intelligibility among Japanese dialects; it is obvious from the data cited above that this intelligibility is almost entirely sociologically determined, not linguistically. Even for the least understood dialects, for which Yamagishiwa considers intelligibility to be a valid measure of dialect distance, he does not provide an alternative measure of dialect distance to correlate with the intelligibility measure.

Yamagishiwa analyzed four kinds of errors in translation. First, vocabulary items that were rarely used or not used at all in standard Japanese were omitted from the students' translations or they were misinterpreted. In other cases, dialect words were confused with phonetically similar but semantically dissimilar standard language words. Sometimes a student misunderstood an entire passage simply because he misunderstood a single word or phrase. Fourth, misunderstanding occurred when a student confused phonetically similar but grammatically distinct particles (1967.8-10).

11. This assumes that (1) the sample mean for the hometown test tape should approximate 100% (Biggs 1957.61) and (2) any variation between pairs of samples on their respective hometown test tapes that is statistically significant has to be considered non-random variation.
Unfortunately, the student was marked wrong if he inserted extra words into the translation or if he used different characters from those that were used for the standard grading translation (1967.11). It seems probable that Japanese allows for enough synonyms and paraphrases that Yamagiwa's mechanical scoring procedure could have been erroneous.

Yamagiwa's answer to Wolff does demonstrate that in a number of cases variables that could lower intelligibility scores did not affect them in the Japanese study. Furthermore, the data from the Japanese dialect study simply underline what Wolff said, that intelligibility is primarily determined by extralinguistic factors and is therefore invalid as a measure of dialect distance (cf. Kirk 1970.205).

4.3.3. Estimating the threshold range. Initial studies have been made to try to justify a particular threshold value or range for a given version of the measuring instrument. These were made to test the plausibility of the 80% intelligibility threshold value that has been postulated for the more recent versions, i.e., those used after the Tzotzil study. The data cited here from the Mazatec and Zoque studies illustrate both the methods used in justifying a particular threshold value and the importance of providing an empirical justification for the threshold used in the interpretation of results obtained from each test design and research situation.

(1) Kirk's test of informant opinions. Kirk's approach was to correlate informant opinions with intelligibility results. He wanted to know the extent to which informant opinions about the relative difficulty of dialects other than his own would support the results of intelligibility testing. At each of the twenty-three test points Kirk asked his informants, independently of one another, two questions: (a) In which towns do you find the dialect easy to understand? (b) In which towns do you find the dialect difficult to understand? (Kirk 1970.210) He recorded their responses, noting which dialects they considered easy to understand and which difficult. He recorded the responses as "easy" (+) or "difficult" (0). These evaluations were entered in an array in which the rows represented the towns the informants came from and the columns represented the dialects about which the informants made their evaluations. Kirk added two columns to the right of the matrix; he labelled these as "Easy Range" and "Difficult Range," respectively. In the "easy range" column he copied the sample mean scores for all of the test tapes that came from dialects that the informant judged easy to understand. In the "difficult range" column he copied scores for the tapes that came from the dialects that were judged difficult. For example, the informant from San Lucas judged San Jerónimo and Huautla to be easy, but Mazatlán and Huehuetlán to be difficult. Sample mean scores of San Lucas subjects on these test tapes were 97% and 80% on the easy dialects and 58% and 32% on the hard ones. The informants' judgments about these dialects were entered in the corresponding cells of the opinion table and the sample mean scores were entered into the two range columns. The opinion data and the corresponding intelligibility scores are presented in Table 11.12 Each row apparently represents the opinions of one informant. (Kirk does not say whether he obtained opinions from several informants or what he did with conflicting opinions that he got at a test point, if he did get them at all.)

12. Table 11 is adapted from one used in a pre-publication draft of Kirk 1970.
To summarize the data: Informants from twenty-three test points rendered a total of sixty-three judgments about the relative ease or difficulty they experienced in understanding eleven Mazatec reference dialects. Of these, twenty-nine were judgments of "easy"; the corresponding sample mean scores of intelligibility are entered in the proper column of Table 11. The thirty-four judgments of "difficult" are also paired with their respective intelligibility scores which are entered in another column of Table 11. A

<table>
<thead>
<tr>
<th>Informant Judgment</th>
<th>Intelligibility Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy Range</td>
</tr>
<tr>
<td>Hu</td>
<td></td>
</tr>
<tr>
<td>Mt</td>
<td></td>
</tr>
<tr>
<td>As</td>
<td></td>
</tr>
<tr>
<td>Mg</td>
<td></td>
</tr>
<tr>
<td>Tg</td>
<td></td>
</tr>
<tr>
<td>Mt</td>
<td></td>
</tr>
<tr>
<td>As</td>
<td></td>
</tr>
<tr>
<td>In</td>
<td></td>
</tr>
<tr>
<td>Ix</td>
<td></td>
</tr>
<tr>
<td>Ay</td>
<td></td>
</tr>
<tr>
<td>Oc</td>
<td></td>
</tr>
<tr>
<td>An</td>
<td></td>
</tr>
<tr>
<td>Zq</td>
<td></td>
</tr>
<tr>
<td>Ac</td>
<td></td>
</tr>
<tr>
<td>Te</td>
<td></td>
</tr>
<tr>
<td>Lo</td>
<td></td>
</tr>
<tr>
<td>Zc</td>
<td></td>
</tr>
<tr>
<td>St</td>
<td></td>
</tr>
<tr>
<td>Hn</td>
<td></td>
</tr>
<tr>
<td>Ms</td>
<td></td>
</tr>
<tr>
<td>So</td>
<td></td>
</tr>
<tr>
<td>Ja</td>
<td></td>
</tr>
<tr>
<td>Cq</td>
<td></td>
</tr>
</tbody>
</table>

Table 11. Informant Opinion of Relative Degree of Intelligibility

Informants in test points listed down the left side of the array indicated whether they considered a particular dialect (indicated across the top) easy (+) or difficult (0) to understand.
row-by-row examination of the table shows us that fifteen informants considered some dialects “easy” and others “difficult.” The others either felt that all of the dialects were “easy” or all “difficult." In eleven of the fifteen cases, the scores associated with the easy-difficult dichotomy do not overlap. Furthermore, the lowest score in the “easy range” column is often considerably higher than the highest score in the “difficult range” column. From these considerations, Kirk concluded (impressionistically to be sure) that informant opinion correlated with intelligibility data (1970.210). He further generalized that dialects for which intelligibility scores measured 80% or above were also those considered easy to understand, and those for which the measured scores were under 80% were those rated difficult to understand. The data from Table 11 show that of the twenty-nine cases rated “easy” by the informants, only three were paired with intelligibility measures of 80% or less, while of the thirty-four “difficult” ratings, only seven were paired with intelligibility measures of above 80%. Thus Kirk concluded that the estimate of 80% as a threshold value for communication between dialects was valid for the Mazatec study.

Kirk’s conclusions are generally supported by tests of significance that I made with the chi-square statistic. The test of the conclusion that opinions correlated significantly with intelligibility consisted of showing that the set of “easy” scores was distinct from the set of “difficult” ones. I used a fourfold table in order to indicate how many scores in each column were higher or lower than scores in the other. (The two groups of scores are not equal in size.) The computed value of chi-square for these scores was 26.78, with one degree of freedom. Since chi-square is significant with a value of 6.635 or more at a probability level of .01 with one degree of freedom (Downie and Heath 1965.299, Table IV), we can safely conclude that the dialects judged “easy” and those judged “difficult” constituted two separate groups which did correlate with high and low intelligibility.

With regard to the threshold value, the “easy” scores were significantly higher than 80%. (The computed value of chi-square was 16.70, which is significant under the conditions stated in the previous paragraph.) In addition, the “difficult” scores were significantly lower than 80%. (The computed value of chi-square was 10.62.) It should be pointed out, however, that there are actually many threshold values which Kirk could have used to distinguish between “easy” and “difficult” dialects. In order to estimate the range of possible thresholds I used twofold tables for testing whether the scores in each column were significantly different from a particular threshold value. I found that the “easy” scores were significantly higher than 90% (the computed value of chi-square was 4.18; its significant value was 3.84 or greater with a probability of .05 and one degree of freedom), but that they were not significantly higher than 91% (chi-square = 2.8). I also found that the “difficult” scores were significantly lower than 75% (chi-square = 5.76), but not lower than 74% (chi-square = 2.94). Any threshold value within the range of 75% to 90%, therefore, seems plausible for separating easily understood dialects from those understood only with difficulty. Finally, these conclusions do not really say whether or not there is communication in the vernacular among speakers of dialects which understand one another less than the threshold values. They merely show that the threshold range is adequate for separating the two sets of scores into two distinct groups.

Kirk’s study is a valuable beginning in demonstrating the validity of the method of dialect intelligibility testing. In particular it suggests an alternative
measure of the relative degree of intelligibility between a pair of dialects. It also suggests a method for validating the threshold range of minimum communication which is postulated for the use of a given test design in a research problem. A possibility for increasing the usefulness of the method might simply be to increase the size of samples of opinions obtained at each test point. As it is, Kirk did not mention the size of any of the samples. The results would have been even more impressive had he obtained opinions from, say, ten informants at each of the test points, and then correlated their opinions with the intelligibility test scores obtained at the corresponding location. Nevertheless, the findings of the pilot study are significant.

(2) The Zoque study. A simpler test of the validity of the 80% threshold value was made by Ralph Engel during the Zoque study. The test consisted of comparing informants' understanding of written materials prepared in a reference dialect with the intelligibility scores recorded on a reference tape from there. The informants were from the Francisco León and Volcán dialects; the written materials and test tape were from the Copainalá dialect. At both points mean scores of above 80% had been registered on the Copainalá test tape. Engel, who speaks the dialect of Francisco León, adapted the Copainalá materials to the phonological system of the other two dialects and read them to informants in Francisco León and Volcán. He found that there were a large number of lexical differences between the dialects and that the local informants could not adequately understand the Copainalá materials (personal communication). The point is, then, that the threshold value must sometimes be raised substantially above the 80% level if one intends to make inferences based on intelligibility measures made by using that particular text. Since test tapes, test formats, and research conditions vary considerably, the need to validate the particular threshold range postulated for a given study seems obvious.

4.3.4. Determining acceptability of vernacular literature. It was stated earlier (p. 70), that if the investigator interprets the index of dialect extendability as a plausible indicator of the acceptability of written materials he should either include additional items in that index (items which will measure the additional variables operative in the system) or drop the measure of intelligibility entirely in favor of some other kind of test. It seems likely that an appropriate addition (or alternative) to intelligibility testing would be the test of the reaction of a literate informant to printed materials from a dialect other than his own. The informant's response, either positive or negative, could be used to estimate the probability that literacy materials printed in this other dialect would be acceptable to other members of his own speech community. Adequate controls would have to be devised so that one could reliably interpret a particular informant's reaction. (It is entirely possible that a negative reaction could be based on an informant's attitude toward some individual in the other dialect, rather than his dislike of the orthography.)

Such a test presupposes the existence of printed materials from each of the dialectal centers and people in the outlying areas who can read these materials. Since, for the most part, neither exists when the tests are made, there has been no direct method for estimating whether members of various

13. Pointed out by Sarah Gudschinsky, personal communication.
speech communities will accept written materials produced in another community.

(1) The Huave question. Milton Warkentin, during the Huave study, briefly considered the question of whether speakers of the San Dionisio dialect would accept materials prepared in the dialect of San Mateo. The results of the intelligibility tests seemed to imply that San Dionisio speakers could understand the San Mateo dialect almost as well as their own. (San Dionisio subjects scored 94% on the San Mateo test tape.) Nevertheless Warkentin knew that there were noticeable phonological differences between these two Huave dialects and that San Dionisio speakers were very sensitive to these differences. Their sensitivity to this may be a means of asserting their group identity which is overshadowed by the eminence of the other Huave dialect.

Although the dialects had not been compared closely to determine which features were common to both, Warkentin felt that relatively few differences would show up if written materials were prepared in each dialect. Because of this he suspected that speakers of the San Dionisio dialect would accept San Mateo materials. Subsequent testing of some San Mateo materials with literate San Dionisio informants tended to confirm his feeling (Stoltzfus, personal communication).

The Huave data are only tentative, especially since no materials have ever been produced in the San Dionisio dialect and we really do not know how San Dionisio speakers would react to such materials were they available. Furthermore, the differences between the Huave dialects may be somewhat greater than the data collected during the study imply (Glenn Stairs, personal communication). Finally, it would help to be able to estimate intrinsic interest values for various kinds of literature since they may be an important factor in assuring whether people will accept literature from a closely related dialect (Margaret Wendell, personal communication).

(2) Gudschinsky's test of common orthography. The Huave situation is probably common in multi-dialect areas. It is especially noticeable in the Zapotec area where practically every speech community has its own dialect variant and often maintains an open distinction between the "real" Zapotec which is spoken by members of that particular community and the Zapotec which is spoken everywhere else. In such areas it will be necessary to take advantage of linguistic features common to several dialects in order to obtain maximum extendability for a particular set of literacy materials. The more similar two dialects are linguistically, the more readily one group will probably accept written materials produced in the dialect spoken by the other group. At this point, therefore, a measure of linguistic similarity, or one based on linguistic similarity, might prove useful to the overall method.

Sarah Gudschinsky proposes a test based on similarity that seems particularly relevant to the problem (personal communication). Her test of common orthography consists of collecting a standard lexicostatistic word list from each of a number of dialects and then computing for each pair of dialects the percentage of words of the list that can be written the same way for each dialect. The measure must be based on the actual percentage of similarity between cognate words, rather than on the percentage of cognates contained in the list; thus, the method of collecting the word lists could be directed toward discovering cognates between the dialects and could ignore the standard requirements for making a lexicostatistic analysis. Analogous to the intelligibility index, this index incorporates a range of
threshold values which should be validated for specific situations. Gudschinsky notes, for example, that it is possible to devise a common orthography for Spanish and Portuguese, but the degree of difference between the two languages is just about the maximum that still would permit such an orthography. Furthermore, since both languages have well-established literary traditions, it would be futile to attempt to employ a common orthography for general use by speakers of both languages.

3) The presence of bilingualism between dialects. The presence of interdialectal bilingualism may be an indicator of which groups will accept materials from some other group. For example, a surprising feature of the testing in San Martin Itunyoso, the Trique dialect intermediate between Chicahuaxtla and Copala, was the high intelligibility measured for the test tapes from the latter two dialects. In addition, the tester noticed that, for several subjects, their responses to the Chicahuaxtla test tape tended to resemble Chicahuaxtla speech, while their responses to the Copala test tape resembled Copala speech. Furthermore, their translation of the test statements into Spanish showed that they obviously understood them. It seems probable that this readiness of Itunyoso speakers to adapt to Chicahuaxtla speech forms might also be accompanied by a willingness to accept written materials from the Chicahuaxtla dialect.

Bilingualism, like intelligibility, is asymmetrical. Therefore, the pattern of bilingualism, i.e., who learns which dialect, is also important. For instance, although Itunyoso subjects tended to mimic the statements of the other test tapes, neither the Chicahuaxtla nor Copala subjects mimicked the Itunyoso test statements.

The above comments imply that the test of an informant's ability to mimic sentences of other dialects might be a valid indicator of interdialectal learning. Gudschinsky points out that recent research in psycholinguistics has demonstrated that ability to mimic sentences of a different dialect is dependent upon one's knowledge of both the grammatical structure and the phonological structure of that dialect (personal communication). Thus it may be valid to reinstate for certain purposes the sentence repeat test used by Crawford in the Mixe study (Section 3.3.4).
5 SENTENCE TESTS

5.0. Preceding sections of this monograph have treated the basis for intelligibility testing, the development of a test to measure intelligibility, the application of the test, and the use of survey results. Now we turn to the discussion of another possible version of the test and compare it with the current one.

5.1. Text Tests vs. Sentence Tests. The approach taken for intelligibility studies in Mexico has been to use elicited texts as the basis for constructing intelligibility tests. Earlier studies, however, used both text and isolated sentences to measure intelligibility.\(^1\) Crawford used a sentence repeat test in Mixe, but felt that it could not be adequately scored where there was a high degree of intelligibility. Bradley compared intelligibility scored on a text with that scored on isolated sentences and noted that the isolated sentences gave significantly higher scores than the text (personal communication). Crawford and Stoltzfus subsequently prepared a sentence test along with a text, but they abandoned it after using it at one test point, where the sentence test scored higher than the text. In successive studies sentence tests were abandoned in favor of text tests modified to yield higher raw scores. Besides, there was a feeling that elicited texts would be more natural and less affected by the eliciting language than isolated sentences. Finally, the absence of context occasioned by the use of isolated sentences was considered a weakness in the method.

5.2. The Purpose of the Choapan Experiment (Casad 1969). The Choapan Zapotec study, conducted by Casad and Lyman, used both the sentence test and the elicited text approach in order to compare the two. The comparison was made not only to determine whether the scores differed significantly, but also to look for differences between other factors in testing such as ease of administration and test sensitivity.

1. For instance, the Hickersons and Turner (Section 3.1.2) and Biggs (Section 3.1.4).
The need for this comparison was felt for two reasons. First, the earlier comparisons were made only between the sentence test and earlier models of the elicited text design. In one case, the evaluations were based only on the ability to repeat the content of both sentences and text. In the other cases, the comparison was made only for subjects at one test point. In all these instances the sentence test scored higher than the elicited text. The reason was probably that the subjects were able to remember the content of short stretches of speech better than the content of a connected discourse. Later modifications in the elicited text test were added to control the effects of factors such as lapse of memory on subjects' scores. There were no subsequent comparisons between the two designs.

The second motivation was the persistent problems observed during administration of the test using the elicited text approach. One problem, already mentioned, was a subject's tendency to repeat the questions rather than to relate them to the story and answer them. This showed that he did not understand what was going on. During both the Mazatec and Totonac studies, some subjects figured out that the answer to a question was contained in the preceding statement, and from then on relied completely on these statements. The administration of the tests to these individuals became simple enough, but hearing the text seemed to be completely irrelevant to the subjects' responses. In other words, these subjects were reacting to the dubbed statements as though they were isolated sentences. For other subjects there seemed to be some interference between the wider context of the story and the dubbed statement. This was manifested by ambiguous answers that may have pointed to different parts of the text. There was no certainty as to what prompted a particular response. Although this type of response can be caused by a poorly worded question, it is fair to add that even when questions are made as specific as possible, the total content of many texts still allows ambiguities in the subjects' responses. The suggestion drawn from these observations, then, was to eliminate the text from the test tape and use isolated sentences followed by short-answer, content questions.

A third motivation was that some investigators were disturbed by the variations among the texts that were used. Carl Wolgemuth, for example, felt that the texts used for the Mecayapan area of the Aztec study were not comparable; they varied too greatly in the number of semantic categories that were sampled via the ten questions, in difficulty for the subjects, and in their content (personal communication).

The sentence test was designed to bring a greater comparability among the tests than had been encountered between texts used for constructing tests in previous studies. On the other hand, the types of structures, the range of semantic categories sampled, and frequency of distribution of these structures and categories, were expected to approximate those encountered in the design which employed elicited texts. Some semantic categories are either systematically ignored in the texts or at least appear only infrequently. For example, conditional statements seldom occur. Simple declarative sentences, on the other hand, are the most numerous. It is probably easy to construct a sentence test that is more difficult for the

---

2. This is probably handled to a large degree by the modification that Stoltzfus added for Zoque (Section 3.4.5). Earlier models are cited primarily to demonstrate the reasoning behind the use of the sentence test.

3. For a fuller treatment of these problems, see Section 6.1.
subject to take than the typical test constructed from a text merely by
overloading the sentence test with questions about conditional sentences or
other types of information not usually found in the texts that we use. 4

5.3. The Construction of the Tests. Two kinds of tests were constructed
for the study; text tests of the format that Stoltzfus employed in Zoque
(Section 3.4.5), and tests based on ten statements elicited in isolation, each
followed by a short-answer, content question. The sample story for the
introduction tape was lengthened from the version Kirk used in Mazatec
(Section 3.4.4), and the number of questions on it was increased from three
to five. The sentence test was analogous to the section of the test tapes in
earlier studies where statements from the text were paired with their
respective questions. In these earlier studies, this section of tape was
spliced onto another tape which contained an uninterrupted copy of the text
from a reference point. In the Mazatec experiment, a subject heard both
sections of the test tape. In the Choapan sentence test, the subject heard
only ten sentences with a question about the content of each one.

The content of the tests was specified subjectively. Solely on the basis of
my own impressions about what kind of information occurs in an adequate
text and how often it occurs, I decided what sentences to use, which
categories of information to test for, and how many times to include a
particular category in each test. Although I am thoroughly acquainted with
the characteristics of the texts used in the studies, this method may have
biased the test design. A more adequate procedure, which I employed in a
later study, involved examining a large number of texts, categorizing the
information which was tested for each one of these texts, and noting the
relative frequencies with which each category occurred in a ten-question
test. In designing the tests, then, I constructed several sentences for each
category, and selected a certain number of them in the proper proportion
for as many sets of ten sentences as were needed for the test tapes (see
Appendix B).

I categorized the sentences and set the number of occurrences as
follows: one temporal statement, one kinship term, one body part, one
locative, two adjectives, two causatives (e.g., reason, direct cause, etc.), and
two events. This categorization was constant for the four sentence tests
used in the Choapan study.

According to the categorization placed on the design of the sentence
tests, I formed ten different sentences from my own knowledge of Spanish.
I wrote each of these sentences down in the form of frames with
substitution slots. I chose substitution lists of several items for use with
each one of the sentence frames. For example, the original sentence might
have been "The men went to the corn field." This sentence would then be
made into a locative sentence frame, "The men went to ______." The
substitution list might include the following: the river, the corn field, their
houses, the oak grove, and the town hall. For the sentence test, one of
these items would then be selected to form a complete sentence, and the

4. This does not imply that these other kinds of information are unimportant. It is only to
say that if we are going to compare sentence tests and text tests, it will be easier at first
if we reproduce in the sentence tests some of the same kinds of systematic biases we
have in the text tests—that is, if we are unable to eliminate the systematic bias from the
text tests. This is because there are so many variables operating at once that it is difficult
to isolate them and measure them individually.
resultant sentence, which might be "The men went to the oak grove," would be translated into the speech of a particular dialect. For the Choapan experiment, the selection of items from the substitution lists was done by Lyman. For reasons discussed earlier, this choice is considered weak from the standpoint of additional bias that may have been introduced into the test by Lyman's prior linguistic knowledge of the dialects in the area.

The forty sentences, then, which were used in the sentence tests, were formed by selecting items from a substitution list and inserting them in the slots of each of the ten basic sentence frames. It was also necessary to ensure that no identical sentences were found among these forty resultant sentences. This involved two steps. First, each of the ten frames was given two to four alternate forms. For example, the frame "The men went ______," might also be stated as "The young boys went ______," "The old women went ______," and "My brother went ______." The substitution list for the first frame, i.e., the river, the corn field, their houses, and the oak grove, might be changed as follows for the second frame: the market place, the spring, the top of the mountain, and their houses. Second, items selected from the substitution set for a particular frame were not included in the substitution list used with other variants of that frame. In this way I made sure that there would be no two questions in the entire set of four test tapes which could be answered the same way, even though the frames were different. In order to prevent bystanders from associating answers with a particular test tape or with a particular question, the order in which the ten basic frames were distributed within a test tape was scrambled for each of the four test tapes. For example, the statements 1, 2, and 3 for one test tape might be 4, 9, and 1 for a second, 2, 5, and 8 for a third, and 10, 7, and 1 for the fourth. Finally, the order in which the individual test tapes were presented to successive subjects generally followed the method in which the hometown test was rotated with the reference point test tapes. This method was selected because both kinds of tests were administered to the same subjects; the elicited text tests were generally administered to a subject before the sentence tests. Regarding these subjects, then, it was assumed that they had adequately learned how to take the tests. For subjects to whom the sentence test was administered before the text test, the hometown sentence test was presented as the first test tape.

After the sentences were selected and questions were formulated for each of them, both sentences and questions were elicited from an informant and recorded on tape. Lyman judged the acceptability of the materials from his own knowledge of one of the Zapotec dialects and from additional information he learned about the various dialects during the course of the study. Both Zapotec and Spanish were used for eliciting the test materials. A control group of speakers was not available to check the procedures.

5.4. Test Administration. The survey team had originally planned to administer the two sets of test tapes to alternate subjects and test fifteen subjects rather than ten. We felt that if we made the first test set longer we would tax the subjects' endurance. Some subjects would probably lose interest and respond poorly because they were tired, in a hurry to do something else, or were uneasy about the procedure. As the testing progressed in Comaltepec, however, it became evident that it would be more feasible to test five subjects on the entire set of tapes than it would be to obtain five additional subjects. To make sure that the overall length of the test did not hamper these subjects' performance, we gave each one
a brief rest period after the first test set was completed. Also, we did not try to administer all the test tapes to any subject who seemed particularly uncooperative or nervous. The sets of tests were alternated between subjects so that some of them received the text tests first and others received the sentence tests first. In this way Lyman was able to judge the relative ease with which subjects were able to learn how to take the tests. The length of time required for administering the tests was also recorded.

5.5. Test Results. The results of the study are presented in Table 12, as raw data matrices. As usual, reference points are indicated down the left side of the table and test points along the top. Hometown scores are on the diagonal. All of the reference point tapes were administered at all the test points, except that the sentence test for Xochiapan was administered only in Comaltepec and Arenal Grande. Both kinds of tests were administered to ten subjects at Jalahuí, eight at Comaltepec, and five each at Choapan, Xochiapan, and Arenal Grande. Text tests were administered to ten subjects at each of these test points.

<table>
<thead>
<tr>
<th>Reference Points</th>
<th>Texts</th>
<th>Test Points</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Co</td>
<td>Ch</td>
<td>Ja</td>
</tr>
<tr>
<td>Co</td>
<td>100</td>
<td>92</td>
<td>98</td>
</tr>
<tr>
<td>Ch</td>
<td>95</td>
<td>100</td>
<td>97</td>
</tr>
<tr>
<td>Ja</td>
<td>89</td>
<td>94</td>
<td>100</td>
</tr>
<tr>
<td>Xo</td>
<td>86</td>
<td>98</td>
<td>98</td>
</tr>
</tbody>
</table>

Table 12. Raw Data Matrices: Choapan Zapotec

A brief consideration of Table 12 shows that mean scores on sentence tests tend to be lower and more variable among themselves than text test mean scores. This tendency is substantiated in a general way by a statistical comparison of the two sets of scores.

I made the comparison to determine whether the differences were significant. I also wanted to find out whether any significant differences that I might find were in some way systematic. To determine this I made tests to try to refute three hypotheses: (1) subjects did not score significantly higher on one text test than on another; (2) subjects' scores were not significantly different from one sentence test to another; and (3) subjects' scores between corresponding text and sentence tests were not significantly different.

To make the tests I used the chi-square statistic. In four of the tests I used the Wilcoxon T along with it. I stated the significance level at .05. (Since this was a preliminary study I did not want to state a higher level. In two cases I show that chi square is significant at levels lower than .05; in most cases, however, the results are the same for levels above .05.) Finally, I based the tests on individual subject scores rather than on sample mean scores. Rather than classify subjects' scores by both reference tape
and test point, I grouped them all together by reference tape only. This allowed me to use a larger sample of observations and carry out a greater variety of tests.

The test to determine whether the test tapes differed significantly among themselves used a six-by-two table in which each of the rows corresponded to a pair of test tapes, A and B. The columns indicated how many subjects' scores on B were the same as or different from A. For the text tests, chi square was 3.598. Since this is less than 11.070, the significant value of chi square at a level of .05 with five degrees of freedom, I conclude that subjects generally scored equally well on all four text tests. On the other hand, chi square for the sentence tests was 18.659. Subjects' scores in this case did vary significantly from test to test.

The next step was to determine the pairs of sentence tests for which subjects' scores differed significantly. For each pair of tapes A and B, I tallied the number of tie scores and different scores (B is higher or lower than A) and computed chi square from the ratio of tie to non-tie scores. The results of these tests are presented in Table 13. They show that subjects understood the Choapan and Jalalui sentence tests equally well, that they probably understood the Choapan sentence test significantly better than they understood Xochiapan, and that they understood both Choapan and Jalalui significantly better than Comaltepec and Xochiapan.

<table>
<thead>
<tr>
<th>Test Tapes</th>
<th>Computed Value</th>
<th>Significant</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch and Co</td>
<td>17.442</td>
<td>Yes</td>
<td>Different</td>
</tr>
<tr>
<td>Ch and Ja</td>
<td>.122</td>
<td>No</td>
<td>Same</td>
</tr>
<tr>
<td>Ch and Xo</td>
<td>2.768</td>
<td>No</td>
<td>——5</td>
</tr>
<tr>
<td>Ja and Co</td>
<td>7.758</td>
<td>Yes</td>
<td>Different</td>
</tr>
<tr>
<td>Ja and Xo</td>
<td>4.924</td>
<td>Yes</td>
<td>Different</td>
</tr>
<tr>
<td>Co and Xo</td>
<td>4.924</td>
<td>Yes</td>
<td>Different</td>
</tr>
</tbody>
</table>

Table 13. Subject Scores on Sentence Tests

The final series of significance tests compared subjects' scores on text tests with their scores on sentence tests. Since I felt that hometown scores for text tests and sentence tests would probably be equivalent, I considered only scores on test tapes from outside dialects, e.g., I did not compare scores of Comaltepec subjects on their hometown test tapes, but I did compare their scores on the Choapan, Jalalui, and Xochiapan tapes. By

5. The computed value 2.768 is significant at a level of .10, however. It is based on only thirteen pairs of scores. I suspect that it would have been higher had a larger sample been obtained.

6. The computation of chi square is based on a formula that takes Yates's correction into account. I did this because chi square is likely to overestimate the significance level when the expected frequency is ten or less (as with the sentence test pairs that include Xochiapan), and when there is only one degree of freedom (cf. Downie and Heath 1965.166).
grouping together scores obtained at all the test points, I was able to base
the significance tests on samples which ranged from thirteen to thirty-three
observations. Otherwise I would have had to base the tests on samples
ranging from five to ten observations. Since the chi square test measures
significance in terms of a dichotomy such as yes-no, same-different,
higher-lower, I also used the Wilcoxon T statistic, which is sensitive to both
the degree of difference and the direction of the relationship between two
groups of scores. The results of both the chi square and Wilcoxon T tests
are presented in Table 14.

<table>
<thead>
<tr>
<th>Tape</th>
<th>Chi Square</th>
<th>Significant</th>
<th>T</th>
<th>Significant Value</th>
<th>Significant</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co</td>
<td>19.36</td>
<td>Yes</td>
<td>26</td>
<td>81 or less</td>
<td>Yes</td>
<td>24</td>
</tr>
<tr>
<td>Ch</td>
<td>.322</td>
<td>No</td>
<td>30</td>
<td>14 or less</td>
<td>No</td>
<td>12</td>
</tr>
<tr>
<td>Ja</td>
<td>.696</td>
<td>No</td>
<td>16</td>
<td>6 or less</td>
<td>No</td>
<td>9</td>
</tr>
<tr>
<td>Xo</td>
<td>4.92</td>
<td>Yes</td>
<td>6.5</td>
<td>11 or less</td>
<td>Yes</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 14. Differences in Text and Sentence Test Scores Using Chi Square
and Wilcoxon T Tests

(As in Table 13, chi square is significant at .05 and one degree of
freedom with a value of 3.841 or more. The significant value of T depends
on the number of ranked pairs used in the computation; N = the number
of pairs of scores used after all ties between text test and sentence test
scores have been eliminated.)

The results of the chi square and Wilcoxon T tests agree perfectly.
Subjects scored significantly higher on the Comaltepec and Xochiapan text
tests than they did on the corresponding sentence tests. They scored as
well on the Choapan and Jalahui sentence tests as they did on the text
tests.

With regard to the three hypotheses mentioned on p. 93, the results of
the significance tests show that the first one must be left standing; the
other two are refuted in part. Subjects did score significantly higher on
some sentence tests than on others; they also scored significantly higher on
some of the text tests than on the corresponding sentence tests. The
significance tests also show that sentence test scores are more variable
than text test scores and that this greater variation is patterned. Both
among the sentence tests themselves and between sentence and text tests,
the variation is significant for Comaltepec and Xochiapan; it is not
significant for Choapan and Jalahui.

The two kinds of tests thus give somewhat different results. For the
Choapan study I think this reflects two different things. The text test shows
that intelligibility is high throughout the entire Choapan area. This is
substantiated by Lyman's ability to converse and be understood in Zapotec
at all the test points we visited. On the other hand, the Comaltepec dialect
is distinguished from the rest of the Choapan dialect by a stock of peculiar
features. The Xochiapan dialect is distinguished by still other peculiar
features, including pronoun shifts that made it difficult for Lyman to understand Xochiapan speakers. These shifts in the singular pronouns are shown in Table 15. The lower scores on the Comaltepec and Xochiapan sentence tests may reflect the effect of such linguistic changes on intelligibility. Sentence tests may be more sensitive in this respect than text tests.

<table>
<thead>
<tr>
<th>Comaltepec</th>
<th>Xochiapan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st person</td>
<td>kya²</td>
</tr>
<tr>
<td>2nd person</td>
<td>kyo²</td>
</tr>
<tr>
<td>3rd person</td>
<td>a. kyebi²</td>
</tr>
<tr>
<td></td>
<td>b. kye²</td>
</tr>
</tbody>
</table>

Table 15. The Singular Pronouns of Comaltepec and Xochiapan

Finally, with reference to the results of the significance tests, the text test scores imply that any one of the reference points would be equally adequate for communicating to all the test points, say at thresholds of 80% and 85% intelligibility. Sentence test scores, however, imply that only Choapan and Jalahuí are equally adequate at these threshold values. Although the sentence test scores may imply a more reasonable estimate for the Choapan area, Lyman feels that Choapan materials will actually have an even more restricted range of usefulness (personal communication). They will probably be unacceptable in Xochiapan unless Lyman makes lexical and orthographic changes in them. In addition to these linguistic differences, there is open resentment between Xochiapan and surrounding towns.

