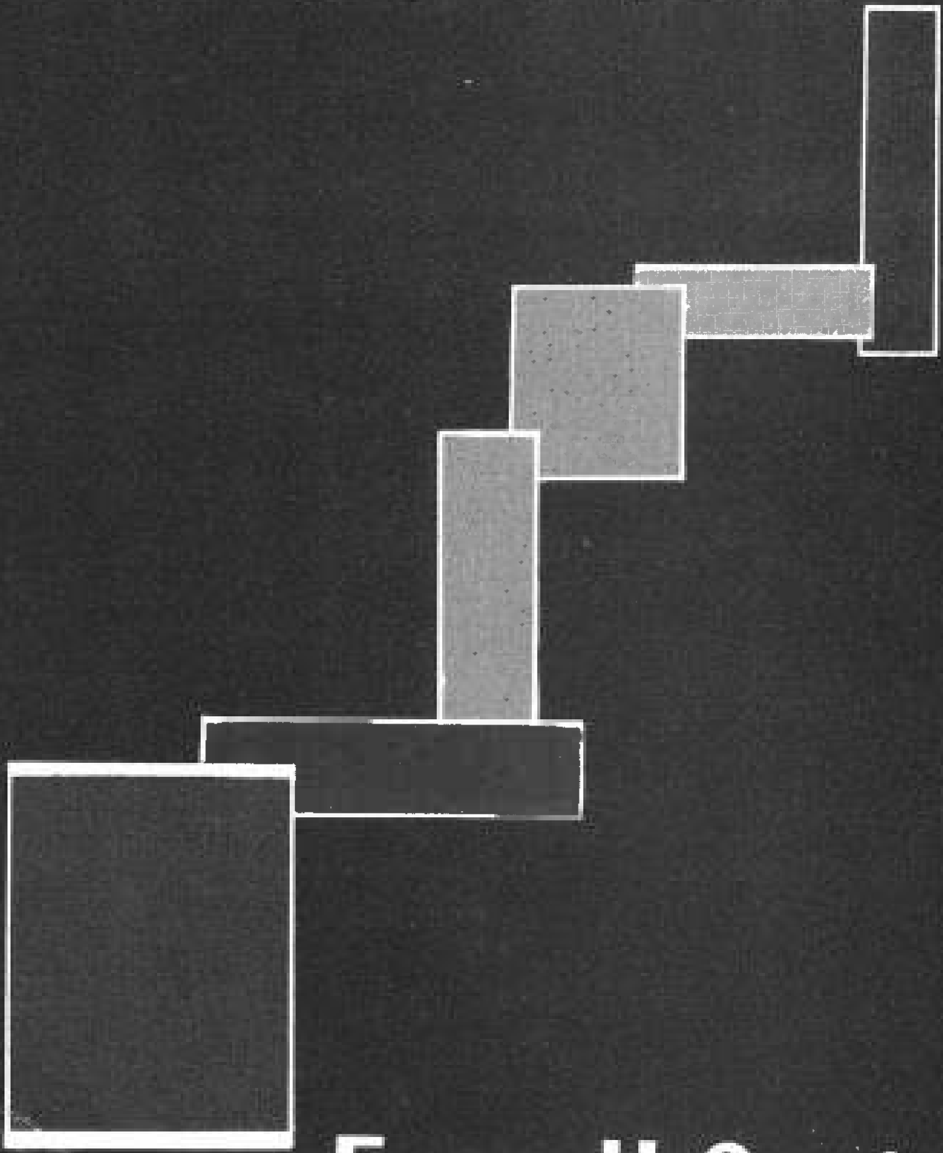
WINDOWS ON BILINGUALISM



Eugene H. Casad

[blank]

Windows on Bilingualism

Summer Institute of Linguistics and
The University of Texas at Arlington
Publications in Linguistics

Publication 110

Windows on Bilingualism

Eugene H. Casad
Editor

A Publication of
The Summer Institute of Linguistics
and
The University of Texas at Arlington
1992

Library of Congress Catalog No: 92–81102

ISBN: 0–88312–809–8

ISSN: 1040–0850

Cover design by Hazel Shorey

Copies of this and other publications of the Summer Institute of Linguistics may be obtained from

International Academic Bookstore
7500 W. Camp Wisdom Road
Dallas, TX 75236

To

Hans Wolff

who first realized the full import of the role of sociolinguistic
factors in determining nonreciprocal intelligibility and
severed its link to measures of dialect distance

and to

John Crawford

who, on the basis of ample field experience, and with much insight,
recognized the validity of Wolff's conclusions and adapted
intelligibility tests as the core of a multifaceted sociolinguistic
index of dialect extendability

[blank]

Contents

Part I: Setting

Part II: Word Lists and Intelligibility

Part III: Windows on Bilingualism

Part IV: Technical Review

Part V: Postscript

Foreword

A picture of the world as a finite number of speech communities, each speaking its own language, is an appealing picture of the world's language situation. For the linguistic theorist and language typologist there would be a finite—though presumably large—number of languages whose characteristics could be checked against hypothesized general principles or parametric settings. For the proponent of vernacular literacy as an empowering mechanism for human communities, one could simply check whether each community had an acceptable writing system for its language, and, if not, proceed to language description, orthography creation, and literacy training. But of course we all know that this picture is wrong in many respects.

Many communities are multilingual: they use several dialects of the same language or several different languages in their everyday communication. Also, we know that vernacular literacy is often resisted or, even if tentatively accepted, fails to take hold. Furthermore, linguists have no adequate professional consensus as to what constitutes "a language" as opposed to dialect variation and no satisfactory measure of just how different any two languages are. And the fundamental linguistic questions remain unanswered of how individual competence relates to community conventionalization and how this relationship changes over time. Finally, speakers' attitudes toward language and toward particular variant forms of language varieties are widely acknowledged to be involved in questions of mutual intelligibility, dialect variation, and language change. Nonetheless no one is yet clear on just how language attitudes affect language structure, language use, and language change, and even the best methods of collecting valid data on language attitudes are not yet understood.