5.6. Observations about Administering the Test. In comparing additional elements of the two test designs, observations were specifically made regarding two aspects of the test administration: relative ease with which the subjects learned how to take the tests, and the time required for test administration.

During the administration of the text test, some of the subjects repeated the questions rather than answering them, the same behavior noticed in earlier studies. Especially for the initial test tape, repetitions of both a statement from the text and the following question were needed to obtain a correct response from a subject. The learning experience provided by the introduction tape apparently was insufficient. Lyman felt an additional cause was the fact that the subjects did not know what to focus their attention on and were handicapped by their inability to remember everything that they had heard (personal communication). The very high intelligibility scores recorded in the Choapan area indicate that the difficulty in teaching the test to subjects was primarily linked to an aspect of the structure of the test tapes rather than to the subjects' inability to understand one of the other dialects spoken within the district.

The sentence test was administered to thirty-five of the fifty subjects tested. Some of these received the text test first, and others received the sentence test first. There was a noticeable difference in the ease with which the subjects learned how to take the two types of test. As the testing

7. I am indebted to Larry Lyman for these data (personal communication).
progressed, if a subject did not seem to respond to a text test, Lyman would begin administering the sentence test to him and afterwards return to using the text test with that subject. In this way, scores were effectively obtained for several subjects who might otherwise have been rejected. None of the subjects who received the sentence test indicated that he was confused by the series of semantically unconnected, isolated sentences. Lyman felt that one possible reason for the ease in teaching the sentence test was the similarity between the format of the sample story and that of each test tape. The additional body of material in the text test may have obscured the connection between the introduction tape and the test itself for some subjects.

The actual time of test administration was noted for some of the subjects. Mean times are recorded for two groups of five subjects: those who took the text test first, and those who took the sentence test first. The mean time required for administering the text test was 25:55 minutes for the first group and 23:13 minutes for the second group. The mean difference was 2:42 minutes. The mean time for administering the sentence test to the first group was 8:22 minutes and 10:44 minutes for the second group, the difference was 2:22 minutes. The total average time for administering both sets of test tapes was 34:17 minutes and 33:57 minutes, respectively.

The difference between the times for administering each set of tapes to both groups is due to the inclusion of the introduction tape with the set of text tests for Group A and with the set of sentence tests for Group B. With this in mind, then, it can be said that the sentence test is nearly three times faster to administer than the text test. Much of this time differential consists in the time that a subject spends listening to the text materials. The rest of this differential is for time required to repeat a particular phrase or question to a subject. Some repetitions were needed for the sentence test also, but they were less frequent and time-consuming than those required for the text test. The difference in total time between the two groups is negligible. The primary suggestion is that a survey team want to increase the sample of subjects from the present ten to, say, fifteen, then in view of time limitations it might be desirable to test using the sentence test rather than the text test, especially if the sentence test can be shown to be easier to teach to subjects and adequate controls can be applied to ensure high reliability.

5.7. Possible Uses for the Sentence Test. The sentence test method may prove to be a valuable supplement to the text method. In the first place, it might be particularly useful in areas where intelligibility is suspected to be high even though there are numerous lexical differences among the dialects. The content of an utterance often enables a subject to understand what is meant even though he does not recognize all the lexical items. In order to determine which lexical items are understood and which are not (this will probably be non-reciprocal), it may be necessary to control the effect of context on the subjects' understanding of the test materials.

In addition, we want to isolate and measure the effect of various kinds of linguistic factors in communication. To do this we might need to eliminate the effect of wider context. A test based on the understanding of isolated sentences might be a more sensitive measure of the role that lexical substitutions and semantic shifts play in intelligibility than a text test (e.g., see Biggs 1957.61 and Section 3.1.4, p. 57). We want to know what role specific differences play under specific conditions, both linguistic and
extralinguistic, in order to construct alternative measures or predictors of intelligibility.

Furthermore, if there is a significant difference between the mean scores obtained by the two methods, this difference might be taken as an index of the quantitative effect of context on intelligibility. This index, in turn, might be used for predicting scores between the two methods. For example, if considerations such as time available for the study and relative ease in administering the tests allow a survey team to use only one of the two methods in a particular study, then the index of differences between scores previously obtained by the two methods in comparable situations could be used to predict what scores would have been obtained had the other method also been used.

It has been suggested that certain kinds of structures may be intrinsically more difficult to understand than others. This idea can be tested. In this case, I think that it would be easier to make valid tests by using isolated statements. Besides controlling context, it will be necessary to control for the informant's rate of speech, his voice quality, and recording and playback quality of the test materials. It will probably turn out that results from this test will vary from language to language.

5.8. Summary. The results seem to indicate that, for the Choapan data, the sentence test may be a more sensitive measure of intelligibility between dialects than the text test. Known linguistic and sociological data support this conclusion. (Jalalui, for example, was formed by a migration from Choapan about two centuries ago. Comaltepec is isolated at the west end of the Choapan district; speakers from there seldom travel to the east end of the zone.)

The results of the study, however, are tentative. Similar studies are planned which may either validate or invalidate the suggestions made from the Choapan data. For example, the assumption that the sentence tests used in the study were comparable is partly based on the assumption that items in a class are likely to be understood equally well (Ladefoged 1968.2; Kroeber 1961). This can be tested. Other tests can be devised to measure the degree of difference between sentence tests that an investigator can build in by purposely including specific classes of items. Finally, a better comparison between text tests and sentence tests might be gained by first matching each sentence test with a text test in terms of the classes of items that are sampled by the questions about the text test. The construction of comparable sentence tests might actually be a subsequent step.
6 CRITICAL REVIEW

6.0. Improvements made with successive studies in Mexico have led to the development of a reliable measure of intelligibility between dialects. The method yields hometown scores closely approximating 100%. The high reliability is suggested by the fact that no significant variation between scores recorded for a set of test tapes from the same reference point was observed in a pilot study in the Mazatec survey. Though some important statistical questions are still unanswered, the method is felt to have some degree of validity.

In this chapter, attention is turned toward some of the problems found in the method. Though some of these are procedural and structural problems, rather than quantitative, they all must be recognized and adequately controlled if the study is to yield valid results. These problems are discussed, not to provide a basis for rejecting the present method, but rather to help ensure that further applications will be fruitfully employed and possibly to suggest modifications for future studies.

The problems will be treated in three sections: those of the elicited text approach, those of the elicited sentence approach, and additional sources of probable error in applying the method. With reference to the text test, some of the problems are more relevant to certain versions of the design than to others. For example, some problems are frequent in the Mazatec test design; but they may have been controlled to a large extent by the interrupted text version. Where this is so, it is made clear in the discussion. Though the interrupted text version is considered “standard” for the moment, reference is made to earlier designs.

6.1. Problems with the Elicited Text Approach. The problems in the elicited text approach that are discussed below are the following: test construction, text elicitation, text comparability, text sampling range, the formulation of the questions, test teachability, scoring of responses, and test sensitivity.
6.1.1. Test construction. The construction of a set of test tapes at any test point is complex. The survey team must complete the following series of steps at each test point: (1) elicit and transcribe an adequate text, (2) formulate a set of questions from the translation of that text, (3) translate the sets of questions for all the test tapes into the local dialect, (4) prepare an introduction tape, (5) submit the translations of the questions to a pre-test panel of speakers of that dialect in order to detect and correct translation errors, (6) make a dubbed copy of a hometown text for constructing the hometown test tape, and (7) record the translated questions and the introduction tape. This work is generally begun during the preliminary trip and is completed just prior to administering the test. This preparation entails a day's work. In practice the pre-test panel is often not utilized, partly for lack of time, and partly for lack of available informants and subjects. For the Mazatec design, it was also necessary to dub cue statements from the text onto a section of blank tape. The corresponding questions were then recorded on that tape following each cue statement.

6.1.2. Text elicitation. An adequate text is often difficult to elicit. The informant may be reluctant to speak about his own experiences. Or he may be unable to think of something to tell, even with suggestions and examples from the survey team. For these reasons, attempts are made to have a surplus of informants available. However, because an entire community may be suspicious of the survey team, it may be that even among a group of informants from that community none will be found who is able, or willing, to provide an adequate text.

Persuading an informant to give a text is only part of the problem. Texts with content that revolves around several people participating in a few events can be quite difficult to elicit. In many cases, it is necessary to accept a text which samples only a narrow range of grammatical structure and lexical content. In these cases, it is hard to formulate an adequate set of questions over certain semantic functions—manner, cause, purpose, and result—that are quite important in linguistic systems. It is desirable to sample as wide a range of the grammar and lexicon as ten questions will allow, and the content of each text directly influences the possible range of structures that can be sampled. Furthermore, when a text does contain this kind of information, it is common to find that there are loan words in the statement desired for use, or that the informant lapses into careless speech at a crucial point in the text. This further reduces the amount of content available for question formulation.

6.1.3. Text comparability. The texts used for constructing test tapes are not always comparable. They differ in the amount of content available for formulating questions, the range of grammatical and semantic structures sampled, interest value for the subject, naturalness of speech style, and quality of recording. Simpler texts may cover a narrower range of the language system than more complex ones. For example, several of the texts collected in the Mazatec study were hunting stories. They yielded questions regarding the size, color, number, or names of animals. On the other hand, there were few regarding peoples' motives or emotional states. Some texts are well organized in thought and are filled with adequate content to form the set of questions. These texts are often the most interesting to the

1. For the discussion on loan words, which may be normal for the language, see p. 13.
subject (Crawford 1965a.10). Other texts are neither rich in content nor high in interest. This incompatibility of texts bears on the ease in constructing an adequate test, the representativeness of the individual text as a sample of a particular dialect, and the subjects' understanding of the tapes of a test set.

6.1.4. Text as adequate language sample. The text used as the basis for an intelligibility test is assumed to be a representative speech sample of a particular dialect. The degree to which the range of structures sampled by any one text adequately reflects the range of structures in the total body of speech from which the text is taken can be said to be a measure of the content validity of that text (Downie and Heath 1965.223). Assuming that a subject's understanding of a dialect can be estimated by a measure of his understanding of a speech sample from that dialect, then, it is also assumed that the accuracy of that estimate will vary to some extent with the content validity of the speech sample. Therefore, it is suggested that an increase in the range of structures sampled by a text will be accompanied by an increase in the accuracy of the measure of intelligibility that will be obtained from a sample of scores on that text (Wilson 1952.252-54).2

The cause for concern is that the sampling ranges of texts vary greatly between themselves, and this variation may be significant. For example, in the Zoque study, a text from the dialect of Copainalá was played to a sample of subjects of the Francisco León dialect. The subjects scored relatively high on that text, i.e., 82.5%. Ralph Engel, who speaks the Francisco León dialect, felt that the intelligibility score was higher than the actual level of intelligibility existing between the two dialects. He had previously read materials translated into the Copainalá dialect to speakers of the Francisco León dialect. These translated materials were not well understood; there were too many lexical shifts between the two dialects. On the other hand, the text used as a measure of intelligibility was quite simple in content and the speaker's style was a slow rate of speech clearly enunciated (personal communication).

The concept "sampling range" is restricted to those structures of the text which are actually used to determine the measure of intelligibility recorded for subjects who respond to that text. This definition implies that portions of all texts used are irrelevant to the tests. The question here is not whether another text will yield the same or different mean score for a sample of ten subjects, as was the case in the Kirk study (see pp.110-111). Instead, it is whether a particular text is a representative sample of the total body of speech of a particular dialect. If it is not, then the mean score of a sample of ten subjects on that test tape cannot be said to represent the level at which speakers of dialect A understand speakers of dialect B. If it is, then the mean score can be said to estimate the level of intelligibility between any two dialects within a certain range of error.3

6.1.5. Questions. The characteristics of a particular text determine to a large degree the selection and formation of the set of questions concerning

2. Note, however, that if the first assumption is wrong, i.e., if we cannot estimate a subject's actual understanding of another dialect by how well he understands a text from there, then it is pointless to attempt a more accurate measurement by increasing the sampling range.

3. This also assumes that the sample of subjects from dialect A is a representative sample of the population of speakers of dialect A. The range of error will depend upon both the size of the sample and the variability of their scores.
its content. Occasionally this interplay has been overlooked, and poorly stated or chosen questions have been introduced into the structure of the test tapes. In addition, it is usually feasible to frame the question to reflect the structure of a particular statement in the text, or to phrase the question to refer unambiguously to that statement. However, it is not always possible to avoid overloading the test with questions about one particular class of items, e.g., color, location, or quantity.

The most serious problem with the questions used in the test tapes is that they are isolated utterances generally elicited through Spanish and translated into a dialect unfamiliar to the investigator. Thus, the structure of these questions may be biased toward the structure of Spanish. Generally, the unnaturalness of these questions has been adequately controlled by the investigator's previous knowledge of a related dialect, or by the use of a pre-test panel of speakers to judge the quality of the translated questions.

An additional problem in eliciting these questions has frequently been that an informant did not fully understand his role, so that instead of repeating in his dialect the question he heard in Spanish, he would answer that question or transform it into a statement. Since informants differ in their abilities, some quickly learn what is expected and others do not seem able to adapt to the particular elicitation procedure. In the latter case, the investigator is less sure about the quality of the translated questions and may seek other informants to check the material. On the other hand, since work with less capable informants has proved time-consuming, investigators who have felt the need to conform to a rigid time schedule have neglected to check the questions with other speakers of the dialect, and faulty questions have thus been incorporated into the test.

As noted earlier (p. 14), the initial role of formulating questions is left to the technician member of the team. It has happened that when this was left to the investigator, some of the items chosen were those suspected to be of limited distribution. This biases the test to give lower scores and could yield quite misleading intelligibility data. Furthermore, in some situations, the investigator is not aware of the constraints that have been put upon the selection of these questions. These constraints are discussed on pages 12-14.

6.1.6. Test teachability. The tests have generally been administered to subjects who are limited in their academic background. The largest educational gap is between young men and middle-aged or older men. Even many of the younger subjects have had only a limited amount of education. Thus, the idea of taking a test is new to many of these subjects, and the test designs employed have been quite sophisticated for them.

For the Mazatec tests, subjects were expected to listen to an uninterrupted text and then to answer questions about that text. Each of these questions was preceded on tape by a statement dubbed from the text which was intended to help the subject remember what the text said. The subject was in effect required to relate the question to the text and the dubbed statement to both the text and the question. Failure to fully relate these elements was manifested by a tendency for some subjects to repeat questions rather than to answer them. In addition, some subjects responded with ambiguous answers that may have come from one of several statements in the text; others seemed to ignore the text completely and answered solely from the dubbed statements.
The interrupted text version was initiated to control these problems. There is no indication that the interrupting of the text for asking the questions disturbs the subjects. The main problem is that the test format requires a subject to discriminate between material necessary for answering a question and material irrelevant to that question. It is complicated by the fact that the subject does not know when the test will stop and he will hear the question. In addition, he does not know what he is supposed to listen for. The problem is not primarily that of the amount of material that he is required to remember; rather it is that of what he must focus on. The rest of the text is somewhat irrelevant.

6.1.7. Scoring subjects' responses. Two main considerations in scoring are the judgment as to what evokes a response and how to evaluate particular responses consistently. The former requires the investigator to discriminate between plausible-sounding guesses and valid responses based on synonyms or paraphrases. Therefore he must be aware of features of the text, semantic characteristics of each dialect, and even cultural associations which might prompt the various responses. This becomes increasingly difficult as the successive dialects which are being studied become more divergent from the one which the investigator speaks. In earlier studies ambiguous answers were encountered because subjects were confusing similar statements within the text. This problem has been minimized, though not eliminated, by the use of the interrupted text test.

Doubtful responses are written on the score sheet along with an evaluation, since investigators have been observed to be inconsistent in some of these evaluations. The investigator's unfamiliarity with the dialect is the primary disadvantage. The limited time spent at each test point generally does not allow sufficient learning of all the information needed for a proper evaluation of doubtful answers. Interpreters are occasionally used; bilingual subjects often are their own interpreters. In addition, the investigator usually keeps transcriptions and translations of the test materials close by for reference during testing.

The tests could be restricted to a small enough corpus of language material to enable the investigator to acquaint himself well enough with pertinent details of the new dialect to make correct judgments in the vernacular. This could be done by either eliminating substantial portions of the text from the test tape or by using sentence tests. Structural restrictions on the test might also cause the subject to focus on a particular segment of speech and to respond unambiguously to the corresponding question. The sentence test seems more amenable to these restrictions than the test test.

6.1.8. Test sensitivity. In earlier studies it was soon noticed that the test scores were reflecting not only interdialectal learning, but that they were also being affected by many individual differences such as differential I.Q., memory ability, limited educational background, and lapse of attention. As a result, hometown test scores were low, and subject scores on the individual test tapes did not follow a clear pattern. Modifications in the test design such as the inclusion of a sample learning tape and the use of cue sentences were introduced into the test design to control these factors simultaneously and to yield more reliable test results. These changes produced a greater uniformity in individual responses and higher scores at the cost of some loss in the sensitivity of the test in its measure of levels
of intelligibility (Crawford 1967). Because of the large number of dialectal areas that were being outlined by the method, and the practical aim of keeping to a minimum the number of centers that would be selected for producing literacy materials, the loss in sensitivity was not felt to be critical.

However, the reliability of a measure does not guarantee its validity. In Mesilla Totonac and Choaapan Zapotec, the areas are characterized by a high degree of intelligibility between the dialects, even though there are known systematic differences between them. The text test mean scores have differed less than had been expected due solely to random causes; thus, the rank ordering of these mean scores cannot be validated on the basis of linguistic differences and social relationships. (The overall high intelligibility can be validated, however.) It seems probable that other tests can be designed that will be as reliable as text tests but will be a more sensitive measure of intelligibility between dialects. The Choaapan experiment suggests that this is so.

Results obtained by using a more sensitive test can be interpreted in the same way as results obtained with the text test. The use of a range of threshold values to express the optimal extendability of dialects can be maintained. However, since it is likely that the threshold range will be more variable for sentence tests than for text tests, it will be necessary to validate the threshold range itself, separately, for each study in which the more sensitive test is employed. The use of a more sensitive test does not imply that more dialects would be distinguished. Instead, it could refine the first approximation furnished by intelligibility figures and come closer to pinpointing the acceptability of written materials from possibly fewer dialectal centers. The measure would continue to register scores of marginal intelligibility between certain dialects, although the range might not correspond to that set for the text method, i.e., 75-85%.

Finally, since the reliability of a test is directly proportional to the number of questions in it, the measure might include fifteen or twenty questions, rather than the present ten. This modification is possible for text tests, but is hard to implement. Reliability can also be increased by testing a larger number of subjects at each test point. Under the text approach this is difficult because of the time required for administering the test set to each subject. Also, available subjects become discouraged by the apparent difficulty other subjects experience in taking the tests. It might be possible to increase the size of the sample of subjects if a test design were employed which was easier to teach and quicker to administer than the text test.

6.2. Problems with the Sentence Test. There are four principal difficulties with a sentence test. These are (1) unnaturalness of the test sentences, (2) loss of redundancy by the elimination of context, (3) loss of linguistic structures above the sentence level, and (4) greater sensitivity to disturbing influences. There are less serious problems in constructing and administering sentence tests.

6.2.1. Unnaturalness of the test sentences. The most common objection to the use of a test based on isolated, elicited sentences is that the sentences will be unnatural in that their syntactic and semantic structures will tend to conform to that of the language used for eliciting. The assumption is made that a freely narrated text tends to be more natural in structure than elicited sentences. A second assumption, based on the first, is that the
tendency toward forced syntactic and semantic structures in the elicited sentences cannot be controlled.

While the assumption that narrative text is natural may be true generally, it is also true that many of the texts used in the studies have not been natural in their structure. The informants' speech styles, including arrangement of content and use of loan words, have been influenced by the elicitation procedures. Often informants have been reluctant to give a text and, at best, there has been a noticeable wavering in the voice quality. Some informants have used an abnormally high frequency of loan words in the text. (Occasionally an informant has repeated the same text with very few loan words for a second recording.)

Regarding the assumption that the naturalness of the syntactic and semantic structures cannot be controlled for statements elicited in isolation, it turns out that, thus far a very important part of the test design has been based on individually elicited utterances, viz., the questions for each test tape. These questions have served well; probably less than ten of the questions translated for the Mazatec study were poorly translated. This means that the elicited questions are within a range of naturalness acceptable to native speakers. This quality of translation has often been reached without the use of a pre-test panel. When a pre-test panel is used, an even higher quality has been achieved. It seems that the quality of translated statements should be as good, if not better, since it is easier to elicit a statement than a question.

With reference to intelligibility, then, there is a permitted looseness in the quality of translated materials, a range of deviance in which intelligibility is unimpaired. Our design should be required only to yield results within this range.

Several resources may be utilized to guarantee an acceptable degree of naturalness in the elicited sentences. The prior knowledge that an investigator has of one dialect of the language may be used for eliciting data in another dialect. Additional language data collected at other test points may suggest analogous structures in the new dialect. With bilingual informants, back translations of the materials into Spanish may prove helpful. Since the best control comes from native speakers, additional informants may check the elicited materials individually. It is also feasible to elicit a free narrative and abstract sentence frames from it that can be used as a pattern for constructing additional sentence tests. Finally, the investigator may not want to insist on a too exact semantic correspondence between the eliciting phrase and the elicited one. For example, in constructing the Jalalui sentence test, the informant substituted "jealous" for "intelligent," which had been specified for that test beforehand. When the testing began in Comaltepec, subjects responded to the question about that sentence with the answer "jealous." Rather than mark the responses incorrect, I recorded all the questionable responses. Later checking in Jalalui revealed that the correct answer found in the statement was "jealous."

6.2.2. Loss of redundancy. Another problem in the elicited sentence approach is the loss of redundancy by eliminating a wider context. The wider context of discourse resolves ambiguities of individual utterances and

4. For a treatment of the elicitation process, see Samarin 1967, chapters 6, 7, and 9.
heightens understanding. The effect is probably of equal value both for a person who listens to oral materials and for a person who reads written materials. Thus it is important to consider its effect on intelligibility.

On the other hand, the wider context of discourse has occasionally been confusing to subjects. In the Mixtec and Chol designs, the test tape allowed the subject to respond to a question with an ambiguous answer that may have been prompted by one of several similar statements within the text. The interrupted text version seems to force the subject to discriminate between features relevant to a following question and features irrelevant to that question, though it does not provide him with direct information about what the relevant material is or on which part of the text he should focus.

Since the wider context may include several different features that resolve a particular ambiguity, it is not always possible to know what a subject is reacting to, even though he may have answered a particular question correctly. Thus it might be helpful, or necessary, to eliminate much of the wider context of connected discourse to gain certainty in evaluating subject responses.

On the other hand, context also conditions the hearer to phonological differences between dialects. It may be desirable to include sufficient context in the elicited materials to control the effect of low-level phonological differences on subjects’ test scores. Kirk suggests using two utterances of each sentence in the tests (personal communication). This could be done by having the informant give them, or by dubbing the sentence twice from the original recording. Data from the Trique study support the suggestion to make such a modification. We included fifteen items in the Trique sentence tests rather than ten. It turned out that in several cases subjects performed significantly better on the last five items of a test than they did on the first five. This apparently resulted from the subjects’ getting used to the phonological differences between the reference dialects and their own (see pp. 113-115). From this I suspect that the Choapan text test and sentence test scores would have been more alike had we included fifteen items in each sentence test.

Finally, as noted on pp. 97-98, the effect of context on intelligibility could efficiently be estimated by carrying out a series of experiments in which both text tests and sentence tests were administered. A comparison of the differences between the scores would yield an index of the effect of context on intelligibility scores. This index could then be used to predict scores for the one method on the basis of those obtained by the other.

6.2.3. Loss of linguistic structure above sentence level. The elicited sentence approach also involves the loss of certain linguistic structures above sentence level that are commonly found in connected discourse. It is felt that most of these relations can be expressed between sentences or within clauses and that they can be incorporated into the elicited sentence approach. For example, a narrative text often contains an introductory paragraph that orients the hearer to the participants involved in certain events and the time, location, and meteorological conditions surrounding these events. These orientations can be expressed by complex sentences such as “The bandit was by the river when the authorities found him.” In some languages, relations such as cause, result, ground of activity, and purpose are expressed by juxtaposing two sentences. Sentence tests can be adapted to include pairs of sentences to express these relations. Trique, for example, exhibits such structures. They are pronounced as phonological
units (Barbara Hollenbach, personal communication). This allows one to elicit them rather easily.

There are a few structures such as plot relations, argument structure, and collateral information that are independent of sentence organization (Grimes and Glock 1970). However, these structures are not even sampled by the questions formulated under the text approach. Therefore, the loss in the range of linguistic structures sampled by using the elicited sentence approach, rather than the text approach, is probably negligible.

6.2.4. Sensitivity and reliability. Although the sentence test is more sensitive to the effects that lexical changes, semantic shifts, and phonological differences have on a subject's understanding of another dialect than the text test is, it is also more sensitive to the effects of differences between informants' rates of speech and the quality of the recorded materials. This is not so noticeable on hometown test scores as it is when the tape is tested in other dialect areas.

For example, in the Trique study we administered two tapes from San Martín Itunyoso to subjects in Chichahuaxtla and Laguna. The tapes were identical except that one informant's speech style was slow with clearly enunciated words while the other was very rapid. Ten Chichahuaxtla-area subjects averaged 87% on the clearly enunciated sentences; they averaged 64% on the ones spoken rapidly. (Five of the subjects were tested on both Itunyoso tapes. Their scores on each tape are comparable to those of subjects who were tested on only one Itunyoso tape.) The more poorly understood tape was used by itself as the hometown test in Itunyoso where the subjects averaged 99% on it. The effect of the two speech styles clearly showed up on Chichahuaxtla scores; it would not have been noticed in Itunyoso itself.

If we had used a pair of Itunyoso text tests that differed in just the same way as the sentence tests (identical content, but very slow versus very rapid speech style), I suspect that Chichahuaxtla subjects would have scored more alike on the pair of text tests than they really did score on the pair of sentence tests. However, this needs to be tested. None of the comparisons of text tests from the same reference point that have been made thus far used tests that differed as much as the Itunyoso sentence tests. This is because controls on speech style and recording quality were already part of the text test design, and the investigators customarily applied these controls in their pilot studies (see p. 110). Regardless of how text tests compare with sentence tests, however, these controls need to be applied rigidly in order to allow a reliable measure of intelligibility with the sentence test.

Although the sentence test is more variable than the text test, it may be made nearly as reliable if the individual tests are comparable in terms of content, recording quality, and difficulty, and if the number of items in the test are increased or the test administered to more subjects. In addition, it may be necessary to build in a certain amount of context so that subjects' scores are not affected by some kinds of ambiguities (see following section).

6.2.5. Test construction and administration. It is somewhat easier to construct and administer sentence tests than text tests. One main difference is that much of the work in constructing the tests is taken out of the informant sessions. The technician must select equivalent sets of sentences to use in the reference tapes. It is probably better to do this before the
actual survey trip begins in order to make sure that the sets of sentences are equivalent. Once the team is on the field, the pressure of time and other immediate responsibilities may cause the technician to become lax or fall behind in specifying comparable sets of sentences for the tests.

Sometimes an informant may not know what a particular sentence or item in the sentence means, making it necessary to invent another sentence on the spot. It is possible that the elicitation of sentence tests requires a more bilingual informant than text elicitation does. For both kinds of tests corresponding sets of questions must be elicited, however.

It is necessary to control some of the ambiguities that arise with the use of isolated utterances. Each question must be phrased specifically for the particular sentence (utterance) it refers to. For example, if the utterance is "The man saw the dog," and the question refers to the object, it may have to be phrased as "What (animal) did the man see?" rather than as "What did he see?" in order to obtain the response "dog." Otherwise, subjects may confuse agent and goal, as they did on one item of a Totonac sentence test (David Persons, personal communication). Part of this lies in the construction and elicitation of the sentences themselves. We may want to elicit some sort of context for the utterances. It is not clear to what extent we want to control for ambiguities.

Although the sentences in the tests do not have any intended relationship to one another, subjects occasionally do connect some of them. For example, one Trique sentence test included the sentences "The man went into his house because it was raining," and "The man's shirt is wet," in sequence. Subjects tended to answer the question, "What is the condition of the man's shirt?" by saying that it got wet because of the rain. So far these unplanned connections have not raised any problems in scoring because the responses have included the desired answer along with the additional information that signals the reference to a previous statement. Nevertheless, it is possible that subjects could make more tenuous connections resulting in ambiguous or obscure answers even though they understand the sentence perfectly. Because of this it seems advisable to try to set up the sentence tests in such a way as to prevent subjects from making obvious connections between sentences.

6.3. Additional Sources of Error. In applying the method there are, in addition to the problems cited above, several non-structural sources that bring error into the intelligibility results. These sources include the sampling methods used as well as psychological bias in both tester and subjects.

6.3.1. Bias in the tester. Bias in the tester can be either conscious or unconscious. An extreme case of the former could be that of an investigator who feels that the method is not valid anyway, with the result that he would not be concerned with maintaining tight controls either on test construction or on test administration. The tendency noted earlier for an investigator to load questions to favor items that he feels will not be understood outside of a particular dialect area is generally a case of unconscious bias. The point of these comments is that some tester bias is there whether conscious or not.

The time schedule that an investigator sets may impose a heavy bias on the results. Often the lack of time is the primary reason for not employing a pre-test panel of speakers to check the translated portions of the test materials. The same pressure may make it impossible to complete the sample of ten subjects at a test point.
The results of one survey were seriously impaired by the decision not to construct hometown test tapes at any of the subsidiary test points. It turned out that when a reference tape from the dialect that the investigator speaks was played at a subsidiary point only one hour’s walk from the reference point town, subjects there scored only 44% on the reference tape, much to the survey team’s surprise. (Because the towns were so close geographically they reasoned that the two dialects must be almost identical so that both dialects certainly would not be needed as reference points. Since no test tape was collected from that subsidiary point, there was no way to determine whether other dialects (including the one the investigator was studying) may have been outliers to it. Interestingly enough, people from the reference point town pass through the subsidiary point town on their way to market, but not vice versa. The decision not to construct subsidiary point hometown tapes was made in order to save time for the survey team. It saved the team two or three days work at most at the cost of adequate coverage of the area.

The complexity of the test design itself may contribute to tester bias. For example, it is possible that an investigator might decide that one aspect of the design is pointless, say, the introduction tape. A poor relationship between members of the survey team and fatigue or strain from the physical and social circumstances of the work also introduce possible bias into the results of the study. Part of the survey team’s responsibility is to ensure that these factors, to some extent unavoidable, do not adversely affect the implementation of the study.

6.3.2. Bias in the subjects. The primary causes of psychological bias in subjects are suspicion of the motives of the survey team and lack of experience in schoolroom-like situations. The motives of the survey team are understandably interpreted by many subjects in the light of previous contacts with people from the outside, which in many cases have been quite negative. The need for the survey team to gain the confidence of local residents is obvious. The use of letters of recommendation from various authorities can be helpful. However, the team’s sense of urgency to complete the study at a particular point may intensify, rather than allay, local suspicions.

A potential subject’s feelings of inferiority for his lack of formal education and for his inability to speak Spanish have often caused him to refuse to take the test. The fact that the testing situation forces a subject to perform in the presence of his friends tends to make potential subjects reluctant to cooperate. They are afraid of losing face. The tester often assures reluctant subjects that it makes no difference whether they answer the questions correctly or not, and that it is not necessary to answer in Spanish. Such assurances help, but seldom overcome in a few minutes the established culture patterns and years of inferiority feelings.

In addition, the intricacy of the test and the difficulty the subjects have in understanding the test tapes compound these problems. A subject’s reaction to the test has influenced the willingness of potential subjects to cooperate. Although the difficulty of understanding other dialects is what is being tested, it is possible to construct test tapes from dialects that are so difficult for a subject to understand that he loses interest and refuses further cooperation (Bradley 1967:5). Thus, it is necessary to construct a test design simple enough that it can be efficiently administered to
unprepared subjects. It is also necessary to choose test tapes that include materials from dialects ranging from high to moderately low intelligibility.

Finally, the lines of authority that a survey team exploits largely determine which persons are included in the sample of subjects at any test point. The groups associated with particular authorities or individuals often vary in terms of educational background, attitudes, and bilingualism. For example, subjects who have been encouraged by the school teacher may be more willing to take the tests than those selected by the town secretary, or vice versa. The more willing subjects, in turn, will probably perform better on the tests.

Thus, subject bias enters testing for a number of reasons. It should be noted again that even the use of a thoroughly random sampling technique would not sidestep these considerations. Though a random selection would allow more classical statistical treatment of the scores for a sample such as the computation of a standard error based on random variation, the fact remains that the "random" sample would consist of those local people who were willing to cooperate and who, therefore, would probably come from the less conservative segment of the population.5

6.3.3. Texts, sampling, and subjects. The method presented here utilizes recorded texts as the basis for measuring a subject's understanding of dialects other than his own. The scores of ten subjects are averaged to yield a point estimate of how well the total population from which the sample comes understands each reference point dialect. The accuracy of the method rests on at least three assumptions: (1) the texts adequately represent the speech of each dialect, (2) the subjects are representative of the population they come from, and (3) the tests are sufficiently long and the samples of subjects large enough to allow reliable measurements. The rest of this section discusses both the kinds of error that result from the methods we use and our justification for making the above assumptions.

(1) Adequacy of the texts. Constraints are imposed both on the class of texts used and on the kind of questions formulated, partly to make sure that the tests cover as wide a range of structures as possible, and partly to make sure that the tests are reasonably comparable. Generally, however, the texts used on any survey vary greatly among themselves. Thus it is pertinent to consider the possible significance of this variability. Would another text yield the same scores? What effect do differing lengths, interest value to the subject, and differing complexities of the texts have on a subject's score? Is a simple text a less adequate sample of the entire language than a complex one?