Fortunately, linguists of the Summer Institute of Linguistics (SIL) not only recognize all these theoretical and practical problems in understanding the nature of human language and the difficulty of predicting language change, but they are also steadily accumulating experience in dealing head-on with these issues. General linguists must be grateful for the commitment of SIL to make Scriptures available to everyone in his or her own language, for this means the SIL linguists must constantly search for answers to such basic questions as these: how different must two languages be in structure, use, and attitudes before there is a need for translation into both instead of just one? What information must be collected in order to reach a reasonable decision? How can a language survey be carried out effectively with a minimum of time, effort, and funds? What measures are most useful in testing for mutual intelligibility of two dialects? How does one determine the incidence of competence in different codes throughout a community and the relevant attitudes of members of the community toward them?

The papers in this volume report on SIL experience with measures of linguistic distance, mutual intelligibility, and language attitudes as well as the necessary information about principles of statistics to deal with such measures. The language situations and the research efforts reported on are quite varied. Ethnographic methods of participant observation are discussed along with analysis of responses from self-report language-competence questionnaires. For linguists unfamiliar with field testing of this kind, it is impressive to see comparisons of cognate counting, sentence repetition testing, intelligibility testing, and the SLOPE approach to proficiency testing. Reports deal with the problems of differentiating between "inherent intelligibility" and the intelligibility that comes from exposure to another language (i.e., partial bilingualism).

Most of the papers included were presented several years ago at an SIL conference on the Pacific area, so they refer more often to language situations in the Pacific than elsewhere, but all of them have implications for cases of societal multilingualism anywhere in the world. In this sense, the papers do indeed offer "windows on bilingualism" and though many of them seem overly concerned with the practical techniques of measurement and with speculations about how to predict the success of vernacular literacy in various language situations, they all serve to remind those interested in linguistic structure of the complexities of human language and the enormous difficulties for predicting changes in patterns of language structure, language use, and language attitudes.

Charles A. Ferguson
Stanford University

Acknowledgments

The present volume of papers is the offspring of a conference in Baguio City, Philippines, sponsored by colleagues of the Asia Area Group of the Summer Institute of Linguistics (SIL) who have been engaged in various language assessment tasks for the last dozen years or more. Much of this work has been carried out under the leadership of Calvin R. Rensch, who brought to this enterprise a solid academic background with a Ph.D. in Linguistics from the University of Pennsylvania and the publication of a landmark volume in Amerindian comparative linguistics titled *Comparative Otomanguean Phonology.* His field experience includes completion of the translation of the Lalana Chinantec New Testament. Thus, Cal is in a special position both to train personnel for language survey related endeavors and for evaluating the propriety of various test instruments.

Several other people were also crucial in seeing this project off the ground and bringing it to a close. Eugene Loos has been a consistent source of encouragement ever since he and Cal asked me to edit the Baguio papers in order to pass on the benefits of the survey conference to the entire SIL field membership. Most importantly, there could be no volume if there were no authors to write the papers. I extend my warmest thanks to each one of them: Barbara F. Grimes, Joseph E. Grimes, John Stephen Quakenbush, Calvin R. Rensch, Randy Kamp, and Dale Savage. They have been both cooperative and prompt and this editor appreciates very much all their contributions.

For various reasons, not all of the papers given at the Baguio Conference are reprinted here. In addition, several papers not presented at the Conference are included because they reinforce ideas presented at the conference and treat issues that must be handled by our organization. They also help keep the content of this volume current. I am very grateful

to Eugene Loos and Joseph E. Grimes for permission to reprint the article *Correlations between Vocabulary Similarity and Intelligibility* from *Notes on Linguistics No. 41.* The author, Joe Grimes, has been in the middle of the development of methods for doing language surveys for twenty years, looking for and testing computer applications to field data. Joe has also been actively involved in analyzing and processing such data throughout that entire period.

I would also like to express my gratitude to Ray Harlow, editor of the *Proceedings of the Fifth International Congress on Austronesian Linguistics, January 1988,* and to the Linguistic Society of New Zealand for their permission to reprint *Surveying Language Proficiency* by Steve Quakenbush.

Finally, this volume has changed form somewhat, partly due to the delays I encountered in editing it. For one, Cal Rensch wrote a report for the 1990 Biennial Conference of SIL, which he has graciously permitted me to incorporate into this collection of papers (See Part V). At about the same time, Joe Grimes showed me his paper on *Calibrating the Sentence Repetition Test* and also agreed to have it published here. And, at my invitation, Dale Savage wrote the two papers included in Part IV. The delays in readying this volume, therefore, occurred in part because of the addition of these articles to keep the collection of papers as current as possible. Be that as it may, I accept full responsibility for the shortcomings of this volume and hope that, in spite of them, it will be a useful one.

Part I: Setting

The Asia Area Survey Conference

Eugene H. Casad

The Asia Area Conference of the Summer Institute of Linguistics (SIL) on Survey Data Collecting and Interpreting was held at the Green Valley Country Club of Baguio City, Luzon, the Philippines on October 12–13, 1987. This conference was of primary importance both for the Asia Area's ongoing research and for SIL's goals in the area of language surveys. In addition, it was a broadening of my own exposure to the range of situations under which surveys are being carried out, as well as an opportunity to interact with a number of my Asia Area colleagues for the first time. I was not only able to contribute somewhat to the conference, but was asked to edit a subset of the papers given at this conference. All parties concerned with various aspects of language assessment could thus share the results of several years of intensive survey work in the Asia area, done under the capable guidance of Calvin Rensch and Carolyn Rensch.

The first day of the conference was devoted to the topic of bilingualism. Randy Kamp began with a paper titled *Bilingualism Testing in the Philippines* (in this volume titled *Inherent Intelligibility, Bilingualism, or Both?*) in which he detailed a comparison of four test instruments used during a survey of the Karao people of Benquet province, Luzon Island.

The test instruments were: a proficiency interview, a self-score evaluation, a self-test questionnaire, and a set of taped comprehension tests. The self-score test required subjects to rate their own ability in the languages that they spoke. This rating was in terms of a five-point scale. The self-test questionnaire consisted of a set of questions on the ways the subject claimed to be able to use a given language. These questions are grouped around a set of various levels of difficulty that are assumed to attach to particular tasks. The proficiency interview followed the procedures in the *Second Language Oral Proficiency Evaluation* (SLOPE) syllabus, a work by Barbara Grimes (1987c) and others to adapt for SIL's use an Educational Testing Service approach to evaluating bilingual proficiency. The taped comprehension tests included a HOMETOWN test tape in Karao and two Ibaloi test tapes.

Kamp's conclusions are very significant. For one, he finds that he cannot always separate bilingualism from inherent intelligibility. He also observes statistically significant correlations between self-score evaluations and interview scores, and between the self-test questionnaire and the interviews. The strength and reliability of these correlations, however, is such as to require separate tests for intelligibility and bilingualism. Finally, he concludes that narrative tests for comprehension probably test no higher than an FSI level 3. This latter conclusion, however, cannot be generalized beyond the Karao survey—cf. the results of the recent study by James, Masland and Rand (1989).

In *Surveying Language Proficiency*, Steve Quakenbush describes two methods he used in a survey among the Agutaynen people of Palawan, Philippines. One was a self-reporting questionnaire which required subjects to give YES or NO responses to questions designed to sample language skills of varying difficulty. The other was an oral proficiency interview modeled on the approach of the Foreign Service Institute and the Educational Testing Service. Quackenbush found a fifty-six percent correlation between the two methods, which compares favorably with the findings of other researchers, including Kamp.

Quackenbush appropriately concludes that bilingual evaluation is not an exact science and that proficiency data must be construed as an approximation of the true state of affairs. The use of one instrument or another depends on the purpose and extent of a survey. Quackenbush prefers the direct test over the self-report interview, but retains both instruments for particular situations. Finally, he concludes that the FSI concept of levels of proficiency provides a usable, valid, and meaningful standard for evaluating proficiency.

Since Barbara Grimes was unable to attend, I presented material from the SLOPE syllabus which was prepared by Barbara and others in cooperation with Dr. Thea C. Brun, head of the language testing unit of the Foreign Service Institute. This had been presented at a workshop held in Dakar, Senegal, in April of 1987.

SLOPE is based on the assumption that bilingualism is most accurately modeled as interactive behavior. This view largely determines the form, content and mode of scoring of the test. A brief description of SLOPE, by Barbara Grimes, is included in this volume which should prove useful for those who would like to know what the main outlines of the test are, without going into all the detail contained in the syllabus (B. Grimes 1987c).

The final paper on bilingualism was a paper by Carla Radloff, *The Sentence Repetition Test,* describing ongoing research in developing a sentence repetition test for evaluating bilingual proficiency. This paper represents work by a team consisting of Radloff, David Marshall, Charles Meeker, and several Pakistani evaluators. The overall project displays a methodological soundness, concern for reliability and validity, and meticulousness second to none in the field.

Radloff's paper was originally intended to be Part I of this volume. However, revisions and additional testing in the field led to a longer version which has been published as a separate monograph (Radloff 1991). In this volume, therefore, I confine myself to the following summary of the approach.

The sentence repetition test consists of a set of fifteen tape-recorded sentences which are played back to subjects individually through earphones. These sentences are determined beforehand to be maximally discriminating for particular levels of difficulty, and are not related semantically. The subject is expected to repeat each sentence verbally. He is scored as follows: perfect score on a sentence is 3; one error on a sentence gives him a score of 2; and two errors on a sentence gives him a score of 1. Three or more errors result in a 0.

The test is said to be a correlated one, i.e., one which takes its meaning from its relation to an independent instrument such as SLOPE or some other evaluation measure. Its function is that of screening for people who are of different proficiency levels.

Radloff has taken special care to tie the team's work in with other kinds of research. They place heavy emphasis on training their testers in order to guarantee the reliability of the results. They found no significant differences between the way both Urdu and Pashto testers scored subjects. Neither the age of the subject nor his level of education had any significant effect on his scores. However, whether a subject had been introduced to the test administrator by an acquaintance or relative, or had merely been encountered among bystanders did make a significant difference in the subject's scores.

Radloff points out that the Sentence Repetition Test (SRT) does not appear to distinguish between speakers who are at FSI levels of 3+ or higher in the evaluation language. In more recent field experimentation, they have tried to modify the SRT, making it more discriminating at higher levels of proficiency in a second language. The results and generalizations from those results are not yet known. There may well be an intrinsic limitation on the discriminability of the method, as the definitions of the FSI levels themselves suggest. I do not see this limitation as sufficient grounds for rejecting the entire approach; rather, the SRT and SLOPE can be used to complement one another. In most cases, a community will probably turn out to be below 3+ in the evaluation language. The SRT could pick this up readily. SLOPE could then be employed for the cases in which the SRT reports a very high level of proficiency.

The Radloff team took consistent and meticulous care in its approach to do things properly. The steps they followed include, among other things, developing the test in the area where the evaluation language is spoken, selecting and training the personnel involved in the study, constructing and testing a long form of the test sentences (40–50), modifying the long form into a short 15-sentence vẹrsion, developing an index of the discriminatory power of the sentences, and rating the participants by an external standard. The fact that the SRT is easy to administer once it is developed is an added bonus.