Kirk made a pilot study of the effects of some of these variations in the Mazatec town of Nuevo Soyaltepec (1970). He used four texts from the Huautla dialect, each differing in length between one minute and 55 seconds and two minutes and 55 seconds. Two different informants had each given one pair of the texts. The longest text was the standard Huautla text used throughout the Mazatec study. All four texts were administered to the same sample of ten subjects. The scores for this sample on the texts ranged from 72% to 77%. Kirk noted that the longer of each pair of texts

5. A random selection of cooperative people might be feasible. In addition, since change comes through the less conservative elements of a group, the method may be on reasonably solid ground in using the scores of cooperative subjects as a first approximation to the acceptability of printed materials published at a particular dialectal center.
scored 2% higher than the shorter, and that the texts from one informant averaged 3% higher than the texts given by the other. Wilcoxon’s T statistic was used as a test of the significance of this variation. It shows no statistical significance for these observations at a level beyond .05 (see Appendix I). For several reasons, Kirk’s experiment should be duplicated several times. In the first place, he implicitly applied some controls on the differences among the tapes. Tests should be made using texts that vary greatly among themselves for several parameters so that we can determine for each one the range of non-significant variation. Furthermore, these tests should be made in terms of various levels of intelligibility from high to low. As it is, his study shows that, in one case, the standard procedures of test construction included adequate controls on text differences such as content, length, informant’s speech style and quality, interest to the subject, and quality of the recording.

(2) Representativeness of the subjects. Regarding the subjects, it is necessary to consider the technique used for obtaining the sample of ten and the possibility of significant differences between groupings of these subjects.

(a) Sampling techniques. The ideal technique for sampling would be to use some random process of selection so that the probability of any one person’s being chosen as a subject would be the same for each speaker in the population. Such a random sample is both a guarantee against biased results and a statistically sound basis for making predictions about the population from which the sample was taken. Decisions are then made on the basis of these predictions. To the degree that a sample is not random one runs the risk of having biased results and making unsound predictions that lead to costly decisions.

This ideal, however, has never been realized in this kind of experiment. Matters of feasibility, practicality, and economics often force an investigator to choose random selection and to exercise his own judgment in selecting a sample. Such considerations have determined the sampling procedures followed in the surveys in Mexico.

A technique for obtaining truly random samples for the Mexican studies would be to select subjects on the basis of a set of random numbers applied to the census lists found in the official records of the reference and test point towns (Crawford 1965b). However, the lists are not always complete, nor are officials in many areas willing to share these records with outsiders who, for all they know, might be tax collectors. Insisting on a random sample might easily compound their initial reluctance to cooperate with the survey team. Furthermore, it would require more active cooperation from these authorities than is often available. Finally, if this method were used, it is quite likely that some of the people selected would live several hours away, and it would therefore be necessary to wait for days until they all came to be tested or until word was received that some of them were not coming and replacements could be tested.

The usual sampling technique consists in testing the first ten suitable subjects who are willing to take the test. It is assumed that these ten subjects are representative of the population. They are frequently obtained through the cooperation of the town officials. In other instances, subjects are obtained through the efforts of an informant or one of the survey team members. Often the subjects are those who happen to be around the town hall. Occasionally, the town authorities themselves have been subjects. In the Mixe study, the investigators accompanied a representative of the town authorities to an arbitrary
point within the town from which the investigators went from house to house until the quota of subjects was filled. In other studies, subjects were brought one by one to the town hall. In segments of the Mixteca study, subjects of three age groups—15-30, 31-50, and 51-60—were obtained from various geographic sectors of the test point (Bradley 1967.10). In several surveys there was at least one occasion when it was necessary to confine the testing primarily to an investigator’s personal friends. There has been no consistent pattern for selecting subjects, even within the same survey.

A suitable subject must be a native speaker of the local dialect; his first language must be that of the test point if he is bilingual; he must be reasonably sober; he must be cooperative; and he probably should score at least 80% on his hometown test tape. A few subjects have been used who scored lower than 80% on the hometown test tape, and a few subjects who scored higher than that have been disqualified for other reasons, either because the subject refused to continue the testing procedure or the investigator felt that a particular subject was not cooperative.

Relatively few women have been tested. This is partly because of their own shyness and fear of the equipment. It is also because the survey team members are generally two men and are operating in an area where they are not known. Besides, many monolingual persons have refused to cooperate; either they felt it was necessary to be able to speak Spanish or they were afraid of the tape recorders. It is also possible that some of these people have been embarrassed about being uneducated and unable to communicate in Spanish. Thus, the majority of the subjects who have been tested have been the more outgoing, somewhat bilingual males, mostly between fifteen and fifty-five years of age.

(b) Sub-groups of the sample. The subjects who are tested are from the more outgoing, less conservative segment of the population. This implies the biased results due to the omission from the sample of a major segment of the population, i.e., the conservative, suspicious speakers of the dialect. Bradley suggests that the significant difference between the two groups for intelligibility testing is probably minimal since the more conservative people of a dialect seem to travel to other dialect areas just as often as the less conservative (1967.10). As noted earlier, even if there were a significant difference between these two groups, a more representative sample would not necessarily result from trying to apply a random selection technique. The subjects who would refuse to cooperate would probably be from the conservative segment of the community (see p. 110).

In regard to differences between sub-groups of subjects, there are a few suggestions from some of the studies. Bradley tested a sample of ten men against a sample of ten women in the Mixteca study. The women averaged 4% lower than the men on six reference point tapes. Even though none of the mean scores for women, taken a reference tape at a time, differed significantly from the corresponding mean scores for men, the tendency for women's scores to be lower than men's must not be overlooked. Since women are seldom tested there is a systematic bias in the intelligibility scores which is undoubtedly greater for some areas surveyed than for others. We need to make more comparisons between men's and women's test scores. (The women who are tested, however, constitute a biased sample in the same way as the men who are tested; they are the better educated and less conservative members of the group. They have also travelled to other dialect areas, if not more widely than that.)
There is no noticeable difference between the scores of monolingual subjects and bilingual subjects, except in the case of bilingual subjects who speak Spanish more than they do the minority language. Subjects as young as thirteen years have been tested, but their age and inexperience (though not that of fifteen- and sixteen-year-olds) are reflected in their test scores. Older men frequently have been difficult to teach how to take the test; therefore, few of the subjects tested have been more than fifty-five years of age. It also seems that travel to other dialect areas does not greatly influence a subject's test scores unless he has prolonged or repeated contact with another dialect. These suggestions are made partly to stimulate further research. The observations on which they are based are sketchy, and tests could easily be designed to clarify some of these issues.

(3) Sample size. The final point on sampling concerns the number of questions in the test and the number of subjects tested. The problem is whether there are enough items in the test and a sufficient number of subjects tested to yield a reliable measure of intelligibility. Pilot studies indicate that for the text test the sample size seems to be adequate and that an increase in either the number of test items or subjects would be uneconomical or infeasible for the gain in accuracy that would be attained. This conclusion may not hold for the sentence test, however.

Stoltzfus studied the effect of more questions on test scores for Chol (1966). He used fifteen questions on each test tape and then compared the mean score recorded for the fifteen questions with the mean score for the first ten questions. The average difference for twenty-five pairs of scores was 2.1%. Since the difference was slight, he concluded that ten questions were sufficient.

Although the Wilcoxon T test shows that there is no significant difference at levels beyond .05, the results are very tentative, partly because some test items were systematically excluded from the comparison. Since certain questions had been poorly translated, the investigators felt that subjects answered them incorrectly, not because the subjects failed to understand what the text said, but because they were misled by the questions. Unfortunately, the elimination of these questions reduced the variability of the test scores considerably.

On the other hand, Trique scores based on fifteen questions were significantly higher than those based on ten. The difference was significant for levels beyond .05 with the Wilcoxon test (T = 22, N = 15). The tendency for subjects to score higher on fifteen sentences than on ten was probably a direct result of making the test longer. The test was made long enough to allow subjects to begin adapting to the phonological differences between the Trique dialects. Thus subjects tended to make less incorrect responses toward the end of a test than they did at the beginning.

To test the significance of this tendency, I divided each test into three groups of sentences: Group I (sentences 1-5), Group II (sentences 6-10), and Group III (sentences 11-15). It seemed reasonable that subjects would make the most incorrect responses in Group I, somewhat fewer in Group II, and the least in Group III. Therefore, I tabulated by group the total number of errors that the 59 subjects tested had made on the test tapes. There were 236.5 errors in Group I, 146.5 in Group II, and 100 in Group III. I then determined the correlation between the relative number of errors and sequential grouping with the Kendall W statistic (Downie and Heath 1965.209-210). (The computation was based on the number of errors that the
entire sample of subjects made on each test tape.) The correlation was .521, which is highly significant at .01, since the critical value of $W$ is .30 when $m = 15$ and $N = 3$ (Downie and Heath 1965:314, Table XIII).6

To determine the particular cases in which subjects’ scores correlated significantly with the sequence of the groups of test items, I computed $W$ for each sample of subjects to which a test tape was administered. The computation was based on the relative number of errors a subject made in each group of five sentences. In accord with the above test with the Kendall $W$, I set the significance level at .01.

The results of the series of tests are presented in Table 16, along with the total errors per group of sentences that the sample of subjects from a particular reference point made on each test tape. In addition, sample mean scores of intelligibility are indicated.

Table 16 shows that $W$ is clearly not significant at .01 for any of the hometown tests and all but one of the non-hometown tests on which the mean scores were above 90% intelligibility. On the other hand, $W$ is clearly significant at .01 for five of the seven cases where the scores were below 90%. For three other cases, $W$ is significant at lower levels. Note that where $W$ is significant, subjects tended to score lowest on Group I, somewhat higher on Group II, and highest on Group III. Where $W$ is not significant, subjects tended to score equally well on all three groups. For two of the three marginal cases, they tended to score lowest on Group I and highest on Group II.

On the basis of the results of the significance tests, it seems that the Copala area scores on the Chicahuaxtla test tape reflect some factor to which subjects adapted during the course of the test so that they were understanding the sentences better during the last part of the test than they did at first. The Chicahuaxtla, Laguna, and Itunyoso scores on the Copala tape may reflect this factor to some extent. It is not reflected in the other scores.

I think that this disturbing factor consists of low-level phonological differences. The best way to estimate its quantitative effect would probably be to make repeated comparisons of sentence tests based on fifteen questions with text tests. This is because the wider context of the test test conditions subjects to such differences so that their test scores are not as greatly affected by them.

Finally, there are several ways to develop the sentence test into a more reliable measure. In the first place we could simply increase the number of test sentences from fifteen to twenty. If we do this, we might not want to score subjects on the first five sentences. On the other hand, since there may be no real advantage to making the test longer (at least, there is a practical limit on how long we can make it), we could make three copies of each test tape so that each group of five sentences came in first position on one of the tapes; i.e., the sequence of Groups I, II, and III on the first tape would be II, III, I on the second, and III, I, II on the third. By increasing the sample of subjects to twelve, each version of the tape could be administered to four subjects. A third possibility would be to elicit short texts from each reference point dialect. Each of these could be played to the subject just before he was tested on the corresponding sentence test.

---

6 To compute $W$ the sets of numbers are converted into sets of ranked values. In this case, fifteen sets of numbers and three ranks were involved.
<table>
<thead>
<tr>
<th>Reference Point</th>
<th>Test Tape</th>
<th>Questions Missed</th>
<th>W</th>
<th>S</th>
<th>P.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>Chicahuaxtla</td>
<td>Ch</td>
<td>3.5</td>
<td>0</td>
<td>1.5</td>
<td>.062</td>
</tr>
<tr>
<td>(10 subjects)</td>
<td>It</td>
<td>16.0</td>
<td>8</td>
<td>11.0</td>
<td>.250</td>
</tr>
<tr>
<td></td>
<td>Co</td>
<td>15.5</td>
<td>13</td>
<td>10.5</td>
<td>.053</td>
</tr>
<tr>
<td>Laguna</td>
<td>Ch</td>
<td>1.5</td>
<td>0</td>
<td>2.5</td>
<td>.013</td>
</tr>
<tr>
<td>(14 subjects)</td>
<td>It</td>
<td>17</td>
<td>5</td>
<td>14.5</td>
<td>.162</td>
</tr>
<tr>
<td></td>
<td>Co</td>
<td>23</td>
<td>9</td>
<td>2.5</td>
<td>.545</td>
</tr>
<tr>
<td>Itunyoso</td>
<td>It</td>
<td>1</td>
<td>0</td>
<td>1.0</td>
<td>.022</td>
</tr>
<tr>
<td>(10 subjects)</td>
<td>Ch</td>
<td>2</td>
<td>0</td>
<td>0.5</td>
<td>.053</td>
</tr>
<tr>
<td></td>
<td>Co</td>
<td>8</td>
<td>3.5</td>
<td>1.0</td>
<td>.260</td>
</tr>
<tr>
<td>Copala</td>
<td>Co</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>.053</td>
</tr>
<tr>
<td>(15 subjects)</td>
<td>It</td>
<td>37</td>
<td>28</td>
<td>15</td>
<td>.490</td>
</tr>
<tr>
<td></td>
<td>Ch</td>
<td>48</td>
<td>36</td>
<td>12.5</td>
<td>.748</td>
</tr>
<tr>
<td>San Miguel</td>
<td>Co</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>.008</td>
</tr>
<tr>
<td>(10 subjects)</td>
<td>It</td>
<td>24</td>
<td>26</td>
<td>13</td>
<td>.646</td>
</tr>
</tbody>
</table>

Table 16. Relative Consistency of Subjects' Errors: Trique Sentence Tests

<table>
<thead>
<tr>
<th></th>
<th>Sentences 1-5</th>
<th>Sentences 6-10</th>
<th>Sentences 11-15</th>
<th>Kendall W correlation coefficient</th>
<th>Significance at .01</th>
<th>Percentage of Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch</td>
<td>Chicahuaxtla</td>
<td>I</td>
<td>II</td>
<td>III</td>
<td>W</td>
<td>S</td>
</tr>
<tr>
<td>Co</td>
<td>Copala</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>W</td>
<td>S</td>
</tr>
<tr>
<td>It</td>
<td>Itunyoso</td>
<td>a. significant at .10, W = .25</td>
<td>b. significant at .20, W = .16</td>
<td>c. significant at .10, W = .25</td>
<td>W</td>
<td>S</td>
</tr>
</tbody>
</table>

(Downie and Heath 1965.314, Table XIII)

Regarding the sample of subjects, Kirk compared a sample of ten against a sample of twenty in the Soyaltepec-Nuevo Soyaltepec area and found no significant difference between the sample means recorded for the standard Huautla test. (The difference in sample means for the two samples was only 1%.) The populations of Soyaltepec and Nuevo Soyaltepec are assumed to have the same composition since Nuevo Soyaltepec was formed from Soyaltepec about fifteen years ago when a river basin was dammed up and most of Soyaltepec was inundated.

As for smaller samples of subjects, a comparison of the mean scores of samples consisting of the first three subjects and the first six subjects with
mean scores of complete samples of ten subjects for the twenty-three towns in the Mazatec area is summarized in Table 17. The purpose is to show the approximate differences from a mean score based on a sample of ten that could be expected if it were necessary to use samples of fewer than ten subjects. The table shows these differences for hometown test tapes and reference tapes, falling into five ranges of intelligibility: 0-39.9%, 40-59.9%, 60-79.9%, 80-89.9%, and 90-99.9%. This breakdown was made partly for ease in computation, and partly because it was suspected that variability would be relatively small at high levels of intelligibility and might increase

<table>
<thead>
<tr>
<th>Intelligibility Ranges</th>
<th>0-39.9</th>
<th>40-59.9</th>
<th>60-79.9</th>
<th>80-89.9</th>
<th>90-99.9</th>
<th>Hometown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean diff A (3&amp;10)</td>
<td>6.7</td>
<td>7.3</td>
<td>8.8</td>
<td>6.1</td>
<td>2.5</td>
<td>2.9</td>
</tr>
<tr>
<td>Mean diff B (6&amp;10)</td>
<td>3.7</td>
<td>5.4</td>
<td>3.7</td>
<td>4.3</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Standard dev A</td>
<td>3.71</td>
<td>4.91</td>
<td>7.95</td>
<td>6.11</td>
<td>1.48</td>
<td>2.90</td>
</tr>
<tr>
<td>Standard dev B</td>
<td>3.61</td>
<td>2.79</td>
<td>2.44</td>
<td>4.01</td>
<td>1.48</td>
<td>1.00</td>
</tr>
<tr>
<td>Expected diff A</td>
<td>1.70</td>
<td>3.30</td>
<td>3.63</td>
<td>2.58</td>
<td>.95</td>
<td>1.24</td>
</tr>
<tr>
<td>Expected diff B</td>
<td>1.65</td>
<td>1.87</td>
<td>1.11</td>
<td>1.69</td>
<td>.95</td>
<td>.44</td>
</tr>
<tr>
<td>Largest diff A</td>
<td>14.8</td>
<td>15.3</td>
<td>19.7</td>
<td>25.3</td>
<td>4.7</td>
<td>10.0</td>
</tr>
<tr>
<td>Largest diff B</td>
<td>14.7</td>
<td>10.0</td>
<td>9.2</td>
<td>17.7</td>
<td>4.7</td>
<td>4.0</td>
</tr>
<tr>
<td>N</td>
<td>21</td>
<td>11</td>
<td>21</td>
<td>24</td>
<td>12</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 17. Samples of Three and Six Compared with Samples of Ten

significantly at lower levels. Included in the table are the mean differences between scores based on samples of ten and scores based on samples of three and six, the standard deviations of the corresponding groups of differences and the expected variation of the sample means of these groups, computed from the student T formula for a 95% degree of confidence. In addition, the largest difference recorded for both groups in each of the various ranges of intelligibility and the number of differences are indicated. The basis of the following treatment is the fact that the differences for samples of three and six subjects are normally distributed around the mean of samples of ten subjects.

In the table, the letters A and B refer to the sample of three and six subjects, respectively. Mean diff A (3&10), 60-79.9 = 8.8 reads that, on the average, if a mean score based on a sample of ten for a particular test tape fell within the intelligibility range of 60-79.9%, then the score recorded for a sample of three on that test tape was 8.8% higher or lower than the mean score based on the larger sample. Standard dev A, 60-79.9 = 7.95 reads that the standard deviation for the 21 differences for samples of three in the 60-79.9% range is 7.95%.

The standard deviation is a measure of the variation that is found within a group of measurements. For example, the average of the series of numbers 3, 3, 3, 3, 3, 3, is 3, while the average of the series 1, 2, 3, 3, 4,
5, is also 3, but the latter series of numbers varies more than the former. Thus a measure of variation is needed to reveal an important characteristic of the two sets of numbers that the arithmetic mean does not express. The degree of variation within a sample of measurements is directly related to the accuracy of the prediction that can be made about the population from which the sample was drawn. From two randomly selected samples of the same size, the more accurate prediction can be made from the sample which gives the smaller standard deviation. The standard deviation, then, is used to compute a predicted range in variation for the mean scores of measures for a sample of a given size at a particular level of probability. Thus, expected diff A, 60-79.9 = +3.63 reads that if additional samples of 21 differences between mean scores for samples of three and ten occurring in the range of 60-799% intelligibility were to be drawn, then 95 times out of 100 the mean score for groups of 21 samples of three would be within 3.63% above or below the mean difference of 8.8% that was registered for the initial samples.

It is necessary to point out that the range estimate discussed above applies only to the mean of an entire sample of differences within one range of intelligibility. In making an estimate of the expected differences between mean scores for samples of three and ten, as in the above table, the prediction should be made on the level of an entire study or, at least, large segments of a given study. Sample means from smaller samples may vary greatly from the mean that would be obtained by using samples of ten or more. The observed extreme differences in sample means for both groups are indicated in the table. Thus, largest diff A, 80-89.9 = 25.3 means that there is one case in which the mean for a sample of ten on a given test tape was in the intelligibility range of 80-89.9% and was 25.3% higher than the mean score of the first three subjects who were tested on that tape.

Note that the largest difference in A does not necessarily correspond to the largest difference in B. It merely means that the widest deviation from the mean for a sample of ten happened to be a certain value for each of the smaller samples at the various categories of intelligibility. N represents the number of cases in each of the groups.

In addition, a weighted mean difference from the mean of samples of ten computed for samples of three and six shows that the mean difference for the entire set of samples of size three is 5.92 and the mean difference for all samples of size six is 2.79. This accords with the general trend toward a true population value as the sample size is increased.

A weighted mean standard deviation computed for the entire sample of 112 differences is 5.5 for differences between the mean scores of samples of three and ten, and is 3.6 for differences between six and ten. Computation of a confidence interval shows that for a sample of 112 differences, the mean difference will be in the range of 5.92 + 1.02% 95 times out of 100 for scores based on samples of three subjects, and it will be in the range 2.79 + .67% 95 times out of 100 for scores based on samples of size six.

Finally, the means for samples of three are approximately normally distributed around the means of samples of ten, and this approximation to a normal distribution becomes quite obvious when samples of six are used. For a discussion of the use of the standard deviation and predictions from this data, see Appendix H.
7 ALTERNATIVE APPROACHES

7.0. In this chapter we consider briefly what to do when it is not feasible to test intelligibility. Since earlier chapters either hint at or treat in detail most of the main points that relate to this topic, the discussion here, to some extent, recapitulates these main points. In short, if it is not feasible to test for intelligibility, then the survey team should attempt to consistently collect linguistic information, certain kinds of sociological data, and informants’ opinions about the relative similarity of other dialects to their own. These data can be used to form ideas about the dialect groupings within which communication takes place in the vernacular. Some of the proposals about how to collect these data and what to do with them after they are collected are highly tentative. They are meant only to suggest some actions that can be taken. They are not complete in any sense of the term.

7.1. Linguistic Relationships. McKaughan has classified the approaches to determining language and dialect relationships into comparative, lexicostatistic, and phonostatistic methods. He considers that each method is most useful within certain ranges of linguistic divergence, ranked as follows: phonostatistic methods are most applicable to languages that are slightly divergent; lexicostatistic methods are still useful for relating moderately divergent languages; and comparative methods, particularly structural studies, are most useful for classifying widely divergent languages (McKaughan 1964.118). Because the ranges overlapped, however, he used the results of one method to validate results obtained by another (1964.99).1

Since we normally expect intelligibility to exist only among highly similar dialects, phonostatistical measures may be more helpful to us than

1. In another paper in the same volume, Wurm makes some comparisons of these methods in their application to New Guinea Highlands languages (Wurm 1964). In this volume also, see Eunice V. Pike (1964). Her typological outline suggests several plausible features to correlate with low intelligibility.
lexicostatistic measures or comparative statements (Ladefoged 1968.5). It would be unwise, however, to ignore groupings outlined by studies in which these other methods were employed. Such groupings provide the broad framework within which we can look for smaller groupings of "mutually" intelligible dialects. Furthermore, we would not restrict the scope of the linguistic data that we collect to only the bare minimum that would enable us to compute some measure of phonological divergence. Close structural comparisons are necessary to determine the kinds of linguistic differences that hamper intelligibility and those that do not.

Grimes and Agard proposed a phonostatistical measure that they applied to the Romance languages (1959). The method, based on comparative phonology, quantified the phonological differences between the sounds in 169 sets of correspondences for all pairings of seven related languages.

To begin with, they made allophonic transcriptions of the sounds in each language and determined for each pair of languages whether these sounds were the same or different in the positions relevant to the sets of correspondences. They also assigned a numerical value to each of these differences. The measure of phonological divergence between each pair of languages was derived by averaging the values of the differences in sound over all the sets of correspondences.

Differences between a pair of sounds were specified in terms of six dimensions of articulation (based on Pike's notion "rank of stricture," 1943): (1) point of articulation, (2) degree of constriction of the air stream at the center line of the mouth, (3) effective timing of the central constriction, (4) secondary oral shaping, (5) velic action, and (6) laryngeal action (Grimes and Agard 1959.601-2). Each dimension was then scaled in terms of the number of phonetic distinctions that occurred in the allophones of the seven languages, and integers were assigned to each distinction recognized along a dimension. For example, the dimension "point of articulation" was scaled as follows: (1) bilabial, (2) labiodental, (3) interdental, (4) apical, (5) laminal, (6) dorsal, and (7) glottal. Apical, laminal, and dorsal correspond to the vowel positions front, central, and back. The value of the differences between a pair of sounds was calculated by finding the numerical difference between the phonetic features of the sounds for each dimension and then summing all differences. Some sound features were irrelevant to the computation. They were therefore omitted from the calculations.

Grimes and Agard showed that the groupings implied by the measure of phonological divergence agreed with those arrived at by standard comparative techniques. In addition, they suggested ways to refine the approach further and the kinds of problems to which the method might give significant answers.

In a subsequent paper, Grimes reported on a refinement in the measure and on additional applications of it (1964). For each pair of languages he plotted on a graph the number of correspondence sets that occurred at each degree of difference from 0 to n. A regression line was then fitted to the points that represented the number of sets of phonological correspondences. The slope of the line was taken to be the index of dialect distance. (The steeper the slope, the closer the two languages were related [1964.45].) Grimes also computed the correlation between the number of sets of correspondences and the degree of difference. Since all the points on the graph did not lie on the regression line (rather they were scattered along both sides of it), he wanted to know the range of random scatter at
different degrees of divergence. He found that the scatter measurement was independent of the slope. From this he concluded that languages tend to change regularly, regardless of the degree of divergence (1964.46).

With regard to the Romance groupings outlined by the use of regression slopes, Grimes pointed out that both groupings were substantially the same except that a cleavage between the western group of languages and the others was more apparent with the index based on regression slopes (1964.47). Finally, he felt that the measure of phonological divergence was a more refined measure of linguistic relationships than glottochronology (1964.48).

McKaughan applied the Grimes-Agard measure to some New Guinea Highlands languages. He used only four scaled dimensions to measure the difference between them: point of articulation, degree of constriction of the air stream, velic action, and laryngeal action (McKaughan 1964.102). He agreed with Grimes that the phonostatistical measure was more refined than the lexicostatistical one. However, he departed somewhat from Grimes on another point. Whereas Grimes had used only material that was definitely shown to be cognate by the comparative method, McKaughan felt that the measure of divergence could just as well be determined from a list of words that included some of doubtful status, that the method could be applied usefully a long time before results from the comparative method became available (1964.103).

Peter Ladefoged recently proposed a second phonostatistical measure and applied it to languages of Uganda (1968). Since he sought to determine cross-language communication (intelligibility), he set up 100-word lists that were rather different from those used in other kinds of comparisons. The words were chosen so as to represent equally several different aspects of the culture common to the groups being studied. In addition, no concern was paid to whether words were borrowed from another language. Thus Ladefoged did not even attempt to elicit some kind of basic vocabulary, used in glottochronology because it is supposed to be resistant to change through contact. Finally, he attempted to elicit generic terms, rather than specific ones, in order to avoid having to use words that might yield several equivalent forms.

He also proposed two ways of finding cognates in order to estimate the effect of semantic shifts on intelligibility. The common equivalents of each word would be elicited first. Then the informant would be given the word used in neighboring languages and be asked to give, if he could, a word in his own language that sounded nearly the same and had a similar meaning. The measure of linguistic similarity would be computed for both lists. Ladefoged felt that the two methods of eliciting cognates would probably lead to different results. The results of either method might correlate with other methods of assessing cross-language communication such as asking informants' opinions or testing comprehension.

To begin with, Ladefoged grouped all the Ugandan languages into four families. For each pair of languages in each family the degree of similarity was computed by quantifying the differences between the corresponding sounds in each pair of words over the entire word list.

To make sure that the comparison would be made between the corresponding sounds of a pair of words, he normalized all the words in the lists. In the first place, he assigned ordinal numbers to the sequence of segments in each word in such a way that the first consonant of the word
stems always occurred in the same position. In addition, each word was transcribed as an alternating sequence of vowels and consonants. He indicated double vowels and double consonants with a superimposed mark for length; he treated compound consonants as single units having associated features such as palatalization or labialization. Finally, he felt that the perceptual difference between a long form and a short form in a pair of cognates would be slight. Therefore he added dummy segments to the shorter members of cognate pairs in order to equalize the number of segments in both members.

To quantify phonological features, Ladefoged fully specified each segment of each word in terms of a set of binary distinctive features, selected specifically for describing Ugandan languages. (He used binary features to simplify computer processing and to avoid certain implications of the use of tertiary and quaternary features [1968.5].) His tentative set of features for consonantal segments were the following: (1) stop, (2) nasal, (3) fricative, (4) approximant, (5) voiced, (6) frontal, (7) coronal, (8) long, (9) labialized, and (10) palatalized. The tentative features for vocalic segments were these: (1) high, (2) mid, (3) low, (4) front, (5) central, (6) back, (7) long, (8) y-glide, (9) high tone, and (10) low tone. He set up this feature system to include certain redundancies. This allowed him to characterize the segments in a way that seemed to correspond with his intuitive judgment about the degree of phonetic difference between them.

The degree of difference between a pair of sounds was calculated as the number of features by which they differed. For example, [p] differed from [pw] in only one feature, labialization, so that the numerical difference was only one; but the difference between [pp] and [pw] was two since they differed in both length and labialization (Ladefoged 1968.6, Table 3). The difference between a pair of words thus consisted of the sum of the differences between each corresponding pair of sounds in them.

Ladefoged's approach to linguistic similarity seems to lead naturally to a method of quantifying linguistic differences on other levels. In particular, it would be feasible to elicit a test list of thirty or more sentences about culturally relevant events, say, five declarative, five intransitive declarative, five transitive imperative, five intransitive imperative, five interrogative, and five equational sentences. Some of these might be further subclassified for number of subject or actor, number of goal, tense, and aspect. The test list would then be elicited from informants of several different dialect areas.

The next step would be to compare the lists in terms of binary phonological, syntactic, and lexical features. These would include some of the following: segmental differences, supra-segmental differences, changes in word order, affix substitutions, stem or root changes, morpheme class differences, and lexical changes.

Presumably some kind of rank order could be assigned to these features in terms of their probable effect on intelligibility. For example, phonetic differences between individual segments are probably the least important kind of differences. Supra-segmental differences (tone or stress) are

2. The consistent difference between long forms and short forms might well correlate with non-reciprocal intelligibility, however. The Totonic results suggest that this is the case.
3. Approximants are sounds produced so that the air stream is unobstructed by the slight constriction of two articulators (Ladefoged 1967.28-29).
4. Coronal sounds are produced by raising the tongue blade from the neutral position (Chomsky and Halle 1968.304).
probably more important than changes of word order. Stem or root changes might be more important than changes in the prefixes (Ladefoged 1968.5). Prefix changes (in terms of the items within a class of prefixes), on the other hand, might be more important than stress shifts, but less important than tone changes. Morpheme class differences may be more important than lexical changes.

The rank order values could serve as weighting factors in quantifying the differences between the lists. For example, one could add up the total number of differences in each feature class, multiply each sum by the appropriate class rank value and sum the products. One could also devise a standard scoring chart of the kind that Pierce used in his study of Algonquian (see Section 3.1.3). Besides using binary features for grammatical and lexical differences, one might simplify further by choosing to exclude certain classes of items. Finally, the index could be stated as the percentage of difference between a pair of lists. The important thing is that the data be handled consistently.

The ranking of features must be validated by correlating detailed linguistic descriptions with intelligibility. This could help us discover combinations of factors that we can use to predict accurately the presence of intelligibility between pairs of dialects.

7.2. Sociological Data. These data are used to look for patterns of dependencies and interaction among dialect areas. These patterns may correspond to intelligibility patterns. If we know the areas to which people travel from other dialect areas, the percentage and subgroups of the dialect population that go to these areas, why they go and how frequently, and what languages they speak in these other areas, we can probably form reasonably correct impressions about which dialects are the most widely understood.

In considering two dialects, A and B, it usually turns out that more people go to area A from area B than go to B from A. The survey team might try to answer certain questions about this pattern. For example: Who travels to B? (All the men? Merchants only? Both men and women?) How many people go there? (Most of them? Only a few?) Why do they go there? (To work? To attend fiestas? To visit relatives? To take care of civil affairs or responsibilities?) What language do they use most often at B? (The national language? A trade language? The vernacular?)

To sum up, the survey team would do well to record consistently for each dialect area all the information that they can learn about the kind, degree, and purpose of social interactions between that dialect and neighboring ones (see Section 2.1.2 and Appendix J).

7.3. Informant Opinion. Kirk’s approach to sampling informant opinions may be a very useful one. However, only one or two informants were questioned at any test point. They were asked to evaluate other dialects in terms of a two-point scale: easy to understand or difficult (see pp. 83-86). The method could be improved by merely obtaining opinions from a larger number of informants at each test point. They should be questioned independently of one another.

An even better procedure would have informants rank dialects along a multi-valued dimension such as the degree of similarity to their own dialect. For example, they might be asked to scale the other dialects as identical, the same, fairly similar, very different, another language. A second
dimension would be the degree of intelligibility with other dialects. Ladefoged proposed a five-point scale based on the question, "How much do people around here understand of dialect A?" He expected to scale the answers: all of it, most of it, the main points, a little, or none (1968.1). I would include a third dimension, how much the other dialect changes. The range of answers might be: not at all, some of the words, a lot of the words, most of the words, and completely.

It would be nice if the three dimensions outlined above turn out to be the same in most cases. I suggest using all three simultaneously because of certain problems that will be encountered as we try to apply the method.

In the first place, the three dimensions may not reveal one underlying scale. If they do not, we may be able to use one of the dimensions to measure attitudes. In any event, we will probably have a more reliable and a more valid measure if we use all three dimensions in the same questionnaire. Ladefoged points out two problems in getting reliable and valid results (1968.1-2). An informant's ability to understand a given dialect may be very different from the general ability of the population viewed as a whole. Because of this, it is wise to collect opinions from informants who represent a cross-section of the population. Furthermore, informants might claim that they do not understand another dialect when they actually do.

One of the most serious problems is that the investigator will have to categorize the degree of estimated intelligibility from the informants' responses. Most of the informants we work with are illiterate. Were they literate, we could make up forms on which each outside dialect would be paired with the informant's, but separated from it by a series of blanks. The informant could be asked to fill in the blank corresponding closest to his own feelings about how similar each outside dialect is to his own. Since this cannot usually be done, however, something is needed to guarantee that investigators will be consistent when they evaluate responses. It is obvious that part of the inconsistency will be in the way subjects describe dialect differences. The categories of the three dimensions include most of the ways I have heard Mexican Indians describe these differences.

If the assumption that all three dimensions are based on the same variable is correct, an investigator can begin by asking the informant three questions about each pair of dialects: (1) How similar are they? (2) How much do they change? and (3) How much do the people understand of the other dialect? He could then plausibly locate the subject's opinion as the point on the scale where two or more of the subject's responses occurred. For example, if the subject answered "the same" to the first question, "a lot of the words" to the second, and "most of it" to the third, the investigator could evaluate the responses as meaning point two on the scale of five. If they were all different, however, he might use the mean of the corresponding rank numbers to represent the subject's opinion.

The team can expect to receive completely anomalous opinions occasionally. For example, the first four subjects might say that they do not understand another dialect at all, whereas the fifth might claim to understand it perfectly. The anomalous response may not be incorrect. The fifth informant may have friends in the other dialect area; the four subjects may be voicing a widespread group resentment against speakers of the other dialect. It may turn out that word lists from both dialects are identical, even phonetically. To reject the one opinion because it is anomalous might be unwise. Under certain conditions, group feelings might
change so that people begin admitting that they understand the other dialect after all.\textsuperscript{5}

Before obtaining opinion data from an informant, the investigator should try to find out which dialects the informant has had personal contact with, and which he has not. Then the informant can be asked about only those dialects that he has actually listened to. In this way it might be possible to obtain fairly reliable data. Note that it is not necessary to look for informants who have firsthand knowledge of \textit{all} the reference dialects of interest at a given test point.

Finally, the rank numbers, representing informant opinions, could easily be averaged and entered into matrices in the same way as raw intelligibility scores. Groupings of dialects could then be derived at thresholds corresponding to the individual ranks.

To summarize this chapter: data about linguistic structures, social interactions, and informant opinions should be collected and analyzed in order to form impressions about dialect groupings within which cross-language communication takes place. Rough measures such as cognate counts provide a first step in grouping the dialects; we do not expect to find intelligibility between two slightly related dialects of a vernacular language. Phonostatistic and structural statistic measures may provide a further approximation to the kind of groupings we want. Data on social interactions could help us find out who actually communicates with whom, and how often, within the tentative groupings. An analysis of informant opinions could substantiate both the tentative groupings and communication patterns. In addition, we might be able to pinpoint negative feelings between groups.\textsuperscript{6}

\textsuperscript{5} Mildred Larson informs me that this actually did happen between the Aguarunas and Huambisas of Peru (personal communication).
\textsuperscript{6} I think that the results from these various approaches can be incorporated into a Guttman Scale analysis. At present, however, I am not competent to go into this approach. For the interested reader I include in the bibliography references to this technique and its application to some problems in anthropology.
8 BILINGUALISM

8.0. In addition to describing the communication patterns between dialect groups of minority languages, we are confronted with exploring relationships between the minority language and the national language. Thus a dialect survey entails the collecting of data on bilingualism. These data play an important role in assigning relative priorities to the dialect areas that are defined. In certain cases it will provide the basis for deciding whether or not it is necessary to conduct a program in the minority language at all.

8.1. Definitions and Goals. Any person who can speak more than one dialect of the same language, or can speak two or more distinct languages, is bilingual. Likewise, any community in which there are two or more subgroups each of which speaks a separate dialect or language is considered a bilingual (or multilingual) community. (A rather different bilingual situation is one in which two variants of the same language are spoken by the entire population in social situations appropriate for the use of one or the other variant. This is referred to by Ferguson as diglossia [1964].) In the following discussion bilingualism will be considered as both an individual and a societal phenomenon.1 In both cases it is modeled as a continuous variable rather than as an absolute characteristic (Diebold 1964). Finally, while the concepts involved apply to studies of bilingualism between dialects of a minority language, the emphasis here is on bilingualism between a given minority language and a national language such as Spanish in Mexico.

To describe bilingualism adequately within a given speech community, it is necessary to establish both a theoretical and a procedural basis for the studies. In the course of the dialect surveys in Mexico, investigators have not coordinated their approach to eliciting data on bilingualism. A standard ethnographic questionnaire which seeks some data on language use is available (Appendix E). Both Paul Mellema and Stoltzfus have used a test

for comprehension of Spanish on a limited scale. Various kinds of rating scales have been applied to judge the degree of bilingualism of subjects, but numerous uncontrollable factors lead to highly unreliable results. We still do not know the type of data that are really needed or the procedures for obtaining these data. This is because there is no well-developed model available with which to approach the problem of bilingualism within the framework of the present methodology.

Some questions are particularly relevant: (1) In a given bilingual community, which language is dominant in the various sectors, the local vernacular, or the national language? (2) If the vernacular is dominant, what proportion of the population speaks the national language either as the first language or as an acquired tongue, and what are the uses of the national language for the community? (3) If the national language is dominant, then what proportion of the population speaks the vernacular and in what situations? (4) Regardless of the relative dominance of either language, is the relationship between the uses of the two stable or changing? (5) If the bilingual condition within the community appears stable, what are possible signs of weakness which threaten the continuance of either language? (6) If the condition is changing, what factors point to the rate and direction of subsequent change in the pattern of language use? (7) If the use of the idiom is diminishing, will a significant portion of the population continue to speak it for a significant period of time?

8.2. Fishman's Approach to Bilingualism. It is immediately obvious that there are no straightforward answers to the above questions. The most probable answer to some of them will be a function of interacting psychological, sociocultural, historical, and geographical factors. An adequate model of bilingualism will thus be largely based on extralinguistic factors (Weinreich 1953.3, 83-110; Diebold 1964.495; Shankara Bhat 1968.20).

In an attempt to organize the diverse approaches to the study of bilingualism, I am following the sociological framework suggested by Joshua A. Fishman in two papers published in Linguistics (1964, 1968).

8.2.1. Criticism of traditional approaches. Studies of bilingualism have been carried out by psychologists, linguists and sociologists (Fishman 1968.21). Fishman claims that studies in all three areas have generally failed to provide a model usable for interdisciplinary studies. These studies in general have touched on bilingualism as a complicating subsidiary aspect of some other problem and have not considered bilingualism per se.

Psychology, for example, has characteristically used a model that views bilingualism as a global capacity, even though at the same time it has increasingly analyzed capacities such as intelligence, memory, and perception as consisting of several different factors. The degree of bilingualism has generally been measured by speed in translation or word association techniques. Fishman regards speed as at best a "peculiarly ethnocentric" criterion. He also points out that such measures are considered to be "context free" in the sense that they are not related to any outside factors that would influence a subject's performance during an experiment (1968.25). Finally, the psychological model views language proficiency in terms of "balance" and "dominance," which Fishman considers unrealistic from a functional viewpoint in that a society in which

2. Word association techniques may be useful, however.
most of the members use both languages for all functions will probably lose one of the languages because it is unnecessary to have two languages which serve the same ends equally.\(^3\)

The linguistic model of bilingualism was constructed on the fiction of two homogeneous speech communities in contact, with the resultant changes in both languages considered to be "interference" between them. This view of language as a pure monolithic structure is at least oversimplified (1968.27)\(^4\). There were two results of this view that especially limited the usefulness of the model for the sociologist. First, bilingualism was viewed as revealing interaction between two homogeneous social groups rather than as reflecting interaction among well-defined subgroups of both communities. For example, social classes and age-groupings may be important parameters for describing bilingualism in both speech communities (Shankara Bhat 1968.18). The second byproduct of the linguist's conception of bilingualism was that the convergence of the two languages led to distasteful changes in both languages.\(^5\) A more realistic position is that an investigation of the types, extent, and purpose of borrowings between languages might produce significant insights into bilingualism (cf. Weinreich 1953.59, 62). Finally, Fishman notes that linguists can effectively contribute to the field by describing the varieties of a language that are spoken by the subgroups within bilingual communities.

Fishman's critique also includes previous sociological studies of bilingualism. The first point he makes is that little attention has been directed toward the relationship between bilingualism as self-reported by individuals and bilingualism as observed or measured by testing. Therefore, inferences from discrepancies between self-reported data and more objective data have not been possible. He adds that the extensive use of census data is appropriate, but that there is generally much doubt as to the validity of such data. (Lieberson mentions four grounds of error in census data: intentional distortions introduced by the census agency, lack of comparability between questions used in censuses by various countries, inconsistency between countries or between successive censuses by one country in phrasing a particular question, and unintentional errors incorporated into the results because a respondent does not know how to answer. He hopes, however, that future censuses may be considerably improved [1967b.139-40].) Finally, Fishman criticizes many sociological studies for depending too much on gross social categories in their analysis of bilingualism. Specifically, do the relevant parameters of a particular social group correspond to those which were actually used in studying the bilingualism of that group?

8.2.2. A static model of bilingualism. Fishman's approach is through a functional analysis of hierarchically ordered clusters of units, analogous to the structuralist model in linguistics. The unit clusters are analyzed as occurring in one of four levels: roles, situations, domains, and culture values (Fishman 1968.46). Roles involve specific people in particular relationships such as husband and wife or teacher and pupil. Included in such

---

3. The concept of "dominance," however, seems to be quite useful even within the functional approach; cf. the discussion on "dominance configurations," pp. 131.
4. Note that Martinet claimed that this view even had its limitations for linguistic studies (preface to Weinreich 1953.vii).
5. This convergence probably ruffles the esthetic feathers of many linguists more than it leads them to consider such phenomena deleterious.
relationships are the relative status of the participants and the purpose of the relationship. The kinds of role relationships are limited by networks which may be open or closed. The network structure may determine when it is or is not appropriate to use a particular language. Situations consist of participants enacting a role relationship in an appropriate time and location, such as town authorities settling a civil dispute in the town hall on a Tuesday afternoon. Domains are the institutions and norms established by a society within which particular situations are realized, such as school. The culture values are assumed to be revealed by attitudes and modes of behavior within the domains, such as a movement by the elite of a minority group to have their language taught on the university level. In this model, it is assumed that a bilingual community remains bilingual only as long as each of the languages is associated with complementary sets of culture values. Overlapping of the sets may lead to instability.

8.2.3. A model of language maintenance and language shift. The above model views bilingualism primarily as a static characteristic of a population. It also describes a population in terms of its internal structure and dynamics, largely separate from external relationships and processes. In an earlier paper, Fishman outlined a more inclusive model of bilingualism as an approach to the study of language maintenance and language shift (1964). He considered the three main areas for research to be (1) synchronic and diachronic studies of language use by bilingual groups in contact, (2) the relationships of psychological, social, and cultural processes to the use of the languages by these groups, and (3) group behavior toward the various languages in the contact setting.

(1) Language use in bilingual groups. Research in the first area would lead to a description of the relative positions of two or more languages in a multilingual community. The description is made in terms of a “dominance configuration” which is based on data about the degree of bilingualism and the domains of language use. The rate and direction of language shift for a population could be ascertained by comparing two similar studies made for that population at two points in time.

Degree of bilingualism would be determined by measures covering three levels at which individuals may differ greatly in language use. The first is media variance, i.e., one's use of language for reading, writing, and speaking. The second parameter concerns the roles in which the language is used, i.e., inner speech, comprehension, and production. In addition, individual language use might be measured for particular types of situations, classified by Fishman as formal, semi-formal, informal, and intimate.

Domains of language use, defined earlier as culturally established institutions and norms, might follow some revision and refinement of the Schmidt-Rohr classification which covers the following nine domains: family; playground and street; school with its subdomains of instruction, languages taught, and recess; the church; literature; the press; the military; the courts; and the government bureaucracy (Fishman 1964.37)6.

(2) Extralinguistic factors. The second question for research seeks to specify the geographical, sociocultural, and psychological factors that influence language use. It may be necessary to consider these factors over

---

6. Note Fishman's caution, however: "No invariant set of domains can prove to be maximally revealing." (p. 47)
time. Though numerous generalizations have been made concerning the extralinguistic parameters which are crucial to language maintenance and language shift, no one of them is invariably associated with either condition. Particular combinations of several factors seem to underlie individual cases of stable or shifting bilingualism.

Fishman illustrates this by citing counterexamples and restrictions to three commonly mentioned generalizations: (a) language maintenance is a function of group loyalty or nationalism; (b) language shift occurs most readily in urban areas, whereas rural areas are most impervious to such change; and (c) the language of lower prestige is replaced by the one of higher prestige. To the first one he points out, for example, that the Guayaquyans of Venezuela surrendered both their language and religion but preserved their group solidarity by maintaining their property relations (1964.51). While admitting that the urban environment tends to change more rapidly than the rural one, he comments that the direction of change is generally unpredictable and may result in a conscious group effort directed toward preserving the traditional language. Regarding prestige languages, he notes that a low German dialect displaced Lithuanian in East Prussia before World War I even though standard German was widely spoken (1964.54). Weinreich demonstrates a similar case between a low Swiss German dialect and Romansch in the face of the standardized German dialect (1953.84-86). Fishman also cites Kloss’s four “ambivalent features,” none of which by itself is connected with either language maintenance or shift. These four features are the absence or presence of higher education in the mother tongue, the relative size of the bilingual group, the degree of similarity between the groups in contact, and the attitudes of the majority group toward the minority (1964.50).

He suggests a three-fold approach to structure the diverse extralinguistic factors that affect multilingual communities. In the first place, studies should be carried out to explicate individual and group involvement in gross sociocultural processes. It is necessary to study how individual and group behavior is affected by processes such as urbanization, industrialization, nationalism, nativism, secularization, etc. The second step would require establishing a typology of contact situations. For example the case of the immigrants to the United States in the nineteenth century is felt to be quite different from that of the Spanish conquest of Mexico. Furthermore, in Mexico after the conquest acculturation assumed distinct forms for the Yaqui, Tarascan, and Mixe groups (Beals 1967.466-467). To set up a typology of these situations, it is necessary to systematize the factors that underlie the intuitive differences between them. Some of these factors would include the power configurations between the dominant and minority groups, type of control exercised by the majority, characteristics of the initial contact between the two groups, plurality patterns, and stratification of the minority group. Assuming that a suitable typology of contact situations were available, and that a relatively uniform and adequate approach had been developed for describing the sociocultural processes active in those situations, the final step would be to compare these language groups and contact situations.

(3) Group behavior toward the languages in contact. The final area of research concerns group behavior which is specifically directed toward the maintenance or shift of a language. Some kinds of behavior considered by
Fishman are attitudes toward the languages and overt implementation of maintenance or shift by group programs.

Regarding attitudes, he suggests studies to describe not only loyalty and antipathy toward a particular language, but also studies relating value judgments such as "beautiful," "ugly," "harsh," or "musical" to the features of given languages. To illustrate the necessity of studies relating attitudes to use, Fishman cites perfect negative correlations between the two. For example, older Polish immigrants to Australia identified strongly with English after a few years' residence there even though they could barely understand English. On the other hand, many of the younger immigrants identified strongly with Polish even though they could not speak it. He reports similar observations for American immigrants (Fishman 1964.60-61).

Fishman points out that overt implementation may have official or unofficial sanction and that most of such efforts occur in the context of language maintenance or shift (1964.61-62). He adds that the relationship between the two conditions and language planning is an unstudied one. Furthermore, studies about the American immigrants have revealed that unexpected relationships may develop because of unusual reinforcement efforts by a community to preserve its language.

8.3. Data Collection. Several modifications must be introduced into Fishman's model in order to utilize it along with field tests. These result in both a simplification of his model and the specification of locally pertinent parameters.

8.3.1. Degree of bilingualism. While the degree of bilingualism within a bilingual community may be evaluated by measures of language proficiency and the description of domains of language use, the framework that we use is much simpler. Because the real world is complex and we actually do not know much about how to measure bilingualism or predict its trends, we will begin with a model that incorporates some unrealistic assumptions. As we attempt to apply it, we can hope to learn enough to make our model more complex and realistic by adding new variables to the scheme or by postulating more complex relationships among the variables included in the present version (Blalock 1969.3-4). Because long-term change studies require that a community be studied at two points in time, say five years apart (and there are other complications), the preliminary approach for describing various stages of change will be to compare several bilingual communities at the same point in time. Additional information can be sought at each of these points which might indicate the direction of change, if any. Questions which might be relevant include the age distribution of those who still speak the minority language, including the percentage of school children who are monolingual in the minority language when they enter the first year of school. Since the directions of sociocultural processes result from the behavior of individuals, data will be collected on both the individual and societal levels (Weinreich 1953.71).

(1) Language proficiency. The approach to describing the degree of bilingualism will involve a single measure of proficiency via comprehension of oral materials. This secures continuity with the measure of intelligibility between various dialects of the minority language. Thus it might be possible to evaluate the position of the minority language relative to Spanish for samples of individual bilinguals from several different test points. In addition, in our tests comprehension takes primacy over speech production.
Finally, measures of comprehension are more feasible than measures of the language a person thinks in. Regarding media variance, the studies concern minority languages that generally have only a spoken form, or at best enjoy only limited use in their written form, and the corresponding cultures are, therefore, mainly pre-literate. For certain of the bilingual communities under consideration, the goal is to create a literature in the minority language. However, data on literacy for these areas are pertinent, since literacy would probably be in the national language and the degree of literacy would be one indication of the extent to which the minority population identifies with majority group values (King 1967). Fishman's situational variance, analyzed into degrees of formality, does not seem necessary to the present method. The description of language use within various domains between individuals enacting specific role relationships may implicitly include the degree of formality. This is because the degree of formality probably depends, to some extent at least, on the interlocutors and on the purpose for enacting the role relationship. Therefore, it seems sufficient to require only the collection of data on comprehension of the national language and on the domains of language use in order to determine the relative "dominance configuration" of languages within bilingual communities at a level of accuracy equivalent to that of the rest of the tests.

(2) The dominance configuration. The term "dominance configuration" is taken from Weinreich, who applied it to the status of two languages relative to individual bilinguals (1953.79). He concluded that the answer to the question of which language was dominant for a particular bilingual may be expressed in terms of a syndrome of characteristics by which a bilingual could be rated: relative proficiency, mode of use, order of learning, age at learning, usefulness in communication, emotional involvement, function in social advance, and literary-cultural value (1953.75-79). He suggested the correlation of types of dominance configurations with types of interference phenomena in order to explain anomalous linguistic changes. In addition he considered that a thorough study of bilinguals and configuration patterns would provide the "understanding of how sociocultural determinants affect the bilingual" (1953.80). For us it is useful to extend the concept to the status of language with reference to an entire bilingual community. It seems reasonable to propose dominance configurations on both the individual and societal levels, with somewhat different characteristics assigned to each (cf. Weinreich 1957.203). For example, the individual configuration might include (a) proficiency in both languages, (b) language use in several domains (this information may be expressed as the frequency with which the language is used within these domains), (c) order and age of learning, (d) attitudes toward the use of the language, and (e) motivations for learning each language.

A dominance configuration for a community might consist of: (a) the relative size of each group in the community that speaks one of the languages as a mother tongue, (b) the proportion of each group that becomes bilingual in the other language, (c) the levels of discourse, or domains, in which the various languages are used, (d) the literary-cultural values of the languages, (e) group attitudes toward the languages, (f)

7. For diglossia it may be necessary to indicate explicitly degrees of formality with language use: Ferguson 1964.431. It seems, however, that a thorough description of language use would reveal diglossia patterns without a specific classification of formality.
direction of language shift, if the bilingual condition is changing, (g) the youngest generation still speaking the minority language, and (h) the proportion of school-age children who do not attend school. A comparison of configuration patterns within communities of a minority language area then would provide an overall view of the relative positions of both languages.

The data required for determining the dominance configuration for an individual or community will be obtained by questionnaires, personal observation, and testing. The variety is necessary to obtain reliable and valid data. Data from questionnaires can be especially problematical. For one thing, the investigator’s lack of knowledge of the local language may affect his conclusions. In early studies in Mexico, investigators often concluded that there was a high degree of bilingualism with Spanish in some communities because almost everyone professed to speak Spanish and no one spoke the vernacular to them. Had the investigators approached the community speaking the local dialect, however, or at least a closely related one, in many cases they might have reached a different conclusion. It is frequently necessary to elicit information about bilingualism indirectly, since people may be sensitive about admitting their use of the vernacular to those who don’t speak it. For example, several informants had said that a dialect of Zapotec was still spoken in a particular town. When the investigators went there, the only person immediately available was a school teacher who denied that anyone in the town still spoke that dialect of Zapotec. However, when the discussion changed to talking about the problems he had with teaching first-year school children in Spanish, the teacher openly admitted that most of these children were monolingual in the vernacular when they first started school. Another problem is that it is easy for inconsistencies to enter into the questioning about bilingualism, as it takes place with different individuals and at successive test points. The investigator must be thoroughly acquainted with the intent of each item on the questionnaire before he begins to use it. Finally, it is necessary to ensure that informants understand the questions adequately. Therefore the questions should be directed toward specific kinds of information; they should be uniformly stated if possible. Even with these precautions, it is reasonable to expect that some individual responses will not be valid. Some respondents may purposely give misleading answers.

(a) Domains of language use. The following questionnaire is suggested as an approach to collecting data on language use in order to determine individual and community dominance configurations. It is adapted for Mexico in particular. Part I seeks data about which language individual informants use within domains of the family, entertainment, local authority structure, district center authority structure, employment, the market, and compadrazgo (a ritual, kin-like system). Part II seeks data about community language use within the domains of the family, school, the town hall, mass media, religious activities, and play and entertainment. The section titled “Additional Data on Bilingualism” provides for the inclusion of demographic information and informal observations.

8. This last may be an important indicator of a high degree of monolingualism in the minority language. For example, a school teacher in one Zapotec town told us that of 250 school-age children in the town, only thirty-seven were attending school. Of these thirty-seven, he added, “Some of them speak Spanish.”

9. Richard Watson reports similar problems in investigating bilingualism between Bahnar and Rengao in the Kontum area of Vietnam (personal communication).
Some of the data that describe community configurations can be partially derived from the sections that are primarily directed toward determining individual language use, especially if the frequency data are based on an investigator's personal observations. For example, while waiting in the town hall one can make note of how many people talk to town officials in the vernacular and how many address them in Spanish. Other sources for observing language use include children playing in the street and people buying at a store. However, it is necessary to consider the possibility that the language used may be determined by the presence of an outsider. That is, people may be embarrassed to speak the minority language in front of an outside investigator and automatically shift to the national language.

Finally, even though the questionnaire is adapted for Mexico, not all important parameters for one area of Mexico are being tested. Other parameters could be included, and some of the domains outlined by the questionnaire could be further analyzed. The goal of the questionnaire is to obtain data about diagnostic items within the limits of the dialect survey program. For bilingual studies outside Mexico, it will be necessary to adapt the questionnaire to local patterns.

**DOMAINS OF LANGUAGE USE**

*Part I: Individual Questionnaire*

<table>
<thead>
<tr>
<th>Name:</th>
<th>Sex:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age:</td>
<td>Education:</td>
</tr>
<tr>
<td>Town:</td>
<td>Minority language:</td>
</tr>
</tbody>
</table>

1. What language do you most commonly use in the home...
   - a. with father?
   - b. with mother?
   - c. with spouse?
   - d. with guardian?
   - e. with siblings?
   - f. with domestics?
   - g. with friends?

2. What language do you most commonly use during entertainment?

3. What language do you most commonly use in the town hall...
   - a. with the town president?
   - b. with the secretary?
   - c. with the judge?
   - d. with the to pilés?11
   - e. with other people?

4. What language do you most commonly use in the district center...
   - a. with the town president?
   - b. with the secretary?
   - c. with the to pilés?
   - d. with the soldiers?
   - e. with store keepers?
   - f. with friends?

5. What language do you most commonly use at work...
   - a. with your employer?
   - b. with your co-workers?

---

10. This questionnaire is based partly on the Schmidt-Rohr classification mentioned earlier, partly on Mackay's classification mentioned by Fishman (1964:38), partly on Lieberson's census questionnaire (1967:143), and partly on my own survey experience.
11. The to pilés is something of an official errand boy, or representative, of the town hall for numerous minor duties.
6. What language do you most commonly use in the market?

7. What language do you most commonly use with your *compadre/comadre*?

**Part II: Community Language Use**

1. What language is most commonly spoken in the home? (These data can be represented in a summary chart derived from Part I.)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Person spoken to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parents</td>
</tr>
<tr>
<td>a. Father</td>
<td></td>
</tr>
<tr>
<td>b. Mother</td>
<td></td>
</tr>
<tr>
<td>c. Guardian</td>
<td></td>
</tr>
<tr>
<td>d. Children</td>
<td></td>
</tr>
<tr>
<td>e. Domestic</td>
<td></td>
</tr>
</tbody>
</table>

2. What languages are used in the school?

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. For first year instruction</td>
<td></td>
</tr>
<tr>
<td>b. For subsequent instruction</td>
<td></td>
</tr>
<tr>
<td>c. Languages taught as subjects</td>
<td></td>
</tr>
<tr>
<td>d. During recess and play</td>
<td></td>
</tr>
</tbody>
</table>

3. What language is most commonly used in the town hall?

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Person spoken to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>President</td>
</tr>
<tr>
<td>a. President</td>
<td></td>
</tr>
<tr>
<td>b. Secretary</td>
<td></td>
</tr>
<tr>
<td>c. Judge</td>
<td></td>
</tr>
<tr>
<td>d. Topil</td>
<td></td>
</tr>
<tr>
<td>e. Other</td>
<td></td>
</tr>
</tbody>
</table>

4. What languages are used for mass media?

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Radio and TV</td>
<td></td>
</tr>
<tr>
<td>b. Press</td>
<td></td>
</tr>
<tr>
<td>c. Literature</td>
<td></td>
</tr>
</tbody>
</table>

5. What languages are used in religious activities?

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Sermon</td>
<td></td>
</tr>
<tr>
<td>b. Prayer</td>
<td></td>
</tr>
<tr>
<td>c. Liturgy</td>
<td></td>
</tr>
<tr>
<td>d. Band</td>
<td></td>
</tr>
<tr>
<td>e. Dances</td>
<td></td>
</tr>
<tr>
<td>f. Other</td>
<td></td>
</tr>
</tbody>
</table>
6. What languages are used for play and entertainment?

<table>
<thead>
<tr>
<th></th>
<th>Idiom</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Adults</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Children</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7. Additional data on bilingualism:

a. Statistics on school children:
   1. Number of school age
   2. Number attending school (by grade)
   3. Number monolingual upon entering school

b. School system:
   1. Federal school
   2. Number of grades taught
   3. Are there bilingual teachers?

c. Proportion of population speaking the idiom:
   1. Older men and women
   2. Younger men and women
   3. Unmarried
   4. Children (up to about ten years)

d. Language breakdown for the community (mother tongue groups: specify dialect and proportion of population):
   1. Spanish
   2. Local dialect of minority language
   3. Outside dialect of minority language
   4. Other minority language

e. Language most frequently heard:
   1. In home
   2. Playground or street
   3. In town hall

(b) Language use on the community level. Whether the position of two languages in a bilingual community is static or changing depends in part upon the functions each language serves for the community members, and in part upon the extent to which the community members agree in the use of each language for each function. Following Fishman, we take it that a state of bilingualism in which both languages serve identical functions, for individuals as well as for communities, will be an unstable one. Conversely, when each of these languages is associated with a complementary set of functions, the condition will be relatively stable. These assumptions provide the motivation for collecting data about domains of language use. They are supported by two kinds of sociolinguistic phenomena. The first one is the specialized form of bilingualism known as diglossia, in which two forms of the same language are spoken by all members of an entire speech community (Ferguson 1964). One of the forms is characteristically viewed as superior. The prestige form, however, is secondarily acquired whereas the lower form is normally acquired as the mother tongue. There is a conventional set of situations appropriate to the use of one or the other variety, and the use of the wrong variety for any situation is cause for ridicule. The description of language choice in diglossia provides a paradigm of language choice at the level of an entire speech community. Table 18 below is a simplification of Ferguson's example. The columns, designated "H" and "L", represent the prestige (high) and vernacular (low) varieties respectively. The "X" indicates which variety is appropriate for the situation indicated at the left.
Besides differential language use and prestige value, Ferguson adds that there are significant differences between the two varieties in literary tradition, descriptive and normative studies, grammatical structure, phonology, and lexicon. The phenomenon is defined from four test cases: Arabic, Modern Greek, Swiss German, and Haitian Creole. From these Ferguson shows that diglossia may have diverse origins and outcomes and that it may remain stable for a thousand years or more (1964.429-430).

While all the examples cited are from literate societies, he considers it possible that a similar situation could exist in a pre-literate one. The relationship between a standard language and its regional or social dialects is similar to diglossia with the distinction that in the former all the varieties, including the standard, are used in ordinary conversation by some group within the speech community, whereas in diglossia the prestigious variety is never used for this purpose (1964.435).

The study of diglossia may also suggest factors influencing the establishment of stable bilingual conditions and their subsequent disruption. For diglossia, in particular, Ferguson suggests three necessary conditions to its establishment: (1) an extensive literature that embodies group values and is in a written form that closely parallels the spoken form, (2) the limitation of literacy (and thus access to the literature) to a small group within the community, and (3) a time lapse of several centuries. He also suggests three trends associated with the disruption of the condition: (1) the desire for more widespread literacy, (2) the need for more adequate communication among the segments of the speech community, and (3) the desire for a standard language as a symbol of group autonomy (1964.436).

The second sociolinguistic phenomenon substantiating Fishman's suggestions is that of two unrelated languages which have specific and complementary roles in a bilingual community. As in the case of the standard language with its regional and social variants, this condition is analogous to diglossia but is distinguished from it by the fact that both languages serve for normal conversation for separate groups within the bilingual community. The bilingualism between Spanish and Guarani, in Paraguay, analyzed by Rubin, is a clear example of a stable bilingual condition of this type (Rubin 1962).

Rubin analyzed the use of the two languages in terms of power and solidarity relations, an approach suggested by Brown and Gilman
(1960). The power relation is asymmetrical and non-reciprocal, i.e., A is more powerful than B politically, socially, religiously, etc. The solidarity relation is symmetrical and reciprocal; i.e., solidarity exists between two persons, or it does not exist between them. It is the means of distinguishing between people who are equal in power. Rubin's application of the paradigm shows that speakers use Spanish to address both equals and superiors with whom they do not feel solidarity. On the other hand, speakers used Guarani to address superiors, equals, and inferiors with whom they do feel solidarity. In addition, a speaker will address in Guarani an inferior with whom he feels no solidarity (Rubin 1962.54-55). Additional observations were that the change from the use of Spanish to Guarani during courtship indicated a change in the solidarity relation; that drunks used Spanish because it made them feel more confident, an expression of the power relationship; and that while travelling abroad, a group of Paraguayans will use Guarani rather than Spanish as an expression of their national identity, i.e., the solidarity relation.

Although she considered the paradigm useful for analysis, Rubin felt that it was necessary to introduce other parameters in order to describe Paraguayan bilingualism. A summary of these shows the following: (1) the Guarani-Spanish usage correlates with the rural-urban distinction; (2) Spanish is the language of instruction in the school and parents may cooperate by speaking Spanish with their children during school years; (3) language use also correlates with the topic of discussion, e.g., school subjects are discussed in Spanish, while jokes are told in Guarani; (4) upper-class women converse among themselves in Spanish while upper-class men tend to converse in Guarani; and (5) there is a correlation of the upper and lower classes with Spanish-Guarani language use (1962.56-57).

8.3.2. Extralinguistic data. Diebold noted that Mexico is an excellent laboratory in which we can study bilingualism (1962). For studying the acculturation processes active in Middle American Indian groups, we can use a broad outline that Beals proposed (1967.450). Historical data are available describing initial and early contact situations between Spanish and the minority languages. Sociological and psychological factors can be abstracted from these data. Dozier, for example, shows that linguistic acculturation may lead to different results depending upon the nature of the initial contact period. For the Yaqui, the result of a permissive approach to the introduction of European patterns was an amalgamation of the two cultures. The Tewa, on the other hand, reacted to forced acculturation by compartmentalizing Spanish elements into a co-existent system separate from the native one (Dozier 1964.509). Data from later periods provide additional background for bilingualism studies. Local mestizos often look down on those who speak a minority language. This attitude may be a carryover from liberal government policies toward the Indian in the nineteenth century which were influenced by a form of "social Darwinism" (Adams 1967). Later, the Revolution of 1910 greatly affected many rural communities. It led to the dispersal and fragmentation of many minority language groups with the result that, in some cases, a complete shift to

12. Beals adapts LaFarge's historical framework and applies it for all of Middle America. He posits six main periods of acculturation: (1) Contact and Consolidation, (2) First Colonial Indian Period, (3) Second Colonial Indian Period, (4) First Republican Indian Period, (5) Second Republican Indian Period, and (6) Modern Indian Period. The limits of the periods are vaguely defined, and he does not assign specific dates to them.
Spanish soon followed. In other cases the result was extensive bilingualism. Finally, as a rule, few outsiders learn to speak one of the minority languages, but many speakers of minority languages acquire at least a minimum degree of proficiency in Spanish. This, in turn, is one measure of the influence of the Mexican national culture on these minority groups.

The immediate need for research consists of describing and analyzing particular bilingual situations. A valuable beginning study on language shift toward Spanish has been done by Clark for three coastal Popoluca dialects (1969). Clark shows how extralinguistic factors, both external and internal, relate to the Oluta, Texistepec, and Sayula dialects. Numerous political, sociological, geographical, and psychological forces, operating throughout the same period of time, have produced different stages of language shift within them: the Oluta dialect is nearly extinct; the Texistepec dialect is not spoken by the younger generation and its use is rapidly declining; the Sayula dialect is still spoken by all the age groups, but it is beginning to enter a state of more rapid decline. The largest Mexican town in the area is Acayucan, a relatively recent boom town. The three Popoluca towns in decreasing ease of accessibility to Acayucan are (1) Oluta, (2) Texistepec, and (3) Sayula. Clark's data are summarized in Tables 19 and 20. The first one presents in chronological order the external influences on the Popoluca communities. The towns primarily affected by each factor are indicated in the second column, and the corresponding results are indicated in the third column. The second table presents some of the internal characteristics of the three communities. The parameters described are the youngest generation of those who speak Popoluca, the attitude of the local people to the idiom and to outsiders, and the children's first language.