To summarize the first day of the seminar, we use Rensch's metaphor that there are "several windows on bilingualism."

The second day of the seminar was devoted to several different topics. It began with a paper by Francis A. Gray of Far Eastern Broadcasting Corporation (FEBC). Titled *The Use of Language Data in Broadcast Research and Project Development*, it was an informative presentation of both the research being carried out by FEBC, the role played by information gathered by other organizations, the nature of RICE (Radio In Church-planting Evangelism), the "World By 2000" declaration, and the scope of Christian broadcasting encompassed by Gray's research.

It is particularly interesting that the development of the personal computer was crucial to making Gray's project feasible. In addition, Gray has found *The Ethnologue* (B. Grimes 1984a) immensely useful and has benefited further from personal interaction with Joseph Grimes and Barbara Grimes.

The next paper was by Mark Taber. His *Survey: A Picture of Maluku* was an informative presentation of what had been done in Maluku up to 1987, the approaches employed in collecting survey data, the kinds of data collected thus far, and a summary of problems in training survey technicians.

Interspersed throughout the second day of the conference were three short, but important presentations by Cal Rensch, which I summarize here as a single block. The first paper, *Calculating Lexical Similarity*, mentions a few considerations in collecting word lists. It makes the point that the seemingly EASY tool of taking a word list in a language one does not know is not so simple after all. One reason is that phonologically similar words may be similar for several different reasons, e.g., they may either be related historically or they may all be related due to borrowing from another language.

Rensch points out that, for our purposes, we do not need to apply the rigorous procedures of historical linguistics to our data. Nevertheless, I suggest that, with the accessibility of the personal computer and the development of programs such as John Wimbish's WordSurv and Don Frantz' Compass G (1970), which Joe Grimes is incorporating into a menu-driven family of programs, it is becoming feasible to describe historical patterns as part of our normal analysis of survey data. Even though some colleagues have mentioned that they do not ordinarily record word lists on tape, I would like to suggest that they do so, especially in view of the tentative state of our investigators' knowledge of the language.

Interestingly enough, the Asia area survey teams have found that they need to take two different word lists from different individuals. This frequently shows from five to ten percent of differences due to sampling error. The first step in calculating lexical similarity is to regularize the discrepancies. The actual calculation is based on applying a set of criteria to successive pairs of words. Rensch devised this set of criteria as a control for variability in researcher evaluations of lexical similarity. No hard and fast decisions are made with respect to words that fail to make the criterion. Such words could still be related historically. The purpose of the calculation at this preliminary stage is simply to identify the most obvious sound correspondences.

Rensch's second paper is called *Sociolinguistic Community Profiles*. Here he addresses the problems we fall into by trying to evaluate characteristics which are not uniformly distributed throughout a population. His observation is that the members of a society who are educated and travel frequently are more proficient in a given second language than their less fortunate fellow citizens. Rensch would like to identify the specific factors related to this greater proficiency. Assuming that a multiplicity of factors underlies second language proficiency, he notes that the size of the subgroups that are associated with these factors varies from factor to factor. His solution is to conduct a census in which a representative of each household is asked about a set of twenty categories outlined by Frank Blair in his *Survey on a Shoestring* (1990).

These data, then, can be used to construct a community profile based on age groups, sex, education level, etc. Profiles of different communities can then be compared, forming the basis for useful hypotheses and arriving at believable conclusions.

Rensch's third paper is a squib titled *Language Proficiency*. Because multilingualism is so widespread in Asia, and because extensive testing for bilingual proficiency is time consuming and costly, the Asia area team is looking for diagnostic patterns of language use which show there is no need for extensive bilingualism testing.

The morning session ended with the presentation of my paper that appears in this volume and that was also presented at the International Language Assessment Conference held in Horsley's Green at the end of May 1989 (Casad 1990). It is titled *State of the Art: Dialect Survey Fifteen Years Later.*

The Tuesday afternoon session was devoted to two papers on language attitudes. The first, *Language Attitude Test in a Multilingual Setting*, was presented by Ronald Krueger. This test was developed as a course requirement for the SIL course, *Sociolinguistic Surveys*, which was taught by Joe and Barbara Grimes at the University of Oklahoma SIL and has more recently been offered at the Texas SIL in Dallas. The test that Ronald and Joanne Krueger designed was intended to sample several areas of interest. These included the domains in which Hindi, Urdu, Gujerati, and English were used, the propriety of the usage of each of these languages in the presence of nonspeakers, the attitudes of people toward each of the languages in question, and their attitude toward the peoples who spoke these languages.

Krueger made a few significant generalizations from his study. For one, language attitudes do not have just simple plus and minus values, but rather reflect a whole range of strengths and meanings. Furthermore, attitudes are not uniform throughout a culture. This means that one must sample carefully and thoroughly if he is going to draw valid conclusions. Finally, controls can be built into an attitude test by including two forms of each question at different points in the test which allowed them to judge the consistency of the subjects' responses. The test consisted of thirty questions, some of which had several parts. Finally, it took about one-and-a-half hours to administer. Although Kreuger did not suggest this test as a model for others, Joe and Barbara Grimes and I have already used it this way in the SIL survey courses.

Rensch stated that in the Asia area surveys they had noticed two patterns of language attitudes. The first pattern is based on the idea of LIMITED GOOD. In this view, people tend to think that if one of the various languages they speak is good, then the other is bad. The second pattern is based on the idea of TOTAL GOOD. All the languages in the multilingual setting are held to be good; one serves for one purpose, the other serves for another. Needless to say, these two situations involve two different strategies for establishing and carrying out field programs.

The last paper in the conference was Roland Walker's *Towards a Model For Predicting The Acceptance of Vernacular Literacy by Minority-Language Groups.* Walker attempted to identify those sociolinguistic variables that best predict if a given vernacular language group will accept literacy materials published in the mother-tongue. Walker's exploratory study cannot be taken to be the formulation of a strongly predictive model of the acceptance of literature by vernacular language groups. What first study of any sociolinguistic topic could turn out that way? Nevertheless, it suggested that certain constellations of variables may hinder the vernacular literatureacceptance of vernacular literature within particular communities. One cannot help but ask whether there are not other constellations of variables that enhance such acceptance. Surely we would want to know about this, too. Walker did not address this question so directly.

In summary, the papers presented at the Baguio conference relate to important topics and reflect careful work on the part of well-trained people. There are several implications to be drawn from all this. Teamwork is one of the most salient—dedicated people working toward a common goal. Sound methodology—there is a consistent and concerted effort to develop test instrumentsreliable test instruments and validate them. These people are using fairly sophisticated statistical methods, but also know when to opt for their intuitions instead of being misled by the numbers. Realism—they recognize the complexity of the task, deem it doable, but know that it cannot be done overnight or even in three or four years. Continuity—rather than throw overboard everything else that previous researchers have done because of their exhilaration over the development of new research in the areas of bilingualism, language use and attitudes, they have taken the best from comparative work, lexicostatistics, and intelligibility testing and have melded it into a coherent program. In short, the members of the Asia area survey team have not only validated what their predecessors had developed in Mexico, but have gone beyond that, teaching us new and useful things.

In conclusion, we hope that this volume, based on the papers given at the Baguio conference and on the work done by the Asia area survey, will benefit others interested in language assessment.

Beyond this volume, I need to mention the fine paper written by Paul Kroeger (1986), called *Intelligibility Patterns in Sabah and the Problem of Prediction*, which was published in *FOCAL I: Papers from the Fourth International Conference on Austronesian Linguistics*; Paul Geraghty, Lois Carrington and S. A. Wurm, eds. This paper represents the finest statistical treatment of survey data that I have seen to date. Finally, Frank Blair (1990) has written a survey handbook called *Survey on a Shoestring*. Following its introductory chapter that gives definitions and outlines the scope of the work, chapters two through six treat the successive topics of survey planning, dialect areas, sampling, bilingualism, and oral proficiency testing. The second half of *Survey on a Shoestring* discusses additional aspects of the assessment task, i.e., recorded text tests, observation, sentence repetition tests, self-evaluation questionnaires, and language use and language attitudes. This is a survey handbook in the full sense of the word and could easily be used in countries outside of the Asia area.

Part II: Word Lists and Intelligibility

[blank]

Calculating Lexical Similarity

Calvin R. Rensch

Most survey work includes the collection of word lists. Pairs of word lists from different dialect areas are compared to determine how similar the vocabulary of two dialects is, based on that sample.

In comparing word lists, some pairs of words seem to be similar and others do not. Sometimes pairs of similar words are thought to be cognate, i.e., similar because they have been derived historically from a common source. That may or may not be so, but determining that two words are cognate requires the analyst to work out a set of sound correspondences among the dialects compared. Then the correspondence rules are applied to the sets of words in order to determine whether the differences are a result of those sound correspondences.

Usually, for survey purposes, it is not necessary to work out this diachronic picture to distinguish true cognates from pairs of words which look alike for other reasons. We simply need to identify those items with similar meaning which are also similar phonetically (or in some cases, phonologically).

Borrowings also may affect the picture which emerges from the comparison of word lists. Borrowings from an unrelated or distantly related language may be easy to identify. However, the South Asia survey team is currently not attempting the more difficult task of identifying borrowings from more closely related languages.

In collecting and processing word lists two kinds of inconsistencies frequently emerge: (a) Dissimilar words are given on the word lists compared when, in fact, similar words are also in use; and (b) pairs of words are inconsistently classified as similar or dissimilar by different members of the team.

For the first kind the South Asia survey team has found it profitable to collect each word list at least twice, i.e., from different individuals on at least two occasions. We have found that from five to ten percent of the items on word lists from any pair of related dialects will be different unnecessarily. Different forms are given when in fact there are similar words in regular use. This may happen for any of the following reasons: (a) Since elicitation is in the language of wider communication (LWC), the response may be the LWC word even though the mother-tongue word is (also) used; (b) one list may have a generic word and the other a specific one; (c) the words may be synonyms; (d) the person who gave the word list may have misunderstood the word used in elicitation.

After the word list is collected a second time, all discrepancies are investigated. Frequently it is found that some of the differences are not real and that one of the forms is preferable, so that a single item can be entered on the word list for that dialect.

Secondly, if surveyors inspect word lists and determine intuitively whether pairs of words are similar or not, the results can vary considerably from one surveyor to another. Therefore, we have found it desirable to adopt a standard criterion for determining lexical similarity. This criterion is applied as uniformly as possible. If a pair of words fails to meet the criterion, it does not mean necessarily that the words are not at all similar or that they are not cognate; it simply means that they have failed to meet the arbitrary standard which has been adopted.

The criterion is as follows:

If the longer word of the pair has this many segments	it should have at least this many segments of category 1	and this many segments of category 2	and at most this many segments of category 3
2	2	0	0
3	2	1	0
4	2	1	1
5	3	1	1
6	3	2	1
7	4	2	1
8	4	2	2
9	5	2	2
10	5	3	2
11	6	3	2
12	6	3	3

Category 1 consists of the following types of segments: (a) contoid (consonant-like) segments which match exactly; (b) vocoid (vowel-like) segments which match exactly or differ by only one articulatory feature; and (c) phonetically similar segments (of the sort which frequently are found as allophones) which correspond in at least three pairs of words.

Category 2 is broadly defined as: all other phonetically similar pairs of segments which are, however, not supported by at least three pairs of words.

Category 3 is constituted of: (a) pairs of segments which are not phonetically similar; and (b) a segment which is matched by no segment in the corresponding word.