From Table 19 it can be seen that contact between Popoluca and Spanish has been more intense and varied for the Oluta and Texistepec dialects than for the Sayula dialect. The former have had access to the outside world for more than seventy years, while the latter was relatively isolated until about twenty years ago. In addition to railroad and highway systems, geographical factors have favored Oluta and Texistepec. The arable land of Oluta attracted immigrants and the sulphur deposits of Texistepec brought in outside laborers. Sayula, on the other hand, had neither highly desirable land nor mineral resources to induce contacts from the outside. From Table 20 it is seen that local attitudes toward outsiders are open for both Oluta and Sayula and are reserved for Texistepec. The difference is probably that outsiders are viewed as a threat by Texistepec speakers, but that residents of the other two towns do not feel threatened because of the attitude of outsiders toward the language. In the case of Oluta, possibility of threat is past; in the case of Sayula, it has yet to appear, if it does at all.

In conclusion, Clark predicts that the Texistepec dialect will be practically extinct by the turn of the century, but says that it is difficult to predict the rate of decline for the Sayula dialect. He believes that it will decline slowly until the speakers of that dialect view their language as a hindrance to progress. He predicts that once that happens, the Sayula dialect will rapidly fall into disuse (1964.14).

Clark's study is a valuable introduction to a generally untreated subject. This kind of study is feasible. It would be helpful for other linguists to make similar descriptions of bilingualism in the communities where they are

13. This statement is strictly true only for the years prior to the construction of the transisthmus highway in 1950 (Clark, personal communication).
### External Influence

1. **Spanish conquest**
   - Towns Affected: All three
   - Result: Acceptance of Spanish as standard language and restriction of Popoluca to local use.

2. **Railroad, roads**
   - Towns Affected: Oluta, Texistepec
   - Result: Coming of soldiers speaking Spanish. Ease in going to outside areas for trade.

3. **Revolution**
   - Towns Affected: All three
   - Result: Dispersion of local people to Spanish-speaking areas.

4. **Epidemics, famine**
   - Towns Affected: All three
   - Result: Further emigration to Spanish-speaking towns.

5. **Immigration of Spanish-speaking people**
   - Towns Affected: Oluta
   - Result: Intermarriage and thus further use of Spanish in the home.

6. **Isthmus highway**
   - Towns Affected: Sayula

7. **Discovery of sulphur**
   - Towns Affected: Texistepec
   - Result: Arrival of outside workers who depreciate those who speak the dialect.

#### Table 19. Factors Producing Language Shift in Coastal Popoluca

<table>
<thead>
<tr>
<th>Youngest Generation of Speakers</th>
<th>Oluta</th>
<th>Texistepec</th>
<th>Sayula</th>
</tr>
</thead>
<tbody>
<tr>
<td>to Age 13-20</td>
<td>Grandfathers</td>
<td>Age 13-20</td>
<td>Children</td>
</tr>
<tr>
<td>Idiom</td>
<td>Inferiority</td>
<td>Inferiority</td>
<td>Pride</td>
</tr>
<tr>
<td>Attitude to Outsiders</td>
<td>Open</td>
<td>Reserved</td>
<td>Open</td>
</tr>
<tr>
<td>Attitude to First Language</td>
<td>Spanish</td>
<td>Spanish</td>
<td>1/2 Popoluca</td>
</tr>
</tbody>
</table>

### Table 20. Internal Characteristics of Popoluca Communities

studying. These studies could take into account the effect of government programs of directed change such as those of INI (Instituto Nacional Indigenista "National Indian Institute"). It is interesting to compare the three conditions suggested by Ferguson as those causing the disruption of diglossia (see p. 136) with Mexico’s drive for literacy among the indigenous groups as a means of further establishing national solidarity. In many cases it can be predicted that relatively rapid assimilation will take place. It can also be expected that there will be interesting cases of language maintenance in the face of government programs and fairly extensive contacts with other languages.
8.3.3. Group behavior toward the languages in contact. Group behavior toward each language in a contact situation is another unexplored subject related to the degree of bilingualism in a community and to the future state of bilingualism.

(1) Attitudes toward language use. A primary goal is to evaluate individual and group attitudes toward the languages. Attitudes expressed about particular features of each language seem to pertain mainly to situations in which two dialects of the same language are in contact. Studies of these attitudes may help determine the acceptability of written materials published in a dialect other than the one that the reader or listener speaks. Attitudes about language use, on the other hand, relate to any pair of languages in contact. These are the ones that are most relevant to the stability of bilingualism.

If language use is actually dependent upon the acceptance of cultural values and these values are manifested by attitudes and overt behavior, then it should be possible to describe the various attitudes encountered and correlate them with various behavior patterns observed in order to predict at least some of the future patterns of language use.

One difficulty is that the same sentiment expressed by two persons may have different meanings, depending on the immediate situation. The reaction “We don’t speak the idiom here,” may reveal the informant’s attitude toward an outside investigator just as much as it does toward his mother tongue. A careful evaluation of attitudes toward the languages in contact, whether positive, negative, or indifferent, must be made to see whether they are valid for the group as a whole (or for a significant subgroup) or idiosyncratic. If they are stereotyped, it might be possible to correlate them with overt group behavior. For example, Weber’s study of Pustunich, a village of Yucatán, suggests a significant correlation of attitudes and values with behavior: the reasons for learning each of the languages and the average age at which pre-school Maya children begin to learn Spanish (Weber 1967). She observed that the reasons given by Maya speakers for teaching their children Spanish were more explicitly formulated than those given for teaching them Maya. Specifically, their reasons were grounded on the hope of enabling their children to find new roles outside of the village context (1967.25-26). Corresponding to this greater explicitness is the fact that the children are learning Spanish at an increasingly earlier age. Thus the number of pre-school children who are monolingual in Maya is steadily decreasing. In addition, the use of Maya is being restricted more and more to the home. Similar observations may be valid predictors of language shift toward Spanish for many speech communities in Mexico.

Additional correlates are suggested in studies that Colby carried out among the Zinacantan Tzotzil (1960, 1961). He categorized individual needs into subsistence, status, and affection requirements. In relating this classification to a theory of acculturation, he considered the needs of status and affection to be more crucial than those of subsistence. Therefore, he said that an Indian would change to the Ladino culture only when he considers that the change will reward him with a higher degree of status or attention, or both, than he has in his present culture (1960.243). For the Zinacantan community in particular, Colby concluded that status, or respect, is more highly valued than both rewards of affection and economic wealth.

One of Colby’s purposes was to determine why the Zinacantans resisted efforts to teach girls to speak Spanish and to learn to read it. He used a
questionnaire, administered by a bilingual literate Zinacantan, to collect data on how the Zinacantans viewed the differences between themselves and the mestizo population, attitudes toward the mestizo, and attitudes toward education of girls in the Spanish language. Regarding the first, Colby concluded that the Zinacantan regard themselves as a primarily pre-literate, agricultural people, while they regard the mestizo society as literate, professional, and somewhat industrialized (1961.81). In summarizing attitudes toward the mestizo, Colby found that 64% of the respondents did not know any mestizos at all and that 32% knew mestizos whom they evaluated as good or bad (1961.81). He also found that attitudes against girls learning to speak and read Spanish were based on the fear that the girls would lose their cultural identity, or that they would upset the male dominance pattern in the Zinacantan society (1960.247). Positive attitudes, on the other hand, were based primarily on economic considerations. In agreement with Weber's findings, they were frequently associated with roles outside of the local community (1961.83). Finally, Colby demonstrated that attitudes toward mestizos were significantly correlated with Zinacantan acceptance or rejection of education for girls (1961.84).

(2) Questionnaire for attitudes about language use. The following questionnaire, incorporating ideas from both Weber's and Colby's studies, is suggested as a preliminary approach to collecting data on attitudes toward language use. It can be used with the questionnaire about domains of language use. It attempts to discover psychological factors that can be correlated with overt group behavior. (In particular, does the use of a minority language negatively correlate with positive attitudes toward those who speak only the national language?) It is also designed to determine the relative values that minority language speakers attach to both their own language and the national one. Finally, it seeks to elicit information that can be used to assess either the relative stability of bilingualism or the extent of language shift toward the use of the national language.

PRELIMINARY QUESTIONNAIRE ON VALUES AND ATTITUDES
FOR STUDIES OF BILINGUALISM

Name:  Minority language:  
Age:  Town:  
Sex:  Date:  

1. Do you think it is good to speak the idiom?
2. In what ways is it an advantage to know how to speak the idiom?
3. Do you speak the idiom with your spouse?
4. What languages do you speak with your children?
   a. Idiom only
   b. Spanish only
   c. Both
5. Do your children speak the idiom? (if answer to 4. is b.)
6. How many people here speak the idiom?
   a. No one
   b. A few people
   c. Most of the people
   d. Almost everybody
   e. Everybody
7. How many people do not use the idiom any longer?
   a. A few
   b. Most
8. When did they start to use Spanish instead?
9. How many still want to use the idiom?
   a. Only a few
   b. Most people
   c. Almost everyone
10. Who still speaks the idiom?
11. Do many adolescents still want to speak the idiom?
12. Do they speak it the same way as the heads of the families do?
13. How differently do the younger people speak it?
14. Now we are going to talk about Spanish. How many people here speak Spanish?
   a. Only a few
   b. Most people
   c. Almost everybody
   d. Everybody
15. In what ways is it an advantage to speak Spanish?
16. Which language is it more important to know?
17. Who taught you to speak Spanish?
18. Are there lots of people who still do not know how to speak Spanish?
19. Is it more important for boys to know Spanish than it is for girls?
20. What advantage is it for boys? (if the answerer is "yes")
21. What advantage is it for girls? (if the respondent says "both" to question 19)
22. Now we are going to talk about Ladinos. Do you have Ladino friends?
23. How often do you visit them? (if answer to 22 is "yes") Do you visit them in their homes or only in their stores?
24. Would you want a Ladino for a neighbor?
25. Do many Ladinos live in your part of town?

For the investigator:
26. Did you hear any spontaneous attitudes voiced regarding the use of either language?
27. If so, what were they?
28. How frequently were they voiced?
29. Who voiced them?
30. What is the attitude of the school teacher(s) toward the use of the idiom?
31. Any other observations, impressions, etc.

This questionnaire is the revision of one that was used in a pilot study of bilingualism among the Chontal of Tabasco. Most of the questions are phrased so as to limit the ways in which a respondent can answer them. Some questions might be rephrased or even be developed into a whole set of questions. Others may turn out to be unproductive so that they should be dropped. Transitions between important blocks of questions are indicated. Finally, the items are presented in a logical sequence. An investigator may not want to follow this sequence rigidly, however.\textsuperscript{14}

It may be necessary to rephrase questions eight and twelve. The purpose of question eight is to find out when local people in general began to use the national language in preference to the minority one. Widespread

language shift frequently follows such things as the opening of a road into the area, the establishment of industry, the initiation of government educational programs, and deliberate efforts by local political leaders to abolish the minority language.

The purpose of question twelve is to determine whether the adult generation is aware of differences between the variety of the minority language it speaks and the variety that the unmarried group (under twenty years of age) speaks. This question is sometimes answered with a comment such as, "They (the younger people) use more Spanish words in their speech."

(3) Group perceptions. It is commonly asserted that a group's awareness of its language and the speakers' loyalty to it are key factors in group survival. Counterexamples have already been cited which indicate that language maintenance or shift may depend on practical and ethnic, or cultural, considerations. In certain cases a group seeking national or regional identity will focus on its language and will emphasize its use as an additional means toward group preservation and outside recognition. In these cases, language standardization may follow.

These suggestions are supported by Wolff's study of the relationships among five Nigerian groups that have been affected by shifts in economic and political dependencies since the nineteenth century (1967). Formerly, the three inland groups were politically and economically dependent upon the two coastal groups. This former relationship was reflected in bilingualism, choice of personal names, choice of place names, and language use. Inland groups became quite bilingual in the coastal languages, but few of the coastal peoples learned to speak the inland languages. Individual names and place names were commonly borrowed from the coastal languages. Finally, communication with coastal peoples was carried out in a coastal language as were religious activities and literacy. As a result of the passing of the palm oil trade, the cessation of colonial rule, and the attainment of national independence, the former relationships among these groups have been radically changed. These changes are correlated with changes in the sociolinguistic pattern. The inland groups are becoming bilingual with non-coastal languages rather than with the coastal ones. Generally, the use of foreign names for individuals is declining and there are movements to restore former inland language place names. Finally, language use for religious activities and education is increasingly shifting toward the inland vernacular. Wolff concludes that these trends are indicative of the use of language as a symbol of ethnic identity and as a weapon for use against a formerly dominant group (1967:23-24).

Solenberger's study of language in the Marianas also shows how a group's awareness of its language may influence retention of the language (1962). The sociological pattern again consists of two mother tongue groups, one of which is politically and economically dominant. As usual, bilingualism between the two languages is one-sided; the dominant language is learned by the speakers of the minority language, but not vice versa. In addition, there is a general distinction between the languages used in the home, and languages used in official capacities. The dominant mother tongue group speaks Chamorro; the minority mother tongue group speaks one of several varieties of Carolinian. English is the most common language in official capacities. Several elements cause an intense jealousy between Chamorros and Carolinians. Solenberger points out that when Carolinians deal with
Chamorro bureaucrats, the Carolinian may use English, Japanese, or German, rather than Chamorro, to put himself on an equal basis with the Chamorro. In addition, Carolinians often speak their own language in the presence of the Chamorros, a means of ostracizing the Chamorros because they refuse to learn Carolinian. Solenberger concludes that the Carolinians will retain their language as long as the Chamorros continue to treat them as inferiors and continue to consider them uncivilized (1962.64).

In summary, it seems obvious that group perceptions will directly influence language use. But it is necessary to examine various factors to determine whether these perceptions will affect group behavior leading to language retention or language shift. In the case of immigrants to Australia and America, the mother tongue has often been dropped. In the two cases described above, the mother tongue is being retained.

8.3.4. Tests for evaluating bilingualism. Additional data on bilingualism can be obtained by administering proficiency tests to a sample of the speech community. The purpose of these tests is to evaluate the relative skill that a bilingual speaker demonstrates for both languages. Haugen points out that the structural differences between languages make it difficult to construct strictly comparable tests. He suggests that tests of production and comprehension could be set up for the levels of phonemics, lexicon, and grammar of both languages. Each of these sub-tests would be scored on some gradient; they might also be weighted against one another, e.g., phonemics might be less highly valued than basic vocabulary. From an overall score for both test batteries, then, an index of bilingualism could be derived for each subject (Haugen 1961).

It is desirable to go one step further and derive a point estimate for group bilingualism by averaging the individual indices. However, there are two reasons why it seems premature to attempt such an extrapolation from present test results. First, the test of comprehension that has been most commonly suggested as the measure of bilingual proficiency is an analogous form of one of the intelligibility tests now employed in dialect studies. (One is already being developed for widespread use.) This test of proficiency is a much simpler measure than that outlined by Haugen and it therefore would be a less sensitive measure than Haugen's. Since the measure that is to be used is a gross one, the second problem, that of sampling, becomes critical. There is a feeling that although it is feasible to employ the same form of the test to measure both intelligibility between dialects of a minority language and proficiency in the national language, the sample of subjects is more representative of the population for testing the former than it is for testing the latter. The sampling method can be termed "judgment sampling" with the restrictions on the subject that he be a native resident of a particular town, have learned the local dialect of the minority language before he learned the national language, and be cooperative. Thus, the subjects we test are probably from the segment of the population most bilingual with Spanish, and the sample mean score from the test on Spanish would permit at best only a rough approximation to a population mean, and that would be most likely only in the case that the sample mean indicated a relatively low index of proficiency. In addition to, or even in place of, randomizing the sampling technique, it might be possible to gather questionnaire data from persons other than those normally used as subjects for intelligibility testing, and later administer the Spanish test to them. If these additional subjects are secured by a house-to-house survey rather
than by the town-hall method, they might tend to be more representative of the population with regard to proficiency in the national language. Finally, it will be necessary to use questionnaire data to validate both indices of individual bilingualism and generalizations about group bilingualism, regardless of the kind of test employed.

(1) Diebold's test for incipient bilingualism. Diebold's Huave experiment shows a way to measure bilingual proficiency on the level of basic vocabulary (1964). The first stage consisted of taking a census of the village of San Mateo. The questions concerned the number and kinship relations of the people in each household, the age, sex, and occupations of the household members, individual language use, and the contacts that the household head had with Spanish-speaking communities. From this census data, Diebold noted for the Huave community that Spanish was generally acquired in adolescence or adulthood, that bilingualism with Spanish was primarily restricted to the men, and that Spanish was never the language used in the home.

The census data also provided the basis for testing bilingual proficiency. The household heads were classified into three groups during the census: coordinate bilinguals, subordinate bilinguals, and monolinguals. The classification was impressionistic, based on whether or not there was a noticeable "accent" from Huave in the individual's pronunciation of Spanish and whether or not an individual could form complete sentences in Spanish. The test, consisting of a 100-word lexicostatistics list, was then administered to random samples of ten, selected from each of the bilingual groups. Each person was required to give a resume of his life history, and was presented with the test. The test was administered orally. The subject was required to give a Spanish equivalent to each Huave item. Responses were judged correct even if the Spanish item given in the response was only partially synonymous with a particular Huave item. The results of the test show that the mean sample scores were distinct for the three groups: 97% for the coordinate bilingual group, 89% for the subordinate bilinguals, and 37% for the monolingual group. The ranges for the three groups, however, overlapped: 89-100% for the coordinate bilinguals, 61-94% for the subordinate bilinguals, an 11-68% for the monolinguals. The results may be interpreted as showing that the coordinate-subordinate-monolingual distinction may be a useful classification, but that bilingual proficiency actually approximates a continuous distribution within a population. Diebold concludes by suggesting the need for further research on incipient bilingualism because of its role in subsequent interference and change. He also suggests that sociological factors are of equal importance with linguistic structures in determining the form of interference between languages in contact.

(2) Comprehension tests. Additional proficiency tests are proposed that will measure only the subject's comprehension of the national language. They may be based either on sentences or texts, or both. An individual's comprehension should probably be measured for two varieties of the national language, a regional one and a standard colloquial one. This requires the definition and choice of regional variants prior to the study (suggested by Bradley, personal communication). The questions could be translated into either the local dialect of the minority language or into the regional variety of the national language. The latter seems preferable; any bias that arises because the questions are phrased in the national language
would be offset by errors in translation from the national language to the local idiom. In the case of tests between dialects of a minority language, the opposite position is maintained because of the general linguistic proximity of the dialects in question. In addition, the subjects would likely be bilingual enough that they could readily answer questions directed to them in the national language. Most monolingual subjects would probably refuse to take the proficiency test, regardless of whether or not the questions were translated into the local dialect. This implies an error in measurement, but the basis for it is broader than the point in question.

A test based on an elicited text was employed by Mellema to measure the degree of bilingualism with Spanish for several samples of subjects during the Otomi study. The text was selected from Lazarillo de Tormes, a Spanish novel of the mid-sixteenth century. Adaptations were made in the text, to conform it somewhat to colloquial Mexico City Spanish. Mellema made the recording from a native Spanish speaker reading the adapted version. The interrupted-text format, equivalent to that used in our dialect intelligibility tests, was adopted and the questions were recorded in Spanish; rather than in a local dialect of the minority language. Mellema subsequently administered the test along with test tapes from various Otomi dialects to subjects in several towns. The mean scores ranged from 94% to 47% with intermediate score of 86%, 73%, and 52%.

Two comments should be made about the test. First, the speech style was rapid; Mellema recommends a second recording in which a slower speech style is adopted before the test is used again (personal communication). Second, the use of a literary work as a source for the text is questionable, if for no other reason than the need to adapt it to a preliterate population. It seems that greater comparability between the minority language tests and the Spanish test would have been achieved had a spontaneous text of one's personal experience been elicited from a Mexico City Spanish speaker. Nevertheless, it provides a basis on which to further develop tests of bilingual proficiency.

In the period since this monograph was submitted for publication, Stoltzfus and his associates have been developing a test based on a subject’s comprehension of a set of fifty sentences. Initial tests were carried out in a Tzeltal area of Chiapas, in Xayacatlán de Bravo, Puebla, and in Tetelcingo, Morelos. The construction and administration of the tests, along with the test results and interpretations of the data are given in preliminary reports by Egland (1972) and Stoltzfus (1972).

8.4. Summary. Questions of feasibility and economy require the methodology to describe and evaluate bilingualism between the minority language and the national language as it occurs in many communities. With knowledge about the degree and stability of bilingualism, the direction and rate of language shift, and group preferences of language use, we can more adequately decide whether or not it is necessary, or even worthwhile, to carry out projects of various kinds in minority languages.

This chapter suggests a sociological approach to begin answering pertinent questions which center on the degree and state of bilingualism. The degree of bilingualism might be determined by collecting data on relative proficiency in both languages and language use within various domains of community behavior. These data would be primarily collected at the level of the individual speaker, but they could be generalized to the community level. Extralinguistic data such as demographic, historical,
sociological, and psychological factors should be collected relative to each community and compared for various communities within a given area. On the basis of these data, bilingualism in a community could be described as stable or as in varying degrees of change. Correlations between attitudinal behaviors, overt behaviors, and other types of data might provide the basis for predicting future states of bilingualism. The various types of data will be obtained by research prior to the experiment, and by administering proficiency tests, using questionnaires, and recording personal observations during the experiment. Questionnaires for data on language use and attitudes toward the languages in contact are included for possible use in Mexico. They might also suggest avenues for research in other areas, though it will be necessary to adapt them to varying degrees.
Appendix A

Sample Texts and Sets of Questions

To provide additional examples of the kind of texts from which we construct test tapes and the kind of questions we use, translations of two texts and their corresponding sets of questions are included. The first text is from Ayoquezco, a reference point for the San Antonio Ocotlán segment of the Zapotec dialect survey. The other text is from Comaltepec, a reference point in the Choapan Zapotec area.

A. Text from Ayoquezco:

"How shall I begin? One time I was really unlucky. I had climbed a guava tree to pick a guava so that I could eat it. Unfortunately, it wasn't any good because it was full of worms. It wasn't any good at all. It sure was pretty to look at, though. That's why I climbed the tree. I wanted to eat the guava, but when I went to get it down, it wasn't any good. As I was coming down, I slipped off the limb of the tree. I fell and broke my arm. I stretched out my arm to the ground and broke it. Afterwards I left that place and went home. I came to my house and told my family that I had broken my arm. I told them that I fell from a tree because of a guava that I saw growing on it. I fell from the limb. I also showed them my arm. I told them, 'Look how it is now. I took care of it myself. I made a splint and tied it up so that it will heal properly. I pulled on my arm to make the bones come together. I did it all by myself.'

"That's how it happened to me. It was in September. I took care of the treatment myself. Since I could do it alone, I did not look for a curer or a doctor. And so the fracture healed. It took nine months for the fracture to heal completely."

B. Questions about the Ayoquezco text:

1. What happened to the boy?
2. Where did he go up?
3. Why wasn't the guava good?
4. Why did he want the guava?
5. What happened while he was in the tree?
6. Where did he stretch out his arm to?
7. Where did he go afterwards?
8. What did he do with his arm?
9. Who treated his arm?
10. How long did his arm take to heal?

C. Text from Comaltepec:

"I am going to tell you about some things that have happened to me. It was when the town called me to serve in the town hall. My wife told me, 'It would be better to run away.' She said this because I had been telling her that I ought to run away and not carry out the town responsibilities. After she thought about it, she said to me, 'No, you're not going to run away. You're going to do what I say. You're going to serve in the town hall just like the people want you to.'

"So I followed her advice to go serve in the town hall. I went there and carried out my town duties. While I was there, however, there was no way I could work. I was not able to earn any money. As a result, the men in the town hall began to get angry. They got very disgusted with me.

"One day when I was in the town hall, these men began to try to make me pay a fine. They were trying to fine me for neglecting my family. Finally they quit trying to get money from me. They grabbed me and took me to court. They were all in agreement with my wife; they accused me of malice. So they put me in jail for punishment.

"There I was in jail. While I was there, my wife did not pay any attention to me. Later, my brother came with the money to get me out. So I came home. But my wife had left and had not come back. She went to her own town. There she got sick and died. She never did come back."
"Another time I went to the town of Lacova to sell candies. There I got in a fight with a witchdoctor. Since I did not believe that he was really a witchdoctor, he put a curse on me.

"Exactly one month later, when I came back home, I got sick. My head hurt me and my teeth ached. It wasn't just a cold, either. I couldn't get well. One night a curer came to visit me. Suddenly a bat flew in the house. It was flying round and round inside the house where I was lying down. It wanted to do something bad to me.

"There were four or five people inside the house. They wanted to hit the bat while it was still in the air. They were striking at it with brooms, but they couldn't hit it. Well, I had a little dog. He jumped up into the air, grabbed the bat, and killed it. That was the end of the bat. The curer then picked up the bat. He went outside to throw it into the weeds. And that's how it happened."

D. Questions about the Comaltepec text:

1. What did this man want to do at first?
2. What did the men in the town hall begin to do at first?
3. Who were these men aligned with?
4. What did this man's brother come to do?
5. What had the man's wife done?
6. What happened to her later on?
7. What did this man do with the witchdoctor?
8. What animal came into the man's house at night?
9. How did the animal get killed?
10. What did the curer do with the bat?
The Trique Sentence Tests

Sentence tests were employed exclusively in the Triere dialect survey. With regard to the design of the instrument, I made some changes from the version that I used in the Choapan study. I increased the number of statements in the test from ten to fifteen. I also changed the classification that was needed to make sure that the individual tests were comparable. Finally, I changed the method that I used to form the statements in the tests.

The classification was based partly on my analysis of the sets of questions that had been asked about the content of eighty survey texts, twenty each from the Mixtec, Chinantec, Zapotec, and Aztec studies. In this way I hoped to make the sentence tests approximately equivalent to text tests in terms of content and difficulty. After I classified the information that the questions sought from the texts, I tabulated the relative frequencies with which the various types of information occurred per text. I derived the final classification from the analysis somewhat subjectively; the information class frequencies were not strictly reproduced in the test format.

The classification was as follows for each of the three tests: one manner, one situational-descriptive, one qualitative-descriptive, one meteorological-temporal, one situational-temporal, one explanatory-causative, one purposive-causative, one subject, one locative, one instrument, one result, one body part, two transitive events, and one intransitive event. This classification can be termed "multi-dimensional" in that the categories overlap and cross-cut one another to varying degrees. Situational-descriptive statements also tended to be adjectival constructions but they referred to psychological characteristics. Meteorological-temporal statements involved particular times of the day and other time divisions while situational-temporal statements expressed one transitive or intransitive event in its time relation to another event. Explanatory-animate and purposive-causatives were expressed by transitive or intransitive events; the distinction between the two is somewhat narrow, but the former indicates the ground of or reason for an action while the latter expresses the intent or goal of an action. Subjects were nouns or noun phrases occurring as the topic of the statement. Result was expressed by an event with the meaning that an individual or object was unintentionally involved in some activity. In some cases the result of an event affected a body part of the person involved.

I constructed enough sentences in the above categories to make up the test set. The sentences within each category differed from one another not only with respect to the lexical items in the information category, but also with respect to the lexical items that served other functions and with respect to features such as singular versus plural, animate versus inanimate, and masculine versus feminine. The sentences generally reflected situations common to rural Mexican life.

An individual text was specified by selecting the right number of sentences from each category. The last step was to scramble the order of sentence types from test to test. Translations of the Triere sentence tests follow:

A. Copala

1. Test Sentences and Questions:
   The men were walking slowly.
   How were the men walking?

2. The man is very worried.
   What is wrong with the man?

3. Paco's parents will arrive here in a week.
   When will Paco's parents arrive?

4. The man could not work because he had another obligation.
   Why couldn't the man work?

5. There are a lot of roasting ears of corn on the table.
   What is there on the table?

6. The boy came to a place above the town.
   Where did the boy come to?

7. The young man returned to his house because he wanted to talk with his brother.
   Why did the young man return home?

8. Benito's house burned down two nights ago.
   What happened to Benito's house?
9. Joe's fingers got hurt when they closed the door.
   What happened to Joe when they closed the door?
10. The men crossed the river by horse.
    What did the men do?
11. When my brother saw the snake, he ran from the corn patch.
    Why did my brother run from the corn patch?
12. My brother is very lazy.
    What kind of person is my brother?
13. Pablo stayed on top of the mountain for three days.
    What did Pablo do?
14. Because the sun came out, the boy wanted to go get firewood.
    What did the boy want to do?
15. The woman is sweeping the house.
    What is the woman doing?

B. Chicahuextla test sentences and questions:

1. The young men were joking among themselves.
   What were the young men doing?
2. When the man learned what his son had done, he punished him.
   When did the man punish his son?
3. There are two birds in that tree.
   What is there in the tree?
4. When I did that job, it was a very nasty day.
   What was the weather like when I did that job?
5. The animals were seen below the spring.
   Where were the animals?
6. Some men went slowly down to the river.
   In what manner did the men go down to the river?
7. The men were going after the 'possum.
   What were the men doing?
8. The young man covered the box with leaves.
   What did he cover the box with?
9. That horse is a very fine one.
   What is that horse like?
10. He began to look for his friend in the morning.
    When did he begin to look for his friend?
11. The man is very happy because his brothers are coming to see him.
    Why is the man happy?
12. The young man had been sitting down for a long time, so he got sleepy.
    What happened to the young man?
13. Chucho came to visit Frank in order to pay off his debt.
    Why did Chucho visit Frank?
14. The men took the roof off of the house.
    What did the men do?
15. The mule kicked the man in the shoulder.
    What did the mule do to the man?

C. Itunyoso test sentences and questions:

1. The man left his machete in the house.
   Where did the man leave his machete?
2. When the man fell, he broke his leg.
   What happened to the man when he fell?

3. The boy is unhappy because his father made him go to work.
   Why is the boy unhappy?

4. The thieves came into the town late at night.
   When did the thieves get to the town?

5. That little dog is very tame.
   What is the little dog like?

6. The child is running along the trail.
   What is the child doing?

7. The man warmed himself with a fire.
   How did the man make himself warm?

8. The young man took money along so that he could buy corn.
   Why was the young man carrying money?

9. The men worked hard all day long.
   How did the men work all day long?

10. The man rested, and then he went out and fed his animals.
    What did the man do after he rested?

11. My brother could not plant his cornfield because it was raining.
    What work couldn't my brother do?

12. There are a lot of boards in the house.
    What is in the house?

    What happened to my friend?

14. It started to rain, so the man went into his house.
    Why did the man go into his house?

15. The shirt is wet.
    In what condition is the shirt?
### Sample Score Sheet

<table>
<thead>
<tr>
<th>Name:</th>
<th>Sex:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialect:</td>
<td>Age:</td>
</tr>
<tr>
<td>Education:</td>
<td>Travel:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A.</th>
<th>1.</th>
<th>6.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>7.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>8.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>9.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>10.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B.</th>
<th>1.</th>
<th>6.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>7.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>8.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>9.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>10.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C.</th>
<th>1.</th>
<th>6.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>7.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>8.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>9.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>10.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D.</th>
<th>1.</th>
<th>6.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>7.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>8.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>9.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>10.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E.</th>
<th>1.</th>
<th>6.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>7.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>8.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>9.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>10.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F.</th>
<th>1.</th>
<th>6.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>7.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>8.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>9.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>10.</td>
<td></td>
</tr>
</tbody>
</table>
Appendix D

Standard Summary Charts

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix E

Ethnographic Questionnaire
Mexico Branch Dialect Survey
(October 1967)

LOCAL AREA INFORMATION

A. Locational Information
   1. map of: topography, communication, distances with walking or driving time
   2. longitude and latitude
   3. reason for being a focus point (if true): mail, telegraph, etc.
   4. altitude and climate
   5. population
   6. history as to origin

B. Prestige factor among villages: give reasons

C. Political and power structure
   1. check one:
      a. Cabecera de distrito
      b. Municipio
      c. Agencia de municipio
      d. Ranchería
   2. organization, membership and rank—
      a. how elected?
      b. term vs. service?
      c. pay?
      d. contrast with neighboring areas?
      e. surface vs. real authority?
      f. comments on succession of office and qualifications

D. Economy
   1. imports: from where? exchange?
      frequency? language used?
   2. exports: to where? exchange?
      frequency? language used?
   3. local resources: what?
      from where?
      how much?

E. Kinship ties
   1. marriage with people from other villages
      a. frequency?
      b. language adaptation: for parents? for children?
   2. where are the compadres?
F. Education
1. extent offered?  2. in what language?
3. percent who take advantage?  4. reason for dropouts?
5. is there government control (local or other)?
6. teachers: local or mestizo? speak idiom?
   attitude to job?  attitude to idiom?
7. graded materials for all grades?
8. attitude of town toward local language? to Spanish?

G. Additional comments:

INDIVIDUAL QUESTIONS

A. Language
1. in the home: parents' language? where born?
   children's language?  do they ever use another?
2. attitude of parents: to native language? to Spanish?
   children: to native language? to Spanish?
3. in religious services, fiestas, weddings, etc.?
4. in the town hall?

B. Migration
1. how many have come?  from where?
   gone?  to where?
2. reasons for above?
3. do they stay?

C. Education
1. of person being interviewed? attitude?
   of his children? attitude?

D. Contact with the outside
1. travel: how much? to where? reasons?
2. reading materials? pictures? records?

E. Additional comments:
Appendix F

Summary of Studies Carried out in Mexico

Partly in order to show how widely the Mexico branch of the Summer Institute of Linguistics has already used intelligibility testing, and partly to give a brief report of work in progress, we have cited data from most of the studies in which the method has been employed. We have omitted some earlier studies whose results are considered to be very tentative, as well as some later ones that are barely started. This summary is also given to acknowledge the help of many field workers, including some outside of SIL, who have graciously cooperated in the dialect survey program. For each study included, the summary gives the names of the investigators, the year in which it was carried out, the number of test points selected, the number of subjects tested, and the approximate percentage of testing completed. Complete reports of all these studies (and others not yet started) are planned for inclusion in a subsequent monograph.