Two relaxations of this criterion are thought to be practical and are now being tested in field survey work.

First, a reduction in the number of segments of category 1 is permitted if it is compensated by a reduction in the number of segments of category 3. For example, the criterion for a word of six segments, as stated, requires at least three segments of category 1, at least two segments of category 2 and no more than one segment of category 3 (3–2–1). However, under this relaxation of the criterion the following distributions of segments are also acceptable: 4–0–2 and 2–4–0.

Secondly, an exhaustive search for corresponding segments will not be conducted to determine those correspondences which are supported by three examples in the word lists. Instead, the correspondences that are readily identified will be noted in reports and the category of such segments adjusted accordingly.

[blank]

Correlations Between Vocabulary Similarity and Intelligibility

Joseph E. Grimes

On the face of it, it seems reasonable that dialects whose vocabularies are similar ought to be able to understand each other rather well. Yet too often they do not. In spite of that, decisions about language programs are sometimes made on the basis of this plausible-sounding but shaky assumption.

At the low end of the scale there is a constant relationship: comprehension is always poor when vocabulary similarity is low. But that relationship does not hold up at the high end of the scale, which is where the program decisions have to be made.

The reason why high similarity is a poor predictor of high intelligibility is that there are other factors besides similarity in vocabulary that influence intelligibility. Even when vocabulary similarity is high, these factors can get in the way—the effect of differences in function words and affixes, syntactic and morphological rearrangements, certain kinds of regular sound shifts, and semantic shifts in both genetically derived vocabulary and loans.

Because intelligibility is so complex, the hopes raised thirty years ago by glottochronology should have faded by now; yet they have not. From one or two hundred forms, some still would see linguistic and cultural history laid out, and would decide at a glance where the paths of communication lie open. One reason the hope stays alive is that it is thought to be more work to calibrate intelligibility accurately than it is to collect a word list and compare it with other word lists, so a shortcut would appear to be welcome.[1] But the energy wasted on invalid shortcuts could be better employed to give us surveys that are done right the first time.

Even when intelligibility tests are given, survey reports occasionally complicate the picture by confusing intelligibility with something quite different: what amounts to bilingual behavior on the part of some of the people tested. When people learn another language, even one that is thought of as a dialect of their own language but is different enough that they cannot treat it as a simple extension of their own mother tongue, all our testing has to be done differently.

The difference comes from the fact that when a community learns a second form of speech, each person in that community does so for his or her own reasons. Some don't feel they need it, and don't learn it; some would like to, but have no opportunity; most learn it well enough for their immediate ends, but no better. This means that bilingual proficiency within a community normally varies greatly from one person to another. The sample needed to test that variation has to cover all the segments of the society, because different social sectors take differently to learning another language or dialect.

Not so with inherent intelligibility. It is an extension of ability to use the mother tongue. As a consequence, what is accessible to one member of the community is accessible to all. Its range of individual variation is fairly narrow,[2] and a smaller sample is statistically adequate for estimating it.

So, in reviewing how well or how poorly intelligibility might be predicted by vocabulary similarity, we do well to remember that when bilingual comprehension is reported as "intelligibility," we are really dealing with something whose distribution in society is quite different from that of intelligibility. In that case, the degree of understanding available to those

[1] Using word lists to study regularities in sound change is not a shortcut comparison. It is at least as time consuming and demanding as intelligibility testing, even with the aid of a computer.

[2] Most surveys fail to report either this range of individual variation or the number of speakers tested in each place. Both are needed in order to interpret correctly the average intelligibility, which is the only one of the three essential figures that usually appears. The standard deviation is a useful measure of the range of variation. It is easy to calculate and is needed for statistical reasoning. It is obtained by taking the amount by which each individual's score deviates from the average, squaring it to keep the deviations that are below the average from canceling out those above it, adding up the squares of the deviations, dividing the sum by the number of test subjects to get the average squared deviation, then taking the square root of that to put it all back onto the original scale. Standard deviations for inherent intelligibility are normally less than fifteen percent (a ball park figure, not yet validated precisely), while a larger standard deviation is typical of bilingual situations.

who have not gone out of their way to learn the other form of speech is lower than the figure given as if it represented uniform intelligibility.[3]

Philippines

Vocabulary similarity estimates and intelligibility test results—subject to the cautions just given—are available for fifty-five pairs of dialects in the Philippines. In these pairs the intelligibility is only weakly correlated with the vocabulary similarity.

The data are from the Tenth Edition of the Ethnologue (B. Grimes 1984), reproduced in (1) and displayed in (2). They are based on field surveys made by the Philippines Branch of the Summèr Institute of Linguistics. The fractions of a percentage point given in the Ethnologue are rounded off here to reflect better the level of accùracy that tests of the kind given yield.[4]

Seven other dialect pairs were left out because the figures reported for them in the Ethnologue for intelligibility are known to involve a substantial amount of bilingualism, yet they come from tests on samples that were too small to be valid for the variation that goes with bilingualism. The languages involved are not even in the same linguistic subgroupings; on purely comparative grounds it would be strange if they understood each other inherently. Atta of Pamplona tested on Ilocano had a similarity of 63% and what was called "intelligibility" of 85%; Itawit on Ilocano, 53% and 68%; Kasiguranin on Tagalog, 52% and 92%; Agusan Manobo on Cebuano, 81% and 88%; Obo Manobo on Cebuano of Nasuli, 35% and 78%; Central Tagbanwa on Cuyonon, 48% and 61%; and Central Tagbanwa on Tagalog, 40% and 54%. Ilocano, Tagalog, and Cebuano are

[3] Casad (1974:177) gives a set of individual scores from the Mazatec survey. The scores for Huautla (Hu) break into three groups. Three people have scores in the fifties: I would guess they represent the real intelligibility. Four have scores 90 to 100: they could be the practiced bilinguals, though this kind of test cannot distinguish high fluency from only moderate proficiency. The other three are in between, possibly reflecting low proficiency in the Huautla dialect. Huautla is a market and cultural center whose speech is learned by people from the countryside. The mean intelligibility is 76%, but the standard deviation is 18%, enough of a scatter to raise suspicion.

[4] Three significant digits gives a spurious impression of the accuracy that can be attained from the form of the test usually given. Rounding the community averages to the nearest five percent would reflect the inherent precision of that type of test even better than rounding to one percent. Tests and sampling procedures could be devised that would be accurate to one percent, but they would be extremely costly, and the decisions about language programs made on that basis would not be noticeably different from tests accurate to five percent.

(1) Philippine dialect pairs, with vocabulary similarity and intelligibility data from the Ethnologue

	VOCAB SIMIL	INTELL	DIALECT TESTED on DATA FROM REFERENCE DIALECT
A	34	69	Bagobo on Tagabawa Manobo
B	48	61	Central Tagbanwa on Cuyonon
C	52	90	Agutaynon on Calamian Tagbanwa
D	54	78	Cuyonon on Tagbanwa
E	57	29	Central Tagbanwa on Lamane
F	57	56	Central Tagbanwa on Calamian Tagbanwa
G	63	60	Obo on Tagabawa
H	65	83	Madukayang on Balangao
I	65	66	Tagbanwa on Quezon Palawano
J	66	63	Yaga on Central Cagayan Agta
K	66	82	Mt. Iriga Agta on Mt. Iraya Agta
L	66	72	Mt. Iriga Agta on Central Bicolano
M	66	92	Kamayo on Surigaonon
N	68	83	Madukayang on Limos
O	68	66	Ambala on Botolan Sambal
P	69	52	Pamplona Atta on Itawit
Q	70	87	Butuanon on Kamayo
R	71	73	Aklanon on Hiligaynon
S	72	64	Ibatan on Itbayatan Ivatan
T	72	78	Batad on Kiangan Ifugao
U	72	91	Piso on Kagan Kalagan
V	74	92	Mansaka on Kagan Kalagan
W	74	31	Ibatan on Basco Ivatan
X	75	82	Kasiguranin on Paranan
Y	75	85	Sibuco-Vitali on Balangingi Sama
Z	76	86	Mt. Iriga Agta on Iriga Bicolano
A	76	78	Karao on Ibaloi
B	76	81	Rajah Kabungsuan Manobo on San Miguel Calatugan Agusan
C	77	84	Lutungan on Balangingi Sama
D	78	87	Ayangan on Batad Ifugao
E	78	88	Hapao on Kiangan Ifugao
F	78	48	Tanudan on Butbut
G	78	47	SW Palawano on Central Palawano
H	78	76	SW Palawano on Quezon Palawano
I	79	70	Butbut on Guinaang Kalinga
J	80	77	Amganad on Kiangan Ifugao
K	80	94	Calamian Tagbanwa on Baras
L	81	81	Guinaang on Balbalasang
M	81	70	Guinaang on Limos

(1) cont.

	VOCAB SIMIL	INTELL	DIALECT TESTED on DATA FROM REFERENCE DIALECT
N	81	86	Madukayang on Mangali
O	82	74	Bangad on Butbut
P	82	81	Rajah Kabungsuan Manobo on Dibabawon
Q	82	76	Brooke's Point Palawano on Quezon Palawano
R	82	96	Brooke's Point Palawano on Central Palawano
S	83	81	Burnay on Amganad Ifugao
T	83	88	Brooke's Point Palawano on Southern Palawano
U	85	87	Mayoyaw on Batad Ifugao
V	85	83	Agusan on Dibabawon
W	85	69	Brooke's Point Palawano on SW Palawano
X	85	76	SW Palawano on Brooke's Point Palawano
Y	85	98	Palanan Dumagat on Paranan
Z	87	94	Casiguran Dumagat on Paranan
A	87	85	Hungduan on Kiangan Ifugao
B	87	63	Sindanga on Tuboy-Salog
C	91	98	Pamplona Atta on Northern Ibanag

used over wide regions of the Philippines, and Pilipino, based mainly on Tagalog, is taught in the schools.

It is likely that unrecognized bilingualism is a factor in some of the other samples reported here as well. If it is, the degree of understanding of the part of the population that has not learned the other language very well is sure to be lower than the averages suggest.

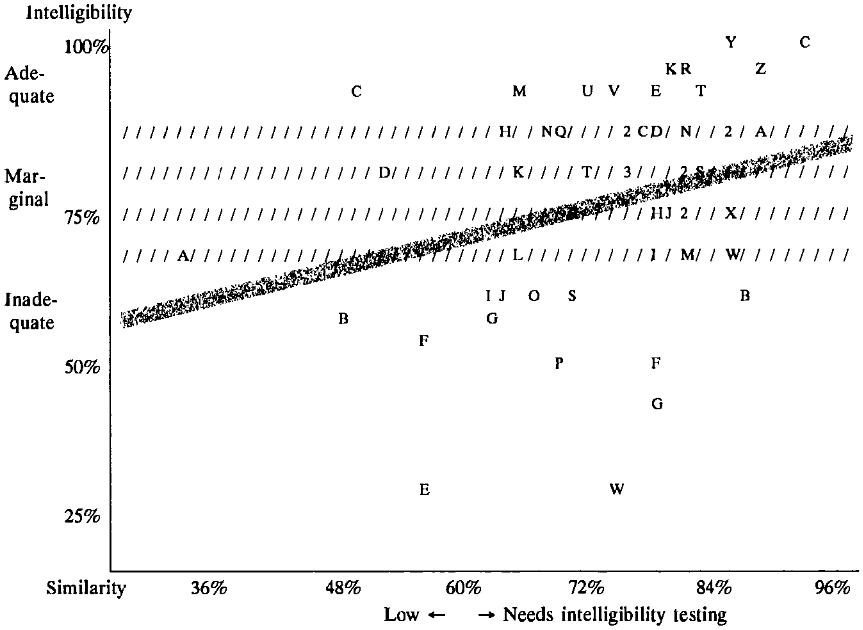
There is also a question about how the vocabulary similarity figures were arrived at. In some cases—I do not know which—the figures probably represent the proven cognates that remain between two word lists after borrowing and internal analogies have haphazardly upset the smooth progress of sound change. Such cognate judgments would be based on the extensive studies of phonological comparison that have been made in parts of the Philippines. In most cases, however, I take the figures to represent proportions derived from impressionistic counts of words that appear to be phonetically similar, without any way to distinguish those that come from a single parent form at an earlier historical stage via demonstrable sequences of sound changes on the one hand, and loans and analogical formations on the other.[5]