A. Participants:

5. Chontal of Tabasco: Kathryn Keller, David and Jan Persons.
15. Totonac I: Herman Aschmann and Kent Gordon.
16. Totonac II: Ruth Bishop, Ella Marie Button, Aileen Reid, Eugene Casad, and David Persons.
17. Trique: Lester Blank, Claude Good, Bruce Holienbach, and Eugene Casad.
### B. Dialect Survey Summary Chart:

<table>
<thead>
<tr>
<th>Language</th>
<th>Year</th>
<th>Test</th>
<th>Points</th>
<th>Subjects</th>
<th>Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aztec</td>
<td>1969-</td>
<td>19</td>
<td>158</td>
<td></td>
<td>75%</td>
</tr>
<tr>
<td>Chatino</td>
<td>1968</td>
<td>11</td>
<td>66</td>
<td></td>
<td>90%</td>
</tr>
<tr>
<td>Chinantec</td>
<td>1967</td>
<td>26</td>
<td>260</td>
<td></td>
<td>85%</td>
</tr>
<tr>
<td>Chol</td>
<td>1966</td>
<td>8</td>
<td>70</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Chontal</td>
<td>1970</td>
<td>7</td>
<td>71</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>HuaAve</td>
<td>1968</td>
<td>1</td>
<td>8</td>
<td></td>
<td>90%</td>
</tr>
<tr>
<td>Mazahua</td>
<td>1970</td>
<td>5</td>
<td>18</td>
<td></td>
<td>40%</td>
</tr>
<tr>
<td>Mazatec</td>
<td>1966-67</td>
<td>25</td>
<td>240</td>
<td></td>
<td>95%</td>
</tr>
<tr>
<td>Mixtec</td>
<td>1965-68</td>
<td>69</td>
<td>683</td>
<td></td>
<td>95%</td>
</tr>
<tr>
<td>Otomí</td>
<td>1968-</td>
<td>14</td>
<td>140</td>
<td></td>
<td>50%</td>
</tr>
<tr>
<td>Pame</td>
<td>1968</td>
<td>3</td>
<td>22</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Popoloca</td>
<td>1970</td>
<td>6</td>
<td>58</td>
<td></td>
<td>75%</td>
</tr>
<tr>
<td>Tepehua</td>
<td>1969</td>
<td>3</td>
<td>21</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Tzephuán</td>
<td>1965</td>
<td>5</td>
<td>17</td>
<td></td>
<td>50%</td>
</tr>
<tr>
<td>Totonac I</td>
<td>1965</td>
<td>15</td>
<td>101</td>
<td></td>
<td>70%</td>
</tr>
<tr>
<td>Totonac II</td>
<td>1967-</td>
<td>8</td>
<td>92</td>
<td></td>
<td>70%</td>
</tr>
<tr>
<td>Trique</td>
<td>1970</td>
<td>5</td>
<td>59</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Tzotzil</td>
<td>1965</td>
<td>6</td>
<td>59</td>
<td></td>
<td>90%</td>
</tr>
<tr>
<td>Zapotec</td>
<td>1967</td>
<td>27</td>
<td>253</td>
<td></td>
<td>60%</td>
</tr>
<tr>
<td>Zoque</td>
<td>1968</td>
<td>4</td>
<td>30</td>
<td></td>
<td>85%</td>
</tr>
</tbody>
</table>
Appendix G

Vowel Shifts between the San Lorenzo and Santa María Dialects of Mazatec

Data from Kirk (1966) show many differences between the vowel system of San Lorenzo and the vowel systems of other Mazatec dialects. A brief comparison of the vowel system of San Lorenzo and Santa María Asunción reveals that San Lorenzo and Santa María both have five-vowel systems (omitting the vowel clusters), while San Jerónimo has a four-vowel system. The high, back, rounded vowel /u/, shared by San Jerónimo and Santa María corresponds to the high, mid, unrounded vowel /i/ in the San Lorenzo dialect. In addition, the low, back, rounded vowel /o/, common to San Lorenzo and Santa María, has fallen together with /i/ in the San Jerónimo dialect (Gudschinsky 1958:14). Finally, the San Lorenzo dialect has only two vowel clusters as opposed to seven for each of the other two dialects. The basic systems are presented below schematically for the three dialects.

a. San Lorenzo
   1 4  ia  i  u  ie  ia  lu
   e  o  ai  e  a  ui  ue  ua
   a

b. San Jerónimo
   1 4  ia  i  u  ie  ia  lu
   e  o  ai  u  ui  ue  ua
   a

c. Santa María Asunción (Jiotes)
   i  u  ie  ia  iu
   e  o  ai  ul
   a  ue
   ua

Santa María was chosen for comparison because it was considered to be well within the Huautla dialect, which is poorly understood by speakers of San Lorenzo, and because the San Lorenzo tape was used there. Santa María speakers scored 25% on the San Lorenzo reference tape. San Jerónimo was included because it is considered the dialect center for reaching the San Lorenzo dialect.

On the surface, the phonological difference between the vowel system of the San Lorenzo dialect and the Santa María dialect would seem to be minimal, since there are the same number of vowels in the inventories of both and the only difference seems to be a topological change between one pair of vowels. (This, of course, ignores the vowel clusters, whose analysis is quite tentative.) However, sets of correspondences between the two systems show a remarkable difference in the distribution of these vowels, and it is suggested here that this difference is great enough to be a significant factor in hampering intelligibility between the San Lorenzo and Santa María dialects.

The rules presuppose a hearer model, since the purpose is to try to explain why a person cannot understand the other dialect. They utilize a labelled-bracketing, raw notation. In addition, phonemic notation is used. Each rule, then, will be read as follows: if a Santa María dialect speaker hears a word of the San Lorenzo dialect in which San Lorenzo phoneme \( v_i \) occurs, he will have to interpret it as one of the following vowels in his own dialect: \( v_1, v_2, \) or \( v_3, \) in order to understand the San Lorenzo word. (This assumes, of course, that the lexical items are cognate and mean the same thing.) In most cases there is more than one possibility for the interpretation of a particular San Lorenzo vowel (symbolized as \( /v/Lo \)). While there are probably some kinds of environmental restrictions determining which Santa María vowel (symbolized as \( /v/Ji \)) is to be paired with a given San Lorenzo vowel, these are not immediately obvious, and it has proved easiest to state the pairings in terms of the relative frequencies for each sub-rule. The following schema illustrates the form of these rules:

\[
(1) \quad /v_i/Lo \quad \{ /v_1/Ji \} .72 \quad \{ /v_2/Ji \} .18 \quad \{ /v_3/Ji \} .10 \quad N = \_
\]
These rules can be read as follows: For a sample of N occurrences of San Lorenzo \( v_i \), the correct frequency for pairing it with Santa Maria \( v_i \) will be 72\%, that for pairing it with Santa Maria \( v_2 \) will be 18\%, and that for pairing it with \( v_3 \) will be 10\%. The Santa Maria vowels will always be put in descending order of frequency from top to bottom within the set of brackets immediately to the right of the arrows, and the relative frequencies will be indicated on the corresponding lines within the right-hand set of brackets in the rule.

The "hearer" rules for vowel correspondences between the two Mazatec dialects are as follows:

\[
\begin{array}{c|c|c}
(2) & (i) & /i/Lo \\
 & /i/ & \text{Ji} \\
 & /u/ & .17 \\
 & /e/ & .045 \\
 & /o/ & .045 \\
 & /a/ & .70 \\
 & \text{N} = 133. \\
\end{array}
\]

\[
\begin{array}{c|c|c}
(2) & (ii) & /e/Lo \\
 & /e/ & \text{Ji} \\
 & /a/ & .08 \\
 & /o/ & .08 \\
 & \text{N} = 25. \\
\end{array}
\]

\[
\begin{array}{c|c|c}
(2) & (iii) & /a/Lo \\
 & /a/ & \text{Ji} \\
 & /i/ & .12 \\
 & /o/ & .08 \\
 & \text{N} = 65. \\
\end{array}
\]

\[
\begin{array}{c|c|c}
(2) & (iv) & /o/Lo \\
 & /o/ & \text{Ji} \\
 & /i/ & .33 \\
 & /u/ & .06 \\
 & /a/ & .89 \\
 & \text{N} = 127. \\
\end{array}
\]

\[
\begin{array}{c|c|c}
(2) & (v) & /i/Lo \\
 & /i/ & \text{Ji} \\
 & /u/ & .14 \\
 & /o/ & .02 \\
 & /a/ & .84 \\
 & \text{N} = 72. \\
\end{array}
\]

In considering these rules notice (1) that there is no topological correspondence between the systems, i.e., /i/Lo corresponds highly with /i/Ji in its distribution, but none of the other pairs of phonemes corresponds in this way, and (2) that there is no strict one-to-one correspondence between any pair of phonemes, though there is a statistical preference shown for such a correspondence by these rules. A reformulation of these rules in the opposite direction, i.e., "production rules" for a Santa Maria speaker learning the San Lorenzo dialect, shows the same kind of correspondences and statistical preferences. Thus, while the inventories of the two vowel systems show almost perfect correspondence, the distributional relationships of the two systems are much more complex. It seems likely that the cumulative effect of all these vowel shifts on intelligibility, then, would be to seriously hamper the transfer of information between the two dialects.

Some of the data need to be reanalyzed and the rephonemicization of them would undoubtedly require a revision of some of these rules, possibly in the direction of simplifying them (Paul Kirk, personal communication). I discarded sub-rules which were supported by only one example from Kirk's cognate sets. Note, however, that these rules cover only a small part of the phonological systems of the two dialects (omitting consonant shifts and tone changes, for example), and linguistic structures outside of the phonological systems are not even treated here. Finally, in most cases, these rules are valid for several dialects, not just the Santa Maria dialect.

This kind of data can be used profitably to help validate intelligibility scores. It might be fruitful to compare vowel distributions for several pairs of dialects and rate the pairs by various criteria of degree of complexity. The next step would be to correlate the indices of complexity for pairs of dialects with the intelligibility scores for these paired dialects. It is reasonable to suspect that the more related a pair of dialects are, the more alike the distribution patterns of their vowel phonemes will be. Thus, there might be a correlation between the two measures in many instances (see Section 7.1).

The following paradigms are abstracted from Kirk's cognate sets. They are included to partially demonstrate the phenomena summarized in (2) above. They illustrate most of the sound shifts and the frequency data incorporated into the rules. The numbers to the right indicate the page number in Kirk (1966) from which the data were taken.

1. Determined from an examination of the cognate sets in Kirk 1966.214-484.
### Appendix G: Vowel Shifts between Mazatec Dialects

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td></td>
<td>(2)</td>
<td></td>
<td>(3)</td>
</tr>
<tr>
<td>four</td>
<td>372</td>
<td>corn (dry ear)</td>
<td>337</td>
<td>two</td>
</tr>
<tr>
<td>Ji</td>
<td>ɲɪhṹ</td>
<td>Ji</td>
<td>ɲɪ́hɨ́</td>
<td>Ji</td>
</tr>
<tr>
<td>Te</td>
<td>ɲɪhṹ</td>
<td>Te</td>
<td>ɲɪ́hɨ́</td>
<td>Te</td>
</tr>
<tr>
<td>Lo</td>
<td>ɲìhɨ́</td>
<td>Lo</td>
<td>ɲɛ̃hé̂̌</td>
<td>Lo</td>
</tr>
<tr>
<td></td>
<td>(unexpected reflex)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td></td>
<td>(5)</td>
<td></td>
<td>(6)</td>
</tr>
<tr>
<td>paper</td>
<td>153</td>
<td>ten</td>
<td>424</td>
<td>near</td>
</tr>
<tr>
<td>Ji</td>
<td>sṹ</td>
<td>Ji</td>
<td>té</td>
<td>Ji</td>
</tr>
<tr>
<td>Te</td>
<td>súhṹ</td>
<td>Te</td>
<td>té</td>
<td>Te</td>
</tr>
<tr>
<td>Lo</td>
<td>sìhɨ́</td>
<td>Lo</td>
<td>tá</td>
<td>Lo</td>
</tr>
<tr>
<td>(7)</td>
<td></td>
<td>(8)</td>
<td></td>
<td>(9)</td>
</tr>
<tr>
<td>twenty</td>
<td>280</td>
<td>animal</td>
<td>254</td>
<td>clay griddle</td>
</tr>
<tr>
<td>Ji</td>
<td>kã́</td>
<td>Ji</td>
<td>cṹ</td>
<td>Ji</td>
</tr>
<tr>
<td>Te</td>
<td>kã́</td>
<td>Te</td>
<td>cṹ</td>
<td>Te</td>
</tr>
<tr>
<td>Lo</td>
<td>kã́</td>
<td>Lo</td>
<td>cí</td>
<td>Lo</td>
</tr>
<tr>
<td>(10)</td>
<td></td>
<td>(11)</td>
<td></td>
<td>(12)</td>
</tr>
<tr>
<td>far</td>
<td>291</td>
<td>toasted</td>
<td>223</td>
<td>thread</td>
</tr>
<tr>
<td>Ji</td>
<td>khĩ́</td>
<td>Ji</td>
<td>s‘ichṹ</td>
<td>Ji</td>
</tr>
<tr>
<td>Te</td>
<td>khĩ́</td>
<td>Te</td>
<td>chṹ</td>
<td>Te</td>
</tr>
<tr>
<td>Lo</td>
<td>khĩ́</td>
<td>Lo</td>
<td>čhĩ́</td>
<td>Lo</td>
</tr>
<tr>
<td>(13)</td>
<td></td>
<td>(14)</td>
<td></td>
<td>(15)</td>
</tr>
<tr>
<td>road</td>
<td>360</td>
<td>long</td>
<td>361</td>
<td>ant</td>
</tr>
<tr>
<td>Ji</td>
<td>ntiyã́</td>
<td>Ji</td>
<td>ntuhṹ</td>
<td>Ji</td>
</tr>
<tr>
<td>Te</td>
<td>ntiyã́</td>
<td>Te</td>
<td>ntuhṹ</td>
<td>Te</td>
</tr>
<tr>
<td>Lo</td>
<td>ntiyṍ</td>
<td>Lo</td>
<td>ntðɨ́</td>
<td>Lo</td>
</tr>
<tr>
<td>(16)</td>
<td></td>
<td>(17)</td>
<td></td>
<td>(18)</td>
</tr>
<tr>
<td>cornhusk</td>
<td>441</td>
<td>cotton</td>
<td>219</td>
<td>uncle</td>
</tr>
<tr>
<td>Ji</td>
<td>čhó</td>
<td>Ji</td>
<td>cankã́</td>
<td>Ji</td>
</tr>
<tr>
<td>Te</td>
<td>čhṹ</td>
<td>Te</td>
<td>cankã́</td>
<td>Te</td>
</tr>
<tr>
<td>Lo</td>
<td>čhã́</td>
<td>Lo</td>
<td>conkö</td>
<td>Lo</td>
</tr>
<tr>
<td>(19)</td>
<td></td>
<td>(20)</td>
<td></td>
<td>(21)</td>
</tr>
<tr>
<td>earthquake</td>
<td>242</td>
<td>owl</td>
<td>271</td>
<td>passes by</td>
</tr>
<tr>
<td>Ji</td>
<td>ɔ́</td>
<td>Ji</td>
<td>hñṹ</td>
<td>Ji</td>
</tr>
<tr>
<td>Te</td>
<td>ɔ́</td>
<td>Te</td>
<td>hñṹ</td>
<td>Te</td>
</tr>
<tr>
<td>Lo</td>
<td>ɔ́</td>
<td>Lo</td>
<td>hṇ̃é́</td>
<td>Lo</td>
</tr>
<tr>
<td>(22) mouse 335</td>
<td>(23) waist 348</td>
<td>(24) sugar cane 349</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td>-------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ji niaé</td>
<td>Ji ntayá</td>
<td>Ji ntihi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Te niaé</td>
<td>Te ntayé</td>
<td>Te ntihi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lo niaé</td>
<td>Lo ntoyá</td>
<td>Lo ntehé</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(25) dough 350</th>
<th>(26) green 378</th>
<th>(27) six 402</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ji ng'ýú</td>
<td>Ji sasé</td>
<td>Ji hç</td>
</tr>
<tr>
<td>Te nayó</td>
<td>Te sasé</td>
<td>Te hç</td>
</tr>
<tr>
<td>Lo nç'i</td>
<td>Lo sosá</td>
<td>Lo hç</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(28) firewood 240</th>
<th>(29) woman, wife 247</th>
<th>(30) glass 250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ji čikč</td>
<td>Ji čhù</td>
<td>Ji čic Ł</td>
</tr>
<tr>
<td>Te čikč</td>
<td>Te čhù</td>
<td>Te čic Ł</td>
</tr>
<tr>
<td>Lo čokę</td>
<td>Lo čhù</td>
<td>Lo čic Ł</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(31) alone 280</th>
<th>(32) root 283</th>
<th>(33) not yet 288</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ji s'ik'č</td>
<td>Ji hamą</td>
<td>Ji khé</td>
</tr>
<tr>
<td>Te kač Ł</td>
<td>Te hamayá Ł</td>
<td>Te khé</td>
</tr>
<tr>
<td>Lo koç Ł</td>
<td>Lo homç Ł</td>
<td>Lo khiá</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(34) cloth 286</th>
<th>(35) needs 299</th>
<th>(36) corn cob 304</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ji nikhé</td>
<td>Ji mąchę</td>
<td>Ji nahną Ł</td>
</tr>
<tr>
<td>Te nikhé</td>
<td>Te mąchę</td>
<td>Te nahnę Ł</td>
</tr>
<tr>
<td>Lo nikhiá Ł</td>
<td>Lo mąchę Ł</td>
<td>Lo nohną Ł</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(37) teeth 340</th>
<th>(38) cave 341</th>
<th>(39) market 353</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ji ni'ńų Ł</td>
<td>Ji nkahó Ł</td>
<td>Ji ntic Ł</td>
</tr>
<tr>
<td>Te ni'ńų Ł</td>
<td>Te nkihó Ł</td>
<td>Te ntic Ł</td>
</tr>
<tr>
<td>Lo ną'į Ł</td>
<td>Lo nkohw Ł</td>
<td>Lo ntač Ł</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(40) early 397</th>
<th>(41) hard 421</th>
<th>(42) cooking pot 434</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ji šitl</td>
<td>Ji tahá</td>
<td>Ji tí</td>
</tr>
<tr>
<td>Te šatí</td>
<td>Te tahá</td>
<td>Te tihi</td>
</tr>
<tr>
<td>Lo šotí</td>
<td>Lo tohó</td>
<td>Lo tehé (unexplained)</td>
</tr>
</tbody>
</table>
Appendix H

Statistical Measures

This section shows how to compute each statistic that has been applied in this monograph. It is meant to help the field technician who does not have a statistics text on hand but would like to make some preliminary analyses while he is actually collecting the data. This section is meant only as a stop-gap measure to use until the technician can obtain an adequate textbook. I have found Downie and Heath, *Basic Statistical Methods* (1965), very useful. It is an elementary text that covers a wide range of statistics and contains problems that are well matched with the discussion in the text in terms of difficulty for the student.

A. The Mean, Median, and Mode:

The mean, median, and mode are three kinds of averages that are used to describe a set of observations. They are called *measures of central tendency* since they in some way represent the most typical valued observation in the set.

The *mean* is what is usually referred to when people talk about the average score, for example. It is computed by adding up the values of all the observations such as individual scores (or measures of distance, length, capacity, or time) and dividing the total by the number of observations in the set. It is represented by the formula

\[ \bar{X} = \frac{\Sigma X}{N} \]

where \( \bar{X} \) (X-bar) is the mean value, \( X \) is an individual value, \( \Sigma X \) is the sum of the individual values, and \( N \) is the number of values in the set. For example, the mean of 100, 90, 90, 100, 70, 80, 80, 90, 90, 80 is

\[ \bar{X} = \frac{870}{10} = 87. \]

The *median value* is the middle value (or the mean of the two middle values) of a set of observations that have been ranked from high to low. If there are an odd number of observations, the value of the middle observation is the median. If there are an even number of observations, the mean value of the two middle scores will be the median. For example, the median value of \{9, 8, 7, 6, 5\} is 7, whereas the median of \{9, 8, 7, 6\} is 7.5. Tie scores may present a further complication which I will bypass here (Roscoe 1969).

The *mode* is the value that occurs most often in the set. For example, the modal value of \{21, 20, 18, 18, 18, 17, 14, 11\} is 18. Whereas there will be only one mean and one median value in a set, there may be more than one modal value, or none at all (Roscoe 1969.41).

The mean is the most stable of these three measures of central tendency. Its value does not vary as much from sample to sample as median and modal values do. Therefore the mean value is likely to be a good estimate of the true population value (Roscoe 1969). However, the mean is affected by extremely high or extremely low scores in a set of observations. In these cases, the median represents the average value more accurately since it is not affected by such scores. (This may be especially true with small samples.)

The median is almost exclusively used for descriptive purposes. The mode is completely restricted in this way (Downie and Heath 1965.44). The mean, on the other hand, is frequently used in further computations to make statistical inferences (Roscoe 1969.43).

In describing data, it may be helpful to report all three measures together. The relationship of the mean, median, and mode may provide clues about the nature of the distribution of a property within a population or, if the true distribution is known to be normal, the relationship of these averages may help determine the extent and direction in which the sample distribution is skewed. For example, in a normal distribution, the three averages have identical values. In a positively skewed distribution (the longer tail of the curve at the right on a graph), the mean value will be higher than the median, which in turn will be higher than the modal value. This order is reversed for a negatively skewed
distribution (the longer tail of the curve at the left of the graph; Downie and Heath 1965,44). Table 21 presents the mean, median, and modal scores for part of the Mazatec test results.

<table>
<thead>
<tr>
<th>Test</th>
<th>Tapes</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Jerónimo:</td>
<td>Te</td>
<td>90</td>
<td>92</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Hu</td>
<td>76</td>
<td>78</td>
<td>90, 55</td>
</tr>
<tr>
<td></td>
<td>Mz</td>
<td>58</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Ja</td>
<td>26</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>Jalapa:</td>
<td>Ja</td>
<td>95</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Hu</td>
<td>73</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Te</td>
<td>46</td>
<td>45</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Mz</td>
<td>35</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Tenango:</td>
<td>Tg</td>
<td>95</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Hu</td>
<td>92</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Te</td>
<td>87</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Lo</td>
<td>23</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Ja</td>
<td>21</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 21. Mean, Median, and Modal Scores: Mazatec

B. Range and Standard Deviation:

The two statistics, range and standard deviation, are members of a group of measures of dispersion. They indicate the extent to which scores in a set differ among themselves.

The range of a set of scores is frequently taken to be the difference between the highest and lowest values in the set (Roscoe 1969,45). Sometimes it is calculated as the difference between the high and low score plus one (Downie and Heath 1965,49). The method selected depends somewhat on how the range will be used. If it is used to set up frequency distributions, the latter method is preferable. Since we use it only to describe a set of figures (rather than basing inferences on it), we need only the simpler definition. Besides, the difference between a range of 40% and 41% is certainly not one of accuracy in our test instrument and 40% is a more convenient number than 41%.

The range is an unstable statistic; its size may vary greatly from sample to sample. It also tends to vary in direct proportion to the sample size. In addition, since the range is based only on the two extreme values in the sample, it does not reveal anything about the distribution of the values that fall between the extremes (Freund and Williams 1964,50). Nevertheless, it is a useful descriptive measure in some cases, especially with small samples.

The standard deviation is one of the most important measures of dispersion. It is used both to describe a set of data and to make inferences about the population from which the data were taken. It can be defined as the square root of the sum of the squared deviation scores divided by the number of scores in the sample, or more simply, the root mean square (Roscoe 1969,51). It can be symbolized as follows:

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

where $\sigma$ is the population standard deviation and $x^2$ is a squared deviation score.

A deviation score is the difference between the value of a particular raw score and the mean value of all the raw scores in a set. For example, if raw score $X_1$ is 15 and the mean value of the scores in the set it belongs to is 23, then the deviation is 8. A deviation score is represented by the formula $x = (X - \bar{X})$. Note that a raw score is represented by an upper-case $X$, a deviation score by lower-case $x$. For each set of scores, some deviation scores will be positive, others will be negative, and they will always sum to zero (Downie and Heath 1965,52).

Since we generally do not know the true mean and standard deviation of a population with respect to some characteristic, we usually have to approximate these values with the sample mean and sample standard deviation. The arithmetic mean ($\bar{X}$) is an unbiased
estimate of the population mean (μ) in that it is just as likely to be above the population mean as it is to be below it. Because of this, it is easy to determine the amount of error that one is probably making when he uses the sample mean to substitute for the population mean (Roscoe 1969.39). In short, the statistic, \( \bar{X} = \frac{\Sigma X}{n} \), is directly substitutable for the parameter, \( \mu = \frac{\Sigma X}{N} \).

The standard deviation of a sample, on the other hand, would turn out to be a biased estimate of the population standard deviation if we directly substituted in the formula given above. This is because the standard deviation is the square root of the variance, a statistic we have not used in this monograph. The variance is defined by the formula

\[
\sigma^2 = \frac{\Sigma x^2}{N},
\]

i.e., the arithmetic mean of the squared deviation scores. It has been found that if the sample variance is calculated by dividing \( \Sigma x^2 \) by \( n \) (substituting for \( N \)), that the resulting value on the average underestimates the value of the population variance (Freund and Williams 1964.52). By substituting \( n - 1 \), rather than \( n \), we obtain a value for the sample variance which is an unbiased estimate of the population variance (Roscoe 1969.50). Thus we estimate the sample variance from the formula \( s^2 = \frac{\Sigma x^2}{n-1} \) and the sample standard deviation from

\[
s = \sqrt{\frac{\Sigma x^2}{n-1}}.
\]

The bias is especially significant for small samples, where \( n \) is less than 30, for instance.

The calculation of the standard deviation proceeds as follows: Calculate the arithmetic mean of a set of scores. Then compute the deviation of each of the scores from the mean and sum the deviations. Square each deviation score and add up the total. Divide the sum of the squared deviations by \( n - 1 \) and take the square root of the quotient (Downie and Heath 1965.53-54). The calculations are indicated in Table 22.

<table>
<thead>
<tr>
<th>X</th>
<th>x = (X - ( \bar{X} ))</th>
<th>x²</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Totals: 72 0 28

(\( \bar{X} = 9; \ n = 8; \ n - 1 = 7 \))

\[
s = \sqrt{\frac{\Sigma x^2}{n-1}} = \sqrt{\frac{28}{7}} = \sqrt{4} = 2.
\]

Table 22. Computation of Standard Deviation

The standard deviation may also be calculated by means of the raw score formula which is frequently easier to apply, especially when some of the raw scores, deviation scores, or mean scores are not whole numbers (Freund and Williams 1964.53). The raw score formula is based on the statistic known as the sum of the squares, written as

\[
x^2 = \frac{\Sigma x^2 - (\Sigma x)^2}{N}
\]

1. Standard statistical terminology uses Greek letters to indicate population values, commonly referred to as parameters, and Roman letters to indicate sample values, referred to as statistics.
(Downie and Heath 1965.55), and can be written

\[ s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}} \]

The following steps must be carried out to calculate the standard deviation this way: Put all the raw scores in one column and add them up. Then square the result \((\Sigma X)^2\). Square each raw score and put the corresponding products in a second column. Then sum the products \((\Sigma X^2)\). The values for the sum of the squares, the square of the sum of the raw scores, the number of raw scores in the set \(n\) scores, and \((n - 1)\) are then substituted in the raw score formula. The next step is to subtract the square of the sum from the product of \(n\) and the sum of the squares. The resulting value is then divided by the product of \(n\) and \((n - 1)\). The square root of the quotient will be the sample standard deviation. These calculations are summarized in Table 23.

<table>
<thead>
<tr>
<th>(X)</th>
<th>(X^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>225</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>121</td>
</tr>
<tr>
<td>12</td>
<td>144</td>
</tr>
<tr>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>13</td>
<td>169</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

Totals: 76 (\(\Sigma X\)) 876 (\(\Sigma X^2\))

Table 23. Standard Deviation: Raw Score Formula

Table 24 presents the standard deviations, ranges, and number of standard deviations per range for the test scores of Mazatec subjects described by the averages in Table 21. In all cases, \(n = 10\).

<table>
<thead>
<tr>
<th>Test Tape</th>
<th>Standard Deviation</th>
<th>Range</th>
<th>No. of (s) in range</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Jerónimo:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Te</td>
<td>8.50</td>
<td>25</td>
<td>2.94</td>
</tr>
<tr>
<td>Hu</td>
<td>17.15</td>
<td>50</td>
<td>2.92</td>
</tr>
<tr>
<td>Mz</td>
<td>13.46</td>
<td>55</td>
<td>4.07</td>
</tr>
<tr>
<td>Ja</td>
<td>14.46</td>
<td>40</td>
<td>2.76</td>
</tr>
<tr>
<td>Jalapa:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ja</td>
<td>6.32</td>
<td>20</td>
<td>3.17</td>
</tr>
<tr>
<td>Hu</td>
<td>20.02</td>
<td>60</td>
<td>2.97</td>
</tr>
<tr>
<td>Te</td>
<td>12.46</td>
<td>35</td>
<td>2.81</td>
</tr>
<tr>
<td>Mz</td>
<td>19.23</td>
<td>55</td>
<td>2.86</td>
</tr>
<tr>
<td>Tenango:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tg</td>
<td>6.71</td>
<td>20</td>
<td>2.98</td>
</tr>
<tr>
<td>Hu</td>
<td>7.43</td>
<td>25</td>
<td>3.36</td>
</tr>
<tr>
<td>Te</td>
<td>12.69</td>
<td>40</td>
<td>3.15</td>
</tr>
<tr>
<td>Lo</td>
<td>14.87</td>
<td>50</td>
<td>3.36</td>
</tr>
<tr>
<td>Ja</td>
<td>13.75</td>
<td>50</td>
<td>3.64</td>
</tr>
</tbody>
</table>

Table 24. Ranges and Standard Deviations: Mazatec

The standard deviation is used partly to describe the dispersion, or variability, of a set of scores. In short, if the standard deviation is small, we know that the scores tend to cluster closely around the mean. If it is large we know that the scores tend to be widely scattered throughout their distribution (Freund and Williams 1964.55). For example, compare the standard deviation and ranges of the scores that Jalapa subjects obtained on their hometown test tape and on the Huastica tape (Table 24).
The relationship between the mean score and the distribution of raw scores in terms of deviation from the mean is stated by Chebyshev's Theorem (Freund and Williams 1964, 58). Specifically, for any distribution of scores, the proportion of them that will be within $k$ standard deviations above or below the mean will be at least $1 - 1/k^2$. (The figure 1 stands for absolute certainty. The theorem is really meaningful only when the value of $k$ is more than 1 [Freund and Williams 1964, 177].)

By Chebyshev's Theorem, then, for any distribution, at least 75% of the values will be within the range of two standard deviations of the mean, at least 88.8% will be within three standard deviations of the mean, and at least 96% will fall within five standard deviations. The importance of the term "at least" is evident from Table 24, where we see that, in every case but one, all of the values fall within two standard deviations. This results from the small sample size; as noted earlier, the range of values in a sample increases with the size of the sample. Also note that as it stands, the theorem does not mention a particular kind of distribution. For normal distributions, i.e., "bell-shaped", we can expect at least 95% of the values to fall within two standard deviations of the mean and at least 99% of them to fall within three standard deviations (Freund and Williams 1964, 55). These facts are important to the problem of determining the probable error in a sample value and the estimation of a range of values within which the true population value can be justifiably assumed to lie.

Finally, the standard deviation provides a basis for comparing distinct sets of scores as well as particular members of distinct sets. For example, in Table 24 compare the standard deviations for the three samples on the Hsuilla test tape. Note that standard deviations for the hometown test tapes are generally considerably smaller than those for reference test tapes. Individual scores can be compared by means of a standard $z$ score that statisticians have developed. A subject's $z$ score is found by the formula

$$z = \frac{X - \bar{X}}{s}$$

where, as usual, $X$ is the individual raw score, $\bar{X}$ is the sample mean score, and $s$ is the sample standard deviation.

C. Confidence Intervals:

Since sample mean scores vary somewhat from sample to sample and from test to test, it would be more realistic to state our intelligibility results in terms of range estimates rather than as point estimates only. A range estimate is expressed as a confidence interval of values clustering around a sample mean. The population mean value is estimated to fall somewhere within the interval at a stated level of probability, otherwise known as the confidence level.

The calculation of confidence intervals is based on two main ideas. Suppose we have a population that consists of thirty members with a mean ($\mu$) of 20 and a standard deviation ($\sigma$) of 5. Suppose that we take repeated, random samples of five from this population, calculate the mean ($\bar{X}$) for each sample, and plot the sample means in a frequency distribution. We will find that the frequency (sampling) distribution of the mean values will itself have a mean value of 20 (which is equal to $\mu$) and a standard deviation of

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

The statistic $\sigma_{\bar{X}}$ is called the standard error of the mean (Freund and Williams 1964, 189). It provides a measure of how much sample means can vary merely due to chance. Applying Chebyshev's Theorem to our example, we can say that there is at least a 75% chance that the mean score of a randomly selected sample of five will be within $2\sigma_{\bar{X}}$ of $\mu$, i.e., between 16.14 and 23.86 ($\sigma_{\bar{X}} = 1.93$). We can drop the second element from the above formula since our samples of subjects usually represent such a small proportion of the population that the second factor is very close to 1. For such cases, there is no consequential difference between the formula given earlier and the following (Freund and Williams 1964, 191):

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

As for the second main idea, suppose that we have a sample and that we construct a theoretical sampling distribution from it, just as we did in the first example. Suppose also that we know what the values of $\mu$ and $\sigma$ are. If the sample is large enough, we can approximate the shape of the sampling distribution by a normal curve that has the mean $\mu$ and the standard deviation $\sigma_{\bar{X}}$. This is called the Central Limit Theorem (Freund and Williams 1964, 206).