[5] There are, of course, many other dialect pairs in the Philippines for which neither similarity nor intelligibility figures have been compiled. The ones given represent dialects that were judged close enough to be worth testing. Most of the others fit into the low intelligibility, low similarity category mentioned in the first paragraph of the paper. They would fill the lower left quadrant of (2).

The figure in (2) shows how vocabulary similarity figures relate to intelligibility measurements in the Philippines. The fifty-five dialects in (1), from which the figure in (2) is derived, are arranged from low to high similarity. They are identified with the letters A, B, C, and so forth for the first 26 on the list, corresponding to the left part of the figure, again as A, B, C,. . . for the next 26 going toward the right, and A, B, C for the last three on the extreme right. Where two or three dialects fall on the same point, a digit is given instead of a letter to show how many dialects are there.

(2) Vocabulary similarity and intelligibility in 55 dialect pairs in the Philippines.

A, B,. . . identify rows in (1), and are repeated after the 26th and 52nd dialects on the list.
Numbers are used where dialects appear on the same spot.
The zone of marginal intelligibility is shown by ////////////.
The regression line for the relation is shown by ▒▒▒.



Normally, vocabulary similarity percentages of 60% and below go consistently with intelligibility measured at 67% and below on simple narrative' material (Simons 1979). That level, for practical purposes, is inadequate for all but the simplest communication. Intelligibility seems to have to be above 85%, as measured on narrative, before much complex and personally revealing communication is likely to take place; Casad's discussion of Kirk's validation tests for Mazatec (Casad 1974:83–86) points to a 90% threshold for being able to extrapolate from a test on narrative to more complex kinds of communication.

It is clear from the figure in (2) that similarity figures above Simons's 60% similarity line go with a wide range of intelligibility: from 31% (W) in the lower right to 98% (C) in the upper right. They are evenly balanced in that ten indicate adequate intelligibility (90% and up) and ten indicate intelligibility (under 70%). What relation there is between vocabulary similarity (s) and intelligibility (i) is indicated by the slanting shaded line, whose equation is

$$i = 0.465s + 41.8.$$

But the scatter of the dialects away from the line, measured by a correlation of 0.34, shows that there is only a weak relation between the two scales.

The actual dialect pairs for the Philippines that are displayed in (2) are given in (1), with the vocabulary similarity and intelligibility figures for each. They begin with the lowest similarity figures in order to make it easy to visualize how the corresponding intelligibility scores vary.

Correlations

The table in (3) and the condensed counterparts of it that are given in (4) for other data pinpoint a few areas for which there actually is a high correlation between vocabulary similarity and intelligibility over a part of the range. As is plain from (3), occasional areas of high correlation show up in very limited combinations of similarity and intelligibility, and comparing it with the tables given later shows that it cannot be predicted generally from one language area to another.

In (3) the data are the Spearman *r* correlations for all dialect pairs whose vocabulary similarity is greater than or equal to the threshold percentage given at the bottom of the column, and whose intelligibility test results at the same time are greater than or equal to the threshold given at the left of the row. The correlations are given to two decimal places in the (a) part of (3), and in a condensed form of one digit with the decimal point removed in the (b) part, which is also the format of the table in (4).

The asterisks (*) represent those parts of the table for which there are fewer than five pairs available. From fewer pairs it is impossible to calculate a meaningful correlation.

(3) Philippine dialect pairs: correlations of vocabulary similarity and intelligibility from the Ethnologue.

All dialect pairs whose similarity is equal to or higher than the figure at the bottom, and whose intelligibility is equal to or higher than the figure at the left, are correlated: (a) to two decimal places, (b) condensed

(a)

PHILIPPINES.DATA
Intelligibility> =I%

90%	0.85	0.83	0.83	0.85	0.52	0.52	*.**	*.**
85%	0.22	0.30	0.30	0.41	0.54	0.24	0.37	*.**
80%	0.19	0.31	0.31	0.28	0.49	0.37	0.48	*.**
75%	0.16	0.18	0.18	0.24	0.40	0.40	0.53	*.**
70%	0.14	0.21	0.21	0.24	0.36	0.46	0.53	*.**
0%	0.34	0.29	0.25	0.23	0.23	0.19	0.34	*.**
	0%	60%	65%	70%	75%	80%	85%	90%

Similarity> =S% for 55 pairs
* N <5 too few to correlate

(b)

PHILIPPINES.DATA (55)
Intelligibility> =I%

90%	8	8	8	8	5	5	*	*
85%	2	2	2	4	5	2	3	*
80%	1	3	3	2	4	3	4	*
75%	1	1	1	2	4	4	5	*
70%	1	2	2	2	3	4	5	*
0%	3	2	2	2	2	1	3	*
	0	6	6	7	7	8	8	9
	0	0	5	0	5	0	5	0 %

Similarity> =S%
0 to 9 R=.0 to .9

The table focuses on correlations in the area of interest, from