The Central Limit Theorem is important because it limits us to a kind of distribution about which we can make strong statistical predictions. Briefly, if we view the area under a normal curve as equal to one, then we can equate a given proportion of that area with
a particular probability statement. For example, one half of the total area under the curve corresponds to 50% probability.

A normal curve is completely determined by its mean and standard deviation. Thus we would get different areas under the curve for each of the infinite number of normal curves that has a distinct \( \mu \) and \( \sigma \), if all of the curves were measured along the base in the same units, that is. For economy, statisticians have selected a standard normal curve with the mean, \( \mu = 0 \), and the standard deviation, \( \sigma = 1 \). Measurement along its base is in standard score units (z scores). Areas under this curve are then associated with particular z scores. Areas under any normal curve can be found by converting measurements into z scores (Freund and Williams 1964,200). The standard normal curve is shown in Figure 6.

![Figure 6. The Standard Normal Curve](image)

With respect to the standard curve, we find that about 95% of the area is within 1.96 \( \sigma \) of the mean, 98% is within 2.33 \( \sigma \), and 99% is within 2.58 \( \sigma \). These percentages correspond to the most commonly selected confidence levels. The z scores 1.96, 2.33, and 2.58 are alternative substitutes for a factor in the formula that is used to calculate confidence intervals. For small samples we use values based on the t distribution which is somewhat different in shape from the normal curve (Freund and Williams 1964,444). To set up intervals at the 80, 90, and 95% levels for the data in Table 24, for which there are 9 degrees of freedom, we use the t values 1.383, 1.833 and 2.262 (1964,438, Table 11).

The basic formula is as follows:

\[
\text{Confidence Interval} = \mu \pm z_t \cdot \sigma/\sqrt{n},
\]

where \( z_t \) is the z value that corresponds to the stated confidence level. For small samples, \( z_t \) will be replaced by \( t_t \). The factor \( z_t \cdot \sigma/\sqrt{n} \) is itself a statistic, the maximum error, which can figure in other kinds of computations.

To apply the above formula we must first substitute sample values for the population values which are generally unknown. Thus our formula becomes

\[
\text{Confidence Interval} = \bar{x} \pm t_t \cdot s/\sqrt{n}.
\]

(Note that if the bias in the sample standard deviation has not already been corrected in the formula for \( s \), it will be necessary to replace the factor \( s/\sqrt{n} \) with \( s/\sqrt{n-1} \).) Table 25 presents 80%, 90% and 95% confidence intervals for the data presented in Tables 23 and 24.

At least two assumptions underlie the validity of using small sample confidence intervals to estimate population values of sample mean scores of intelligibility. We must assume first that the samples were randomly selected; second, that each population is approximately normally distributed with regard to how well its members understand the dialects of other speech communities (Freund and Williams 1964,241, 243).

Because of field conditions we have been unable to meet the first assumption; at best our sampling technique is random only in the sense that it is a "happenstance" selection tempered by a few restrictions on subjects. Furthermore, there is a greater or lesser degree of systematic bias since certain groups of people are almost always excluded.

As for the second assumption, it is not clear how justified it is. It may be that the distribution of intelligibility scores is most nearly normal when intelligibility is very high due to a close linguistic relationship and also when it is very low. The distribution may be most skewed when intelligibility is in an intermediate range, say 80% to 85% (based on men's scores) and when it is very high due to interdialectal bilingualism.
However, the data of Table 21 show that the sample distributions tend to be negatively skewed at all levels of intelligibility. In order to test the significance of this tendency I coupled these data with similar data that I had derived from Table 1 (p. 33). For 28 cases, classified as either positively skewed, negatively skewed, or approximately normally distributed, chi square, with a value of 15.55, was significant beyond a level of .01. (At .01 with 2 degrees of freedom, chi square is significant with a value of 9.21 or more.) In short, the second assumption seems doubtful.

We have avoided calculating range estimates because of the problems in our sampling techniques. As a result we have relied entirely upon point estimates. At the risk of offending the professional and orthodox statisticians, I make the unprofessional suggestion that we might do better to state our results in terms of range estimates anyway. I think that total reliance upon a point estimate is more misleading than guessing that the true value probably lies within a given range of values even if I cannot guarantee the strongest statistical probability of it. Furthermore, I think that range estimates can be as easily handled as point estimates with regard to interpreting the data. Admittedly, it is easier to compute mean scores than confidence intervals.

Eventually we may approach the whole problem through Bayesian methods (1964.243). However, there is another class of statistics that rests on different assumptions from those of classical statistics. We briefly discuss three such statistics in the following section.

D. Non-Parametric Statistics:

These statistics are also called distribution-free statistics. This is because they do not require any assumptions about the shape of a population's distribution with respect to the property of interest. Thus we can apply non-parametric methods to our survey data at a smaller risk of violating basic assumptions than we face when we apply parametric methods. In addition, non-parametric statistics are easy to calculate and are especially applicable to samples of ten or less. Finally, they are almost as efficient, mathematically speaking, as parametric measures, particularly when the assumptions underlying parametric methods cannot be met (Downie and Heath 1965.235-236).

In this study we have used three non-parametric measures: the chi square, Wilcoxon T, and Kendall W. The following discussion shows how to calculate each one.

a. Chi Square.

Although chi square has many uses, I have applied it only for significance tests. This statistic can be used to test several differences simultaneously as well as to test the difference between a single pair of values. It is applied to discrete frequency data or to continuous data modeled as discrete categories (1965.160).

To begin with, data are sorted into two or more categories and the number of observations per category is tallied. A set of expected frequencies per category is also computed. All of these values are then entered in a table as follows:

<table>
<thead>
<tr>
<th>Test Tape</th>
<th>X</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Jerónimo:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Te</td>
<td>90</td>
<td>86.3-93.7</td>
<td>85.1-94.9</td>
<td>83.9-96.1</td>
</tr>
<tr>
<td>Hu</td>
<td>76</td>
<td>68.5-83.5</td>
<td>66.1-85.9</td>
<td>63.7-88.3</td>
</tr>
<tr>
<td>Mz</td>
<td>58</td>
<td>52.1-63.9</td>
<td>50.2-65.8</td>
<td>48.4-67.6</td>
</tr>
<tr>
<td>Ja</td>
<td>26</td>
<td>19.7-32.3</td>
<td>17.6-34.4</td>
<td>15.7-36.3</td>
</tr>
<tr>
<td>Jalapa:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ja</td>
<td>95</td>
<td>92.2-97.8</td>
<td>91.3-98.7</td>
<td>90.5-99.5</td>
</tr>
<tr>
<td>Hu</td>
<td>73</td>
<td>64.2-81.8</td>
<td>61.2-84.6</td>
<td>58.7-87.3</td>
</tr>
<tr>
<td>Te</td>
<td>46</td>
<td>40.6-51.4</td>
<td>38.8-53.2</td>
<td>37.1-54.9</td>
</tr>
<tr>
<td>Mz</td>
<td>35</td>
<td>26.6-43.4</td>
<td>24.8-46.2</td>
<td>21.2-48.8</td>
</tr>
<tr>
<td>Tenango:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tg</td>
<td>95</td>
<td>92.1-97.9</td>
<td>91.1-98.9</td>
<td>90.2-99.8</td>
</tr>
<tr>
<td>Hu</td>
<td>92</td>
<td>88.8-95.2</td>
<td>87.7-96.3</td>
<td>86.7-97.3</td>
</tr>
<tr>
<td>Te</td>
<td>87</td>
<td>81.5-92.5</td>
<td>79.6-94.4</td>
<td>77.9-96.1</td>
</tr>
<tr>
<td>Lo</td>
<td>23</td>
<td>16.5-29.5</td>
<td>14.4-31.6</td>
<td>12.4-33.6</td>
</tr>
<tr>
<td>Ja</td>
<td>21</td>
<td>15.0-27.0</td>
<td>13.0-29.0</td>
<td>11.2-30.8</td>
</tr>
</tbody>
</table>

Table 25. Confidence Intervals for Mazatec Data
The observed frequencies for categories A and B are in the left-hand column; the corresponding expected frequencies are in the right-hand one. We obtain the expected frequencies by assuming that if chance alone were operating, the frequencies would be the same for each of the two categories, or one half of the total observed frequencies (1965.161). In the computations, the sum of the expected frequencies always equals the sum of the observed ones.

The value of chi square depends on the magnitude of the difference between each observed frequency (O) and its corresponding expected frequency (E). Chi square is obtained by means of the following formula:

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

For the above example, then,

\[ \chi^2 = \frac{(7 - 17)^2}{17} + \frac{(27 - 17)^2}{17} = \frac{(-10)^2}{17} + \frac{(10)^2}{17} \]

\[ = \frac{100}{17} + \frac{100}{17} = 5.88 + 5.88 = 11.76 \]

The significance of chi square depends on its value, the particular significance level chosen, and the number of degrees of freedom \( \nu \). For a 2 x 1 table, as in the example, the number of degrees of freedom is one. For larger tables, we use the formula:

\[ \nu = (r - 1)(c - 1) \]

where \( \nu = \) degree of freedom, \( r = \) the number of rows in the table, and \( c = \) the number of columns (1965.164). If we chose the significance level of .01, we find from a table that the critical value of chi square, with 1 df, is 6.64 or greater. We therefore conclude that chi square is significant.

It turns out that when the expected frequencies are 10 or less and there is only 1 df, the value of chi square is usually an overestimate. To handle this bias, statisticians use Yates’s correction for continuity. Chi square is then found by means of the formula:

\[ \chi^2 = \sum \frac{(O - E - .5)^2}{E} \]

By Yates’s correction, chi square in our example becomes 10.62. This is still significant at .01. Roscoe recommends the use of Yates’s correction for all cases where \( \nu = 1 \) (1969.192).

The calculation of chi square for 2 x 2 tables, as well as for larger ones, is somewhat more complicated. After deciding how many categories of observations to include, how to classify the observations (yes, no, and undecided, for example), and tallying the observations for the right cells of the table, it is necessary to construct a corresponding table of observed frequencies. This is done for each cell \( x_{ij} \) by multiplying the sum of the observation in column \( j \) by the sum of the observations in row \( i \) and dividing the product by the total number of observations in the sample. Table 26 illustrates these steps with data from the Choapan study (see p. 94).

In Table 26, note that the marginal values are (approximately) equivalent for both parts of the table and that the sum of the expected frequencies (in the lower right-hand corner) equals the sum of the observed frequencies. The expected frequency 24.2 equals the product of 101 and 33 divided by 138. The other expected frequencies were derived in a similar fashion. In this case there are five df: \( (6 - 1)(2 - 1) = 5 \).

Chi square is computed according to the formula given first. (Yates’s correction is not used when there are two or more degrees of freedom.) There are twelve terms in the summation, most of which are omitted from the abbreviated calculations which follow:

\[ \chi^2 = \frac{(29 - 24.2)^2}{24.2} + \frac{(15 - 24.2)^2}{24.2} + \ldots + \frac{(2 - 3.5)^2}{3.5} \]

\[ = \frac{29.04}{24.2} + \frac{94.64}{24.2} + \ldots + \frac{2.26}{3.5} = 18.669; \]

2. The "degree of freedom" refers to the number of observed values of a variable minus the number of certain mathematical restrictions on them. These restrictions turn out to be statistics that must be calculated prior to computing the particular statistics that figures in the test of an hypothesis (Roscoe 1969.161-162).
therefore chi square is significant at a level of .01 (critical value = 15.086).

In addition, note that in three cases an expected frequency is 3.5. I could have combined rows 3, 5, and 6 into one row in order to get a larger expected frequency. This

<table>
<thead>
<tr>
<th>Pair of Tests</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diff.</td>
<td>Same</td>
</tr>
<tr>
<td>Ch and Co</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>Ch and Ja</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Ch and Xo</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Ja and Co</td>
<td>25</td>
<td>8</td>
</tr>
<tr>
<td>Ja and Xo</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Co and Xo</td>
<td>11</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 26. 6 x 2 Contingency Table: Choapan Sentence Tests

step is recommended when any E in a large table falls below 5 (Downie and Heath 1965.170). Downie and Heath recommend further individual tests to determine precisely where the significant differences occur in the event that chi square in a large table turns out to be significant (1965.170).

b. The Wilcoxon T.

The Wilcoxon T is a test that is applicable to correlated data such as two sets of scores for the same subjects (Downie and Heath 1965.133). Since the Choapan test scores and the Mazatec reliability test scores (see Section 6.3.3) are correlated data, I have used this statistic to analyze them. It is one of the best non-parametric alternatives to the student t test (Roache 1969.174). However, it assumes that the property being measured is continuously distributed (1969.183).

To calculate the Wilcoxon T, we first construct a table with columns in which we enter both the raw data and the results of subsequent operations on them. The individual values for each pair of scores are entered in the corresponding cells of columns X and Y.

The second step consists in finding the difference between each pair of scores. Sometimes this difference will be positive; other times it will be negative; and occasionally there will be no difference. The corresponding sign of the difference is entered, along with its value, in the proper cell of the third column (column D).

Then, ignoring the signs of these differences, the absolute values of the differences between each pair of scores are ranked and the value of each rank is entered in the appropriate cell of the fourth column. The smallest difference found in the third column is given the highest rank (1), which is entered in the corresponding cell of the fourth column (Abs. Rank Difference). Succeeding ranks are then assigned ordinally to the increasing values of the differences. Whenever a difference of 0 occurs, it is not entered in the fourth column and the total number of ranks is reduced by one. If two pairs of scores yield the same difference, they are given the same rank. The value of this rank is determined by dividing the sum of the ordinal rank positions that would be occupied by these differences (say, fourth and fifth position) by the number of differences that share a particular value. Thus, if fourth and fifth positions were shared by two pairs having a difference of 2, they would be assigned the rank of 4.5.

The ranked values of pairs showing positive differences (the value of X is higher than that of Y) are entered in the column R(+). Ranked values of pairs showing negative differences (X is less than Y) are entered in the column R(-). The calculation is completed by summing the values in these columns. The smaller sum is taken as Wilcoxon's T statistic.

For interpreting the significance of T, it is helpful to compute the mean sum of the values of the ranks assigned to each difference. This is obtained by the formula

\[ T = \frac{N(N + 1)}{4}, \]

where N is equal to the number of pairs of scores yielding differences, and \( \bar{T} \) is the mean sum of the ranks. If \( T \) were equal to \( \bar{T} \) there would be no significant difference between the two groups of data, obviously. However, since there generally is a difference, the question is, is this difference significant? Since the mean scores of the other tapes may be either above the mean for the first tape, or below the mean, the test will be two-sided, i.e., it will be necessary to specify a limit of acceptability on either side of the mean, outside of which the difference between \( T \) and \( \bar{T} \) is considered significant. A diagram of the test is shown below:
The limit of acceptability is set at 5%. This means that if the difference between T and \( \bar{T} \) exceeds a certain numerical limit, there is a 95% probability that the difference is significant. Finally, it should be noted that the absolute values of T are symmetrical around the axis of the line, increasing from 0 at either end of the line to the value of T at the axis. (T can come from either of the two ranks.) Thus a table of critical values for the Wilcoxon T statistic reads that, for a set of ten ranked pairs of numbers, an absolute value of 8 or less for T constitutes a significant difference from the mean sum of the ranks \( \bar{T} \), which is 26. Suppose, then, that for such a set of ranked pairs, the value for T was 12, as in diagram (2) above. According to the diagram, this difference, i.e., 28 - 12 = 16, is not significant. The two sets of data, therefore, probably come from the same population.

Note that since the T statistic is always the smaller of the rank sums, its absolute value is always smaller than the mean of the rank sums \( \bar{T} \). Its sign will be either positive or negative depending on which of the two rank sums it is derived from.

Table 27 shows the computation of T for non-hometown subject scores on the Xochiapen test tapes (see Section 5.5).

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>D</th>
<th>Abs. rank Difference</th>
<th>R(+)</th>
<th>R(-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>10</td>
<td>8.9</td>
<td>1.1</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>2.</td>
<td>9</td>
<td>6.7</td>
<td>2.3</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>3.</td>
<td>10</td>
<td>8.9</td>
<td>1.1</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>4.</td>
<td>7</td>
<td>7.8</td>
<td>-0.8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5.</td>
<td>7</td>
<td>5.6</td>
<td>1.4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6.</td>
<td>8</td>
<td>5.6</td>
<td>2.4</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>7.</td>
<td>10</td>
<td>7.8</td>
<td>2.2</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>8.</td>
<td>8</td>
<td>8.9</td>
<td>-0.9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9.</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>11.</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>13.</td>
<td>9</td>
<td>10</td>
<td>-1</td>
<td>3.5</td>
<td>3.5</td>
</tr>
</tbody>
</table>

\( T = 6.5, \bar{T} = 35.5, N = 11 \). With \( N = 11, \bar{T} = 35.5, T \) is significant with a value of 11 or less at a level of .05. Therefore we conclude that subjects scored significantly lower on the Xochiapen sentence test than they did on the text test.

Table 27. Wilcoxon T for Xochiapen Test Scores

Note that there are two tie scores between X and Y values so that the number of ranks in the calculations is reduced to eleven. There are also two pairs of tied differences between X and Y values. Roscoe considers the first kind of tie an argument for retaining the null hypothesis and accordingly suggests assigning all such ties to the rank with the smaller sum. He also considers this kind of tie to be counterevidence to the assumption that the measured property is continuously distributed. The argument that such ties may result from a crude measurement of the property he finds unsatisfactory (1989,184).

However, since the tie scores come from subjects tested in the areas closest to Xochiapen and since the probability that a subject will get the same score on both tests increases with the proximity (both geographical and linguistic) of the test dialect to the one he speaks, we should not be surprised to find this kind of tie. Moreover, our test
consisted of only ten items. The range of scores was four for the text test and less then six for the sentence test. Since the sample size was ten and evaluations were made only in terms of whole and half units, we are almost assured of obtaining the second kind of tie.

c. The Kendall W.

The Kendall W is a correlation coefficient that measures the overall relationship among three or more sets of ranked numbers (Downie and Heath 1965:209). With this statistic we can measure for each sample of subjects the extent to which their individual scores on a set of test tapes tend to have a profile that corresponds to the profile of sample mean scores on the test tapes (see Section 2.3.2). We have already used this statistic to measure the degree to which Trique subjects agreed in making the most errors on the initial five questions of the sentence tests, somewhat fewer errors in the second five, and the fewest errors on the last five questions.

<table>
<thead>
<tr>
<th>Tapes</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10</td>
</tr>
<tr>
<td>Te</td>
<td>80 100 85 80 95 95 90 95 75 100</td>
</tr>
<tr>
<td>Hu</td>
<td>50 70 55 95 100 80 75 90 55 90</td>
</tr>
<tr>
<td>Mz</td>
<td>25 80 55 60 60 55 60 70 50 60</td>
</tr>
<tr>
<td>Ja</td>
<td>35 50 10 25 25 30 10 20 20 50</td>
</tr>
</tbody>
</table>

Table 28. Raw Score Matrix A

The calculation begins by converting sets of raw scores into rank scores. Suppose that we have a raw score matrix A in which a set of test tapes is listed down the left side, the subjects are listed along the top, and each subject's test scores are entered in the cells of one column. Table 28 presents one such matrix.

We derive a ranked score matrix B by ranking from highest to lowest the scores in each column of A. The highest-ranked score is assigned the value 1, the next highest the value 2, and so on. Tie scores are assigned the same rank value. This value is the mean of the values of the rank positions they normally would have occupied had they not been ties. For example, both scores of 55 in the third column of A (Table 28) receive the rank value 2.5, which is the mean of rank positions two and three. The rank scores are then entered in matrix B (Table 29), which has the same number of rows and columns as A.

Matrix B is entered into a table that has three additional columns: one for the sum of the ranks in each row of B (column X), one for the difference between the sum of the ranks and the mean sum of the ranks (D), and one for the squared differences between the two values (D^2). This display is presented in Table 29.

<table>
<thead>
<tr>
<th>Tapes</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 X D D^2</td>
</tr>
<tr>
<td>Te</td>
<td>1 1 1 2 2 1 1 1 1 1 12 13 169</td>
</tr>
<tr>
<td>Hu</td>
<td>2 3 2.5 1 1 2 2 2 2 2 19.5 5.5 30.25</td>
</tr>
<tr>
<td>Mz</td>
<td>4 2 2.5 3 3 3 3 3 3 3 29.5 4.5 20.25</td>
</tr>
<tr>
<td>Ja</td>
<td>3 4 4 4 4 4 4 4 4 4 39 14 196</td>
</tr>
</tbody>
</table>

Σ = 100 Σ: = 415.50

Table 29. Ranked Score Matrix B

The values in X are obtained by adding up the rank values in each row of B. They themselves are summed to give the Total sum of the ranks, in this case, 100. The total sum of the ranks is divided by the number of rows in B to yield the mean sum of the ranks, i.e., 25. (This value is not indicated in the table.) The values in D are found by subtracting each value in X from 25, the mean sum of the ranks. The squares of the D values are calculated and entered in the final column. The D^2 values are then added up.
The Kendall $W$ is computed from the formula

$$W = \frac{12 \sum D^2}{m^2 (N)(N^2 - 1)},$$

where $m$ is the number of judges (subjects) and $N$ is the number of projects (test tapes). For our example, the calculations are as follows:

$$\frac{12(415.50)}{100(4)(15)} = \frac{4986}{6000} = .831.$$ 

From a table of critical values we find that $W$ is significant with a value of .35 or more at a level of .01 when $m = 10$ and $N = 4$. Therefore we conclude that subjects' relative performances on the test tapes were highly correlated with one another.

The Kendall $W$ ranges in size from 0 to 1. A coefficient of 0 means that the subject responses were completely disparate or that each subject obtained the same score on all the tests he took. (One person does not necessarily have to score just the same as some other person, however.) A coefficient of 1 generally means that all of the subjects obtained the same relative scores over the set of test tapes.

The measure was computed for twenty-nine samples of Mazatec scores. (In some cases it was computed for several sub-samples of the original one since the entire test set had not been given to all ten subjects.) In all but one case, $W$ was significant at a level of .01. It ranged in size from .390 to .842. In sixteen cases it was larger than .75; in ten cases it was smaller than .60. Generally, when $W$ is large, sample mean scores are widely scattered; when it is small, most of the mean scores cluster together closely.

It turns out that we are unable to arrive at reliability coefficients for our tests by such standard methods as test-retest and split-halves. Usually we are unable to carry out a field experiment a second time. Furthermore, our tests are probably a kind of speed test; at least this is true for hometown tests (Downie and Heath 1965.219). The Kendall $W$ seems a likely substitute. It is easily applied when we have only one available sample. As a coefficient of correlation it behaves like those used to measure reliability. That is, it is affected by the range of scores in the sample; the more homogeneous the sample, the smaller the coefficient. Finally, its size varies from sample to sample (1965.221). It would probably be wise to report, along with the Kendall $W$, the number of subjects in the sample and the sample mean scores of all the tests in the test set.

d. Critical Values of Chi Square, $T$, and $W$.

The following tables give the critical values of chi square, $T$, and $W$ that are most likely to be needed (these values are taken from Tables IV, IX, and XIII of Downie and Heath 1965).

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.20</td>
</tr>
<tr>
<td>1</td>
<td>1.642</td>
</tr>
<tr>
<td>2</td>
<td>3.219</td>
</tr>
<tr>
<td>3</td>
<td>4.642</td>
</tr>
<tr>
<td>4</td>
<td>5.989</td>
</tr>
<tr>
<td>5</td>
<td>7.289</td>
</tr>
<tr>
<td>10</td>
<td>13.442</td>
</tr>
</tbody>
</table>

Table 30. Some Critical Values of Chi Square.
### Significance Level (Two-Tailed Test)

<table>
<thead>
<tr>
<th>N</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>30</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>52</td>
<td>43</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 31. Some Critical Values of the Wilcoxon T

### Significance

<table>
<thead>
<tr>
<th>m</th>
<th>.10</th>
<th>.05</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>.44</td>
<td>.58</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>.36</td>
<td>.42</td>
<td>.56</td>
</tr>
<tr>
<td></td>
<td>.32</td>
<td>.37</td>
<td>.49</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.35</td>
<td>.45</td>
</tr>
<tr>
<td></td>
<td>.29</td>
<td>.33</td>
<td>.42</td>
</tr>
<tr>
<td>9</td>
<td>.31</td>
<td>.35</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td>.23</td>
<td>.28</td>
<td>.38</td>
</tr>
<tr>
<td></td>
<td>.21</td>
<td>.26</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td>.24</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>.19</td>
<td>.23</td>
<td>.29</td>
</tr>
<tr>
<td>10</td>
<td>.25</td>
<td>.31</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td>.21</td>
<td>.25</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>.19</td>
<td>.23</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>.18</td>
<td>.21</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>.17</td>
<td>.20</td>
<td>.26</td>
</tr>
</tbody>
</table>

Table 32. Some Critical Values of the Kendall W
Appendix I

Significance Testing and Kirk's Reliability Study

Kirk's reliability study has already been mentioned briefly (Section 6.3.3). Kirk elicited a pair of texts from each of two informants, speakers of the Huastle dialect of Mazatec. These texts differed in length and content. They were then used to construct a set of intelligibility tests which were administered individually to ten subjects in the town of Nuevo Soyaltepec. The test tapes are designated Hu, Hu, Hu, and Hu; Hu is the standard Huastle test tape that was administered at all test points in the Mazatec survey. Hu, 2:55 minutes long, and Hu, 2:20 minutes long, were given by one informant. Hu and Hu, given by the other informant, were 1:55 and 2:10 minutes long, respectively. The motivation for the study was to determine whether the differences between the texts would significantly affect mean scores. The mean scores were 74.5%, 72.5%, 75.0%, and 77.0% for the respective test tapes. Kirk suggested from these data that the differences between the means were not significant and that such factors as differential length of the story, content, and interest value for the subjects do not greatly affect a subject's scores (Kirk 1970:211).

Kirk's suggestion is testable statistically. The discussion that follows gives a brief summary of certain aspects of significance testing and presents the results of the test of the Nuevo Soyaltepec data that I made with the Wilcoxon T statistic.

A statistical test of significance is based on four elements: (1) the statement of a hypothesis and an alternative to that hypothesis, (2) the choice of a significance level, (3) the choice of a particular statistic for testing the hypothesis against the alternative, and (4) the decision as to whether to accept the hypothesis in case it cannot be rejected, or to withhold judgment (Freund and Williams 1964:258-260).

The hypothesis is generally known as the "null" hypothesis, and it is stated in a form meaning "there is no difference between an observed characteristic of A and the corresponding observed characteristic of B." Almost always, this statement of no difference is known to be false before the test is applied. It is obvious that there is a difference, but it is not known whether this difference is merely due to random variation or whether to the two samples having been taken from different populations. It might be said that the experiment is designed to disprove the null hypothesis, thus indicating that there is a significant difference between the two observed characteristics (Fisher 1956:1512-1521). For this reason, the alternative hypothesis is stated so that the rejection of the null hypothesis entails the acceptance of the alternative (Freund and Williams 1964:258).

Occasionally, however, it may be the intention of the experiment to allow the null hypothesis to stand. The statement of the alternative is either one-sided or two-sided. For example, the null hypothesis might be \( \mu = 18 \), which means "the true mean value for both samples is 18 even though the observed values are 15 and 20." The alternative hypothesis might be \( \mu \neq 18 \), a two-sided alternative because it means that \( \mu \) is either greater or less than 18. A one-sided alternative might be \( \mu > 18 \), which means that \( \mu \) is 18 or more; values less than 18 are not relevant to the test. The terms "one-tailed test" and "two-tailed test" are applied to tests in which the alternative hypothesis is correspondingly one-sided or two-sided. The actual statement of an alternative depends on the nature of the problem which is not always intuitively obvious.

The choice of a significance level is related to the possible outcomes of the experiment. These are that the null hypothesis may be accepted or rejected, and in either case the action taken may be correct or incorrect. The desired actions are that a true null hypothesis should be accepted and a false one rejected. There are also two possible incorrect actions: to reject the null hypothesis when it is actually true, known as a Type I error, and to accept a null hypothesis when it is actually false, known as a Type II error. An example of a Type I error would be to say that there is a significant difference between two mean scores on a test tape when the difference is really not significant. A Type II error would consist in saying that there is no significant difference between the two mean scores when there actually is.

The significance level is the statement of the degree of probability that either a Type I or a Type II error is being committed. The choice of this level depends upon the penalties that are attached to committing either kind of error. It is made on the basis of prior knowledge about particular aspects of the problem.

For example, a Type I error in the dialect survey program would mean that an extra literacy team would unnecessarily be assigned to a multilingual area. The penalty would be that of overloading one area for literacy work while some other area would be
neglected. On the other hand, a Type II error would mean that some dialect group would be overlooked within that multilingual area. In both cases there will be inadequate coverage of some multilingual area since there is not an unlimited supply of literacy teams available. The main difference is that with a Type I error a literacy team’s effort would be largely wasted by working in an area where it was not really needed. With a Type II error at least the team would eventually be assigned to some distinct dialectal area. It seems, then, that the penalties attached to making a Type I error in this aspect of the survey program are slightly higher than those attached to a Type II error.

Background information collected before and during the preliminary survey trip should be helpful in estimating the probability that any two dialect areas are identical or distinct. If this information strongly suggests that the two areas are the same, then the significance level for a Type I error should be set quite low. If the information suggests that the areas are distinct, then the level for committing a Type I error might be set somewhat higher with the result that the probability of committing a Type II error is diminished (Wilson 1952:59). With these considerations in mind, the significance level should be set before the experiment begins. Normally-chosen levels are .10, .05, and .01, though there is actually variation from .40 to .0001, depending upon the nature of the experiment (1952:60). The chosen level is maintained for that particular experiment, once it has been selected. If it is changed, it is changed for a future replication of the experiment as a result of information gained from the initial one.

\[ \mu = 25 \quad \bar{X} = 26.2 \]

**Figure 7. Significance Test with Parametric Measure**

Many significance tests are made using parametric statistics. This class of statistics requires the assumption that the population which is being sampled is normally distributed with respect to a particular characteristic. This normal distribution can be approximated by a bell-shaped curve, with the area under the curve considered to be one. Then if a section of the curve is cut off by two lines parallel to the axis of the curve, the area between the two lines is some proportion of the total area. This proportion can then be equated with a particular probability. For example, in Figure 7 two lines cut off the tails of a normal curve. The shaded area represents 5% of the total area under the curve; the unshaded area, therefore, represents 95% of the total area. The diagram represents a significance test at a .05 significance level for a test in which the true value \( \mu \) is hypothesized as 25 and the observed value is 26.2. Calculations with a sample size of 16 and a standard deviation of 3.6 show that the difference is not significant since it does not exceed the limits of rejection at a .05 significance level. This test leaves a 5% probability for committing a Type I error. Type II errors are more difficult to calculate, but the same type of reasoning is basically followed.

It is also possible to perform significance tests with non-parametric measures. A major difference between the two is that the former class of statistics assumes a normal distribution, whereas the latter does not. Referring to Figure 7, it is correct to say that of a sample of mean scores, there were probably many with values of almost 25, but there were few with a value of about 27.5. However, this statement is inapplicable to a non-parametric measure which makes no reference to the distribution of a given characteristic of a population. Furthermore, non-parametric measures are more applicable to small samples, \( N \leq 10 \). Therefore, this kind of measure was used for testing the Nuevo Soyaltepec data. Since the data are test scores for the same sample of subjects on four different test tapes, they can be considered correlated data. Because the Wilcoxon T statistic is applicable to correlated data and since the occurrence of tie scores has a minimal effect on the computation, I chose it for testing the significance of the differences between the mean scores.
Finally, with regard to the null hypothesis, the decision must be made whether to accept it, or withhold judgment if it is not possible to reject it. It is common to withhold judgment in this case, since by making no decision one does not risk committing a Type II error (Freund and Williams 1964:260). Further testing is necessary if one does not want to accept the null hypothesis and is required to make some decision. If a two-tailed test is used (as in Figure 7), it is wise to consider the region around the rejection limits as a "no-decision" zone (Wilson 1952:80).

For the test with the Mazatec data, the decision was to accept the null hypothesis if it could not be rejected. The significance level chosen was .05. Prior knowledge included a test by Bradley in which he obtained greatly varying scores with two different test tapes from the same test point. Kirk previously had tested two tapes from Huautla in San Miguel, in which the difference between mean scores was slight. The conflicting results are explained by realizing that two different test designs were involved, the Mazatec test being a more reliable measure than the Mixtec. The hypothesis under test was that the tests came from the same population. The alternative was that they came from different populations. Thus the test is two-sided.

The Wilcoxon T statistic was computed for each of four pairs of tapes from Huautla: 1 and 2, 1 and 3, 1 and 4, and 2 and 4. The last pair represents the highest and lowest values of the sample means.

Table 33 summarizes the Wilcoxon test for these pairs of test tapes. Data are indicated for T, T (the mean of the rank sums), critical value of T, decision regarding significance, the size of the sample of ranked scores, and the number of tie scores.

<table>
<thead>
<tr>
<th></th>
<th>Hu₁ &amp; Hu₂</th>
<th>Hu₁ &amp; Hu₃</th>
<th>Hu₁ &amp; Hu₄</th>
<th>Hu₂ &amp; Hu₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>19.5</td>
<td>16.5</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>T</td>
<td>22.5</td>
<td>18</td>
<td>27.5</td>
<td>18</td>
</tr>
<tr>
<td>Critical value</td>
<td>6 or less</td>
<td>4 or less</td>
<td>8 or less</td>
<td>4 or less</td>
</tr>
<tr>
<td>Significant</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N</td>
<td>9</td>
<td>8</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Tiea</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 33: Summary of Wilcoxon T Test of Significance

In short, the T tests support Kirk's conclusion that the mean scores did not differ from one another significantly. This conclusion is corroborated by both an analysis of variance and the calculation of Kendall's W.
Appendix J

A General Theory of Intelligibility

Although it is premature to formulate a mathematical model of cross-language communication, we have learned enough from our own field experience in collecting intelligibility data and from examining studies of bilingualism that we can state a preliminary verbal theory of some of the important variables and their relationships. Unfortunately, the theory we present here will be too broad to test directly. Nevertheless, there are three reasons for submitting the theory as it now stands: (1) to allow a semi-formal restatement of it so that its inadequacies can be discovered, thus leading to refinements in the theory, (2) to present a rationale for the suggestions made earlier in this paper to collect certain kinds of sociological data, and (3) to stimulate the development of ways to measure the variables. As we begin to measure the variables that are included in the theory, we can carry out tests of their relationships. The results of these tests will undoubtedly lead to further (probably extensive) revisions in the theory.

This treatment leans heavily on the approach outlined by Blalock (1968a; 1968b; 1969) that we mentioned briefly in Section 4.3. We pointed out that it is based on a general theory that consists of an inventory of postulated variables that are related to one another in specific ways. The general theory is linked to an auxiliary theory that adapts it to a particular population, a given set of measuring instruments, and a specific research design. The rest of Section 4.3 discussed several studies partly to provide an empirical basis in terms of which we can begin to apply Blalock’s method.


Blalock’s general causal theory attempts to provide a theory broad enough to explain numerous occurrences of some social phenomenon, each of which may occur in a distinct context (see Jackson and Curtis 1968:144-45). A major desideratum of the theory is that it lead to specific predictions which can be tested, thus allowing its rejection (Blalock 1968b:155). A major problem in stating the theory is that the real world is extremely complex; many variables are relevant to the problem. If we include too many in our theory, it becomes too complex to test. If we include too few, our theory is rendered incapable of explaining many of the phenomena we want to understand (Blalock 1969:6-7). Kemeny suggests that the real problem is that the kind of mathematics we need to adequately handle complicated systems has not yet been developed (Kemeny 1969:26).

A generalized causal theory applied to intelligibility would proceed through the following steps: (1) the specification of the full set of theoretically interesting variables, (2) the postulation of relationships among these variables, (3) the description of primary networks of dependencies that fit (1) and (2), and finally (4), the statement of simplifying assumptions and what they relate to. The meaning attached to an index of maximum extendability in this framework is that the concept represents the maximum domain within a multilingual area in which a speaker of a given dialect can make himself understood in the vernacular. (This meaning is adopted rather than one which might view dialect extendability in terms of the potential area within which a single orthography can be used, for example.)

A set of ten variables is presumed to underlie the dependent variable, intelligibility. Five are taken to be completely independent; five are dependent and intervene between the independent variables and intelligibility itself. The independent variables are (1) history of intragroup relations, (2) socioeconomic relations, (3) alternatives for language use, (4) relative size of the groups, and (5) degree of linguistic similarity between a pair of dialects. The dependent variables are (1) nature of intragroup contact, (2) societal attitudes, (3) language attitudes, (4) type of bilingualism, and (5) degree of bilingualism.

In some cases, the label I have used to refer to a particular variable actually refers to a complex of phenomena. For example, “socio-economic relations” covers the whole sphere of social, political, and economic affiliations. In the theory, these are expressed in terms of dominances, dependencies, and rivalries. “Nature of contact” refers to the intensity, purpose, and primary location of interaction between members of distinct communities. “Alternatives for language use” refers to the possible use of a national language, trade language, or pidgin in addition to another vernacular dialect. “Societal attitudes” refers to a group’s attitudes toward itself as well as toward another group. “Language attitudes” likewise is based on the distinction between own-group and other-group. Both societal and language attitudes are expressed in terms of the strength of positive affect, negative affect, or neutrality. “Degree of bilingualism” means the proportion
Figure 8. Variables Underlying Intelligibility
of the population of group A that customarily uses a form of speech recognizably distinct from its own dialect to communicate with members of group B. "Type of bilingualism" refers to language use when people from groups A and B talk together. Do they communicate in the national language? In a trade language? In the dialect of A? Or the dialect of B? Finally, "cross-language communication" refers to the general ability of speakers from A to understand speakers from B regardless of the genetic relationship between language A and B. This formulation reflects my view that intelligibility is simply a special case of bilingualism.

The relationship among these variables can be described as a network of causal chains which link independent variables to intelligibility by paths of varying degrees of indirectness. This is shown in Figure 8. Each variable is identified by a label. Specific causal relations are indicated by the arrows linking the boxes, e.g., "socioeconomic dependencies \( \rightarrow \) intensity of contact" means that an increase in the strength of socioeconomic dependencies leads to an increase in the degree of contact between two communities on some level of society. This can be symbolized as \( X \rightarrow Z \). There may be more than one cause of a particular variable, i.e., \( X \rightarrow Z \) and \( Y \rightarrow Z \). There are causal chains of the form \( X \rightarrow Y \rightarrow Z \), which signal indirect relationships that are increasingly tenuous, e.g., "relative size" \( \rightarrow \) "bilingualism" \( \rightarrow \) "intelligibility." Another relationship is reciprocal causation, i.e., \( X \rightarrow Y \rightarrow X \), which can be symbolized as a double-headed arrow linking a pair of boxes. There may also be a case where a variable acts both indirectly on a dependent variable through an intervening variable and directly on the dependent variable, i.e., \( X \rightarrow Y \rightarrow Z \) and \( X \rightarrow Z \). Finally, it is possible for one independent variable to influence directly more than one dependent variable, i.e., \( X \rightarrow Y \) and \( X \rightarrow Z \).

The relationships displayed in Figure 8 can be organized into sets of primary and secondary dependencies. Primary dependencies hold between those variables that are linked directly; secondary dependencies result from including an intervening variable in the linkage between a pair of variables. Ignoring the secondary dependencies, then, there are several sets of primary dependencies: first, societal attitudinal attitudes are determined by socioeconomic relations and the history of intragroup relations. Second, language attitudes are determined by societal attitudes. Third, degree of bilingualism results from the nature of the contacts between two groups and from the relative size of the groups. Fourth, type of bilingualism results from societal attitudes, language attitudes, the alternatives available for language use, and the degree of linguistic similarity between two dialects of the vernacular. Finally, the degree of cross-language communication (intelligibility) is a result of the degree of linguistic similarity between the dialects of the vernacular that are spoken by the two communities, the type of bilingualism, and the degree of bilingualism.

There are a number of simplifying assumptions which must be made in order to make the testing of the propositions of this theory feasible. The first is that the theoretical system is closed; that is, all the variables that determine intelligibility are related within the system. The second is that the populations are also closed; that is, the characteristics of neither population are being affected by migration between the two groups. The third is that the theoretical system adequately approximates reality (Blalock 1966b:192). Finally, the system is assumed to be static; otherwise the second assumption would not be justified.

An additional set of assumptions relates the variables explicitly included in the system. The first is that all the independent variables specified by the general theory really are independent of one another; i.e., no two of them have a common cause outside the system. The second is that all variables, independent and dependent, that are directly linked to the same dependent variable are independent of one another; (Wiggins 1968:392).

For example, we assume that present patterns of socioeconomic relations can change independently of patterns seen in the past history of the groups. The third assumption is that all the variation in any dependent variable is accounted for by the behavior of the variables which are directly linked to it (Siegel and Hodge 1968:28). For simplicity's sake, we begin by assuming one-way causation between each pair of variables (as opposed to reciprocal causation). We also assume that the variation in a dependent variable is accounted for in terms of an additive relationship among the variables that are its direct cause. (The relationship could well be multipliciative, for example.) In addition, we assume linearity between all the pairs \( X \) and \( Y \) such that \( X \rightarrow Y \). Finally, we assume that the correlation between the end points of an indirect linkage is zero. Thus for two causal chains \( X \rightarrow Y \rightarrow Z \) and \( X \rightarrow W \rightarrow Z \), the correlation between \( X \) and \( Z \) is zero if we control for both \( Y \) and \( W \) simultaneously (Blalock 1966b:168-69).

There is the possibility of some kind of effect on \( Z \) which occurs because \( X \) and \( W \) are interacting together (1968:180). This effect does not occur when only one of the two variables is influencing \( Z \) (1969:156). An interaction that involves \( X \) and \( W \), each linked to \( Z \) through separate intervening variables, is a first-order interaction. Two first-order interactions can themselves interact to produce a second-order interaction (1969:163). Although we do not yet know what they are, it is probable that there are significant interactions among some of the variables of Figure 8. We assume, however, that there are
no significant second-order (or higher) interactions (1968b: 183, 185; 1969: 163). In looking for first-order interactions, then, we begin by assuming that all variables directly linked to the same variable do not interact. However, we still should test all possible pairs of variables for first-order interactions. If we find some that the theory does not predict, we should probably look for an additional variable that intervenes between one of the two interacting variables and the dependent one (1969: 164).

We need these assumptions because the theory is complex, and yet we would like to be able to formulate testable hypotheses from it. The large number of variables and the indirect relationships between important ones allow for a large number of indeterminate predictions along with the specific ones we would like to make. This is to be expected of systems with more than five or six variables. We must confine ourselves to oversimplified versions of reality which represent isolated subsystems, each composed of a small set of variables assumed to be related to one another in certain ways (1968b: 158-59).

By means of the above assumptions, we break up an initial set of eleven variables into five independent subsystems, each of which corresponds to one of the primary dependency sets and none of which contains more than five variables. We can now state propositions both about pairs of directly linked variables and pairs of indirectly linked ones. Propositions about direct linkages can be viewed as the axioms of the system; those about indirect linkages are the theorems. We will also try to state the direction of the covariations for the directly linked variables; that is, whether B increases or decreases with a corresponding increase in A.

Before we can test these propositions, however, we have to obtain adequate measures of indicators of the variables (1969: 18-19). For most of the variables we will have to begin with a crude measure such as a dichotomy. Thus, for “type of bilingualism” we might use the values 1 and 2, where 1 means “bilingual with another dialect of the vernacular” and 2 means “bilingual with a genetically unrelated language.” Attitudes may be evaluated as (+1), positive affect; (0), neutrality; and (-1), negative affect. Dependencies can be described as high (1) versus low (0); dominance as strong (1) versus weak (0); and rivalries as intense (1) versus weak (0). History may be measured as bad (-1), good (+1), or neutral (0). Both the degree of contact and degree of bilingualism can be treated in terms of intense (1) versus none (0), or on an even finer scale. Alternatives for language use could be scaled from 1 to n, depending on the number of possible choices. The location of contact might also be measured as (1) and (0), with (1) meaning “contact in home town” and (0) being “contact in another town.” Finally, the proportion of the population involved in the contact might be scaled as (2), large; (1), small; and (0), none.

The following propositions are the axioms of our theory. This does not imply, however, that since they are axioms they are inherently untenable (1969: 11). For any pair of language communities, either of which can be variously the “hometown group” or the “outside group,” the following propositions are postulated:

1. Good historical relations result in positive affect toward both the hometown group and the outside group.
2. Poor historical relations result in positive affect toward the hometown group and negative affect toward the outside group (Dozier 1964).
3. Socioeconomic dominances lead to positive affect toward the hometown group and negative affect toward the outside group.
4. Socioeconomic dependencies lead to neutral affect toward the hometown group and positive affect toward the outside group.
5. Socioeconomic rivalries lead to positive affect toward the hometown group and negative affect toward the outside group.
6. Positive affect toward the hometown group leads to negative affect toward the outside group.
7. Negative affect toward the hometown group leads to positive affect toward the outside group.
8. Positive affect toward the hometown group leads to positive affect toward its own language.
9. Positive affect toward the outside group leads to positive affect toward the outside language.
10. Negative affect toward the hometown group leads to negative affect toward its language.
11. Negative affect toward the outside group leads to negative affect toward the outside language.
12. Positive affect toward the hometown group tends to make the outside group bilingual with the hometown group’s language.
13. Negative affect toward the hometown group tends to make the hometown group bilingual with the outside language.
14. Positive affect toward the hometown language tends to make the outside group bilingual in the hometown group’s language.
15. Negative affect toward the hometown language tends to make the hometown group bilingual in the outside group's language.
   (All the above axioms can be stated in the form "the more intense X affect, the more likely result."
16. The more similar the hometown and outside dialects, the more likely there will be bilingualism with one of these.
17. The more dissimilar the hometown and outside dialects, the more likely there will be bilingualism with a genetically unrelated language.
18. The more alternative languages there are available, the less likely it is that bilingualism will involve the vernacular.
19. The more socially and economically dominant the hometown group, the more likely it is that people will come from outside areas to the hometown area.
20. The more dependencies there are between two groups, the greater the proportion of people from one group that will be brought into extensive contact with people from the other one.
21. The more intense the rivalries between groups, the less contact there will be between them.
22. The more extensive the contact between two groups, the higher the incidence of bilingualism.
23. The smaller one group is relative to another, the higher the incidence of bilingualism in the smaller group (Weinreich 1957.211).
24. The higher the incidence of bilingualism, the more effective the cross-language communication.
25. The more types of bilingualism there are, the more likely intelligibility will be low between vernacular dialects.
26. The more similar the vernacular dialects, the higher the intelligibility between them.
   Rather than attempt to list even a significant fraction of the full set of theorems derivable from the above set of axioms, I will mention a few of the more obvious and important possibilities that the theory must handle. These are stated as the conditions which we can expect to exist if communication between dialects is to be carried out via the vernacular and, for these conditions, which dialect most likely will be the preferred one.

1. The more self-sufficient a vernacular group is socially and economically (the fewer external socioeconomic dependencies), the more likely bilingualism will be between dialects of the vernacular rather than between a dialect of the vernacular and some external language.
2. The extent of interdialect bilingualism is inversely proportional to the availability of a trade language or national language.
3. If two dialects are quite similar, but there is extreme resentment between the speech communities, an alternative language may be adopted by the more dependent group for communicating with members of the dominant group (cf. Solenberger 1962).
4. If the two dialects are similar enough, and if social attitudes are favorable, people from one community will communicate in the vernacular with people from the other.
5. If a pair of dialects is different enough, it is likely that communication will be in the vernacular only if there is no alternative language available.
6. If there is an alternative language, and if the vernacular dialects are very distinct, the vernacular will be used in exchange between members of the two speech groups only if the groups form a highly self-sufficient and independent unit. In such a case, the total unit is likely to have a fairly large population.
7. In the case of 4, 5, or 6, the dialect selected for communication will be the one spoken by the group that is dominant economically, politically, and socially.

b. Auxiliary Theories.

The general theory can be tested only indirectly. This will be done by constructing and testing auxiliary theories that adapt it to specific research designs, particular populations under study, and one set of measuring instruments. The auxiliary theory allows one to state specific, refutable hypotheses. After testing these hypotheses, we will have to reformulate the general theory. It may turn out, for example, that some of the variables in Figure 9 are so closely related that they all can be taken as indicators of a single underlying variable; thus we can simplify the theory. On the other hand, the tests may reveal significant interactions that we did not expect and force us to make the theory more complex either by adding more variables to the system or by postulating non-additive relationships among the interacting ones. Finally, the results of these tests will enable us to eliminate inadequate theories but not establish any one theory once and for all (Blalock 1969.152). However, with respect to our practical goals, it will probably be sufficient simply to eliminate enough unlikely theories that we are left with one that is highly plausible. To
Figure 9. Auxiliary and Main Theories: Trique
achieve exact mathematical solutions may be too costly, if not impossible (Kemeny 1969:27).

An auxiliary theory consists of an inventory of indicators (measured variables) that are linked to corresponding theoretical variables of the general (or main) theory. We have to assume the nature of these linkages. The auxiliary theory must also include assumptions about omitted variables that cause measurement errors as well as specific variables that cause systematic bias in the measurement process. In most cases, the auxiliary theory will be linked to only a few of the theoretical variables of the main theory (Blalock 1968a:23-27; 1969:4-5, 151-53).

In the rest of this section I will try to illustrate what an auxiliary theory applied to the Trique data (Section 4.3.2) might look like. Because of the preliminary state of the art, the treatment is partly based on fictional values that correspond to measured values and correlations. Figure 9 presents the combination of main and auxiliary theories for the Trique situation (Section 4.3.2). The dotted line separates the main theory (the upper set of boxes) from the auxiliary theory. Each measured variable is assumed to be linked directly to its theoretical counterpart. The arrows that point to variables in the auxiliary theory, but that do not have labelled points of origin, represent variables that cause error in the measurement process. We assume that these measurement errors are both random and uncorrelated. The sixth variable in the auxiliary theory represents systematic sampling error due to the measuring instrument as well as to the segment of the population sampled.

We assume that the level of measured intelligibility between a pair of dialects is an additive function of weighted measures of (1) linguistic similarity, (2) intensity of contact, (3) location of contact, (4) bilingualism with Spanish, and (5) systematic sampling error. We can represent this by the formula

\[ Y' = aX_1 + bX_2 + cX_3 + dX_4 + e, \]

where \( Y' \) represents the predicted level of intelligibility, the lower case letters represent the correlation between an \( X_i \) and \( Y' \) (intelligibility), and \( e \) represents measurement error. In order to obtain the value of \( Y' \), therefore, we need to have values for each measured variable and the dependent one (cf. Blalock 1969:61).

The auxiliary theory applies to the five test points that were included in the study: San Andrés Chicahualtica (Ch), Laguna (La), San Martín (Ma), Sabana (Sa), and San Miguel (Mi). Sabana and San Miguel are both part of the Copala dialect (Co). The intelligibility scores are taken from Table 10 (p. 80). The measure of linguistic similarity is based on the 250-word list whose results are given in Table 9 (p. 79). La and Ch are identical; so are Sa and Mi. These pairs are thus given values of 100% similarity.

I rated intensity of contact as between 0 (none at all) and 1 (very extensive). I assigned the value 1 only to members of identical dialects. This means, for example, that these Copala barrios at least have more contact with some other Copala barrio than they do with an outside dialect. I state it this way since people from Sabana and San Miguel rarely cross paths. For each Trique dialect I ranked the relative frequencies with which people from there interact with people from each of the other dialects. This left me with twenty pairs "from A to B" which I ranked and grouped somewhat impressionistically. Although most of the ranking was derived from the facts given in Section 4.3.2, I had to estimate the frequencies for San Miguel. Next I combined each "A to B" with its corresponding "B to A" and averaged the rank values. (For instance, if "A to B" was 6 and "B to A" was 4, then A-B was 5.) This resulted in a symmetrical measure of intensity of contact. I made some adjustments so that local variations would show up. For example, I knew that Laguna subjects had more connections with Copala speakers than San Andrés subjects did, so I rated the Chicahualtica dialects differently with respect to intensity of contact with Copala. In addition, since the primary contact between Chicahualtica and Copala involved the Copala barrios of Tierra Blanca and Tiapa rather than either Sabana or San Miguel, I rated the Copala dialects lower on intensity of contact with Chicahualtica than I had rated the Chicahualtica dialects on Copala. Finally, in a few cases I gave the average rank a higher or lower value to make the overall ranking conform more closely to what I already knew about the area.

The measure of location of contact allowed me to build in the asymmetrical relationships that exist in the Trique area. If all the interaction took place in the hometown area, I rated location as 0. On the other hand, if all of it took place in an outside area (some other town), I rated it as 1. If almost all contact was in the other town, I rated location as .9. Reciprocal location of contact is shown by a value of .5; .7 means that most contact is in another town, but a fair amount also takes place in the hometown.

I used a single factor to represent bilingualism with Spanish. First I set up a scale of .01 to .10 such that .01 would represent complete bilingualism and .10 no bilingualism at all. (This has the effect of contributing a larger value to \( Y' \) when there is a lower incidence of bilingualism.) Since few Triques speak Spanish well, I simply assigned the value .09 to all of the dialects. This assumes uniformity in the incidence of bilingualism.
which is certainly not the case (Weinreich 1957:207). For example, the highest incidence of bilingualism probably occurs in Chicahuaxtla; it is somewhat lower in San Martin; and lowest in San Miguel and Sabana (Bruce Hollenbach, personal communication). As it is, the differences in the incidence of bilingualism may not be enough to worry about. This variable could probably be dropped from the auxiliary theory without harm. However, it could be a very important variable for an area much larger and more heterogeneous than the Trique, such as the Mixtec.

As for error terms, I basically followed Hollenbach’s off-the-cuff remark that he felt that the results from San Martin were “95% correct.” I also took into account the data presented in Table 16 (p. 115) and whether the test scores were for hometape or reference test tapes. I felt that it was reasonable to assume smaller errors on hometape, larger errors on reference tapes. Thus I used error terms of .02 and .05.

The coefficients for the variables in the equation are completely subjective; they were selected specifically to yield predicted values that would correspond fairly closely to the measured values of intelligibility. I felt that linguistic similarity was by far the most important variable, that intensity and location of contact were of equal importance, and that measurement error affected the scores more than bilingualism with Spanish. Thus I weighted them as follows: linguistic similarity, .75; intensity of contact, .20; location of contact, .20; bilingualism, .10; and error, 1.0.

Our prediction equation thus assumes the form

\[ Y' = .75X_1 + .20X_2 + .20X_3 + .10X_4 + e. \]

Applied to the San Martin relationship to Copala, it yields

\[ Y' = .75(.70) + .20(.70) + .20(.90) + .10(.09) + .05 = .535 + .140 + .180 + .009 + .050 = .914 = 91.4\% . \]

The measured value \( Y \) was actually 92%.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ch: Ma</td>
<td>77</td>
<td>.9</td>
<td>.1</td>
<td>.09</td>
<td>.05</td>
<td>83</td>
<td>84</td>
</tr>
<tr>
<td>Co</td>
<td>65</td>
<td>.5</td>
<td>.5</td>
<td>.09</td>
<td>.05</td>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td>2. La: Ch</td>
<td>100</td>
<td>1.0</td>
<td>0</td>
<td>.09</td>
<td>.02</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>Ma</td>
<td>77</td>
<td>.9</td>
<td>.1</td>
<td>.09</td>
<td>.05</td>
<td>86</td>
<td>84</td>
</tr>
<tr>
<td>Co</td>
<td>65</td>
<td>.8</td>
<td>.5</td>
<td>.09</td>
<td>.05</td>
<td>83</td>
<td>81</td>
</tr>
<tr>
<td>3. Ma: Ch</td>
<td>77</td>
<td>.9</td>
<td>.9</td>
<td>.09</td>
<td>.02</td>
<td>98</td>
<td>97</td>
</tr>
<tr>
<td>Co</td>
<td>70</td>
<td>.7</td>
<td>.9</td>
<td>.09</td>
<td>.05</td>
<td>92</td>
<td>91</td>
</tr>
<tr>
<td>4. Sa: Ma</td>
<td>70</td>
<td>.4</td>
<td>.1</td>
<td>.09</td>
<td>.05</td>
<td>64</td>
<td>69</td>
</tr>
<tr>
<td>Ch</td>
<td>65</td>
<td>.2</td>
<td>.2</td>
<td>.09</td>
<td>.05</td>
<td>57</td>
<td>63</td>
</tr>
<tr>
<td>5. Mi: Co</td>
<td>100</td>
<td>1.0</td>
<td>0</td>
<td>.09</td>
<td>.02</td>
<td>99</td>
<td>98</td>
</tr>
<tr>
<td>Ch</td>
<td>65</td>
<td>.1</td>
<td>.1</td>
<td>.09</td>
<td>.05</td>
<td>58</td>
<td>59</td>
</tr>
<tr>
<td>Ma</td>
<td>70</td>
<td>.0</td>
<td>0</td>
<td>.09</td>
<td>.05</td>
<td>56</td>
<td>59</td>
</tr>
</tbody>
</table>

Table 34. Estimated Variables and Predicted Intelligibility for Trique Data (see Figure 9)

(Variables (1) and (6) are the most adequately measured ones. The measurements of (2) and (3) are reasonable, although the values for San Miguel are actually an educated guess. \( Y' \) was obtained with an equation that is based partly on the values in this table and partly on certain constants whose values are purely subjective.)

Both “measured” and predicted values for the Trique data are presented in Table 34. The agreement is actually more than I would have expected to find had I obtained correlation coefficients by objective methods. I think we can label my hocus-pocus approach an exploratory fit of measured variables to one possible mathematical formula (that probably contains a host of unjustifiable assumptions). As I mentioned earlier, the purpose is only to show what an auxiliary theory might look like. What I want to emphasize is that there are ways to measure the variables I have postulated for the general theory. If we are careful about developing ways to measure them and consistently collect data by means of such measures, then we can calculate correlation coefficients,
state theorems about the relationships among variables, test these theorems, and revise our general theory on the basis of these tests.

A practical result of this process is finding an easier way to determine dialect extendibility. For example, if I had reliable measures of the right set of variables for each of several dialects, a sound empirical evaluation of the correlation between each one and intelligibility, a plausible notion of their intercorrelations (whether completely independent or only slightly correlated), and an appropriate mathematical model (involving additive relationships, multiplicative ones, or both), then I could predict the level of intelligibility between each pair of dialects. If, on actually measuring intelligibility, I found substantial agreement between the two sets of values, then I might stop testing for intelligibility altogether and simply rely on data from the other sets of measuring instruments.

We close this discussion with a few comments about tests of hypotheses. There is a common view that the goal of science is not to accept or to prove hypotheses, but rather it is to disprove them (Wiggins 1968.390). For one thing, even if a given hypothesis cannot be disproved, there are generally enough unknowns in a problem that one can always formulate additional alternative hypotheses that imply the same predictions (Siegel and Hodge 1966.57; Bialock 1966b.158). This view permits an objective evaluation of the internal or empirical validity of a hypothesis; by increasing the number of alternative hypotheses that are disproved, one also increases the degree of internal validity of a given hypothesis (Wiggins 1968.390).

A hypothesis is disproved if the systematic variation that the theory assumes between an independent and a dependent variable is contradicted by observation. Wiggins points out two conditions that must be met in order for a test of a hypothesis to be valid (1968.390). The first is that the range of variation in the independent variable must be very wide, because variations in the dependent variable may be observable only near the high or low limits of the independent variable. For example, the effect of interdialectal learning on population mean scores shows up only if a large enough proportion of the speakers of one dialect communicate often enough in the vernacular with speakers of the other. Likewise, intelligence differences will be reflected only in the test scores of extremely bright or extremely dull subjects. The second condition is that there must be controls in the design to handle sleeper effects. For example, subjects' performances on the test tapes may improve as they become better acquainted with the method and as they gain more confidence in the investigator. This could mean that the first set of intelligibility scores obtained at a test point would be low, not because of negative social conditions between the local people and those from the reference point areas, but because of lack of education on the part of the subjects as well as a general mistrust of non-vernacular-speaking people.

The primary reason that we have not yet been able to isolate the effects of some variables on intelligibility is that experiments thus far have not taken into account the high and low values of independent variables that are required to produce measurable changes in the dependent variables. For example, Kirk's study with four Huautla texts showed that change of speaker, different content, and different length of story did not significantly affect a subject's understanding of Huautla speech. This does not mean, however, that differences in these factors could not produce significant differences in intelligibility, if other factors are present. If these characteristics were built into a similar test. As to the second condition, sleeper effects were observed in earlier studies and controls for them were incorporated into later tests (Section 3.4.2). To supplement Kirk's study we can now cite a clear example that showed up recently in tests made during the Aztec study in the state of Puebla. Because subjects had been scoring much higher on a test from San Antonio Alpanocan than the investigator thought they should have, the survey team formulated a second set of ten questions on that text and collected another text from which they constructed a second test tape. Then they went to Santa Maria Coapan and administered the two tests to twelve subjects. Six of these were tested on the version of the first test that used the original set of questions and six on the version that included the second set. All twelve were tested on the second test tape as well. The subjects averaged 92% on the original version and 95% on the version with the second set of questions. However, all twelve subjects averaged only 66% on the test based on the second text. The investigator concluded that the difference in subject matter (a hunting story, whose content was basically predictable, [see p. 11, paragraph (b)]) versus a story about an event involving some soldiers, whose content was not predictable) was the primary reason for the substantial difference in scores. In addition the narrator of the hunting story spoke slowly and separated the successive sentences in the text by noticeable pauses. The narrator of the second text spoke in a more animated style with fewer pauses. Thus the first text was also easier for the subject to follow with understanding (David Perea, personal communication). Thus, because strict controls were not maintained on either the content of the texts or the narrator's style of delivery (or both) a substantial difference in test scores was observed for this pair of tests.
BIBLIOGRAPHY

Abbreviations

AA    American Anthropologist
AL    Anthropological Linguistics
ES    Explorations in Sociolinguistics, edited by Stanley Lieberson
HMAI  Handbook of Middle American Indians, edited by Manning Nash
IJAL  International Journal of American Linguistics
LCS   Language in Culture and Society: A Reader in Linguistics and Anthropology, edited by Dell Hymes
MP    Measurement and Prediction, edited by Samuel A. Stouffer et al.
MSR   Methodology in Social Research, edited by Hubert M. and Ann B. Blalock
PA    Practical Anthropology
SJA   Southwestern Journal of Anthropology

Adams, Richard N.

Beals, Ralph L.

Bellman, Richard E. and Dreyfus, Stuart E.

Bender, Marvin L. and Cooper, Robert L.

Biggs, Bruce

Blalock, Hubert M., Jr.

Blalock, Hubert M., Jr. and Blalock, Ann B., eds.

Bradley, C. Henry

Brown, Roger and Gilman, Albert
Bull, William E.

Casad, Eugene H.

Chomsky, Noam and Halle, Morris

Clerk, Lawrence

Colby, Benjamin N.
1961  "Indian Attitudes Towards Education and Inter-Ethnic Contact in Mexico." PA 8.77-85

Crawford, John C.
1967  "Theoretical and Methodological Beginnings of Dialect Survey Program." Lecture presented to Dialect Survey Field Workers' Conference, Cuernavaca, Morelos, Mexico.

Diebold, A. Richard, Jr.
1962  "A Laboratory for Language Contact." AL 4.41-51.

Downie, N. M. and Heath, R. W.

Dozier, Edward P.

Dundes, Alan

Egland, Steven Thomas
1972  "Informe Sobre La Alfabetizacion Tzeltal." Ms.

Ferguson, Charles A.

Fisher, Sir Ronald A.

Fishman, Joshua A.
1964  "Language Maintenance and Language Shift as a Field of Inquiry." Linguistics 9.32-70

Freund, John E. and Williams, Frank J.

Goode, William J. and Hatt, Paul K.
Grimes, Joseph E.
"Finding Networks of Communication." JIAL, in press.

Grimes, Joseph E. and Agard, Frederick B.

Grimes, Joseph E. and Glock, Naomi

Gudschinsky, Sarah C.

Haugen, Einar

Hickerson, H.; Turner, Glen D.; and Hickerson, Nancy P.

Hymes, Dell, ed.

Jackson, Elton F. and Curtis, Richard F.
1968 "Conceptualization and Measurement in the Study of Social Stratification." In: MSR. pp. 112-49.

Kemeny, John G.

King, Arden R.

Kirk, Paul L.
1967 "Recommendations for Mazatec Allocations Based on Dialect Survey." Ms.

Kroeber, A. L.

Ladefoged, Peter

Ladefoged, Peter; Glick, Ruth; and Criper, Clive

Leal, Otis and Robinson, Dow
Lieberson, Stanley, ed.  
Publication Forty-Four, Indiana University Research Center in Anthropology, Folklore, and Linguistics.  

Mckaughan, Howard  

Naroll, Raoul  

Nash, Manning, ed.  

Newman, James R., ed.  

Olmsted, David L.  

Pierce, Joe E.  
1954 "Crow versus Hidatsa in Dialect Distance and in Glottochronology." IJAL 20.134-36.

Pike, Eunice V.  

Pike, Kenneth L.  

Roscov, John T.  

Rubin, Joen  

Samarin, William J.  

Saporta, Sol, ed.  

Sebeok, Thomas, ed.  

Shankara Bhat, D. N.  

Siegel, Paul M. and Hodge, Robert H.  

Solenberger, Robert R.  

Stoltzfus, Ronald D.  
Bibliography

Stubblefield, Morris and Stoltzfus, Ronald D.

UNESCO

Upshaw, Harry S.
1968 "Attitude measurement." In: MSR. pp. 60-111

Voegelin, C. F. and Harris, Zellig S.

Weber, Rose Marie

Weinreich, Uriel

Wiggins, James A.

Wilson, E. Bright, Jr.

Wolff, Hans

Wurm, S. A.

Yamagiwa, Joseph K.

ADDENDA

References to Works on Bilingualism
The following list presents a number of key items that I have not used in this paper but that should be taken into account by those seriously interested in pursuing the study of bilingualism with the view of answering questions such as those posed in Section 8.1.

Bright, William, ed.

Capell, A.

Ferguson, Charles A. and Gumperz, John

Fishman, Joshua
1968 Readings in the Sociology of Language. The Hague: Mouton. (This volume contains reprints of most of the studies discussed in Chapter 8).
Fishman, Joshua; Cooper, Robert L.; and Ma, Roxanne

Fishman, Joshua; Ferguson, Charles A; and Gupta, J. Das

Gumperz, John, ed.

Haugen, Einar

Kelly, L. G.

Labov, William

Mackey, W. F.

Macnamara, John, ed.

Rubin, Joan

Vildomec, Verocej

References to Guttman Scalogram Analysis

The following references include both technical discussions of the Guttman method and a few of its applications to problems in anthropology.

Aberle, David F.

Edwards, Allan L.

Festinger, Leon
1947 "The Treatment of Qualitative Data by 'Scale Analysis.'" Psychological Bulletin 44.149-61.

Goode, William J. and Hatt, Paul K.

Goodenough, Ward H.
Guilford, J. P. 

Guttman, Louis 

Hage, Per 

Kay, Paul 

Schaeffer, Harold W. 

Stouffer, Samuel A. and others, eds. 

Torgerson, W. S. 

Upshaw, Harry S. 

Additional Studies of Intelligibility

The following items treat studies that I have not taken into account in the text. I have largely ignored the role that specific linguistic features play in determining intelligibility. We should relate this area to our main interest in some way, however.

Bansal, R. K. 

Masica, Collin P. 

McQuown, Norman A. 

Sarles, Harvey B. 

Measures of Linguistic Similarity

Here are more papers that relate to Chapter 7.

Capell, A. 

Ladefoged, Peter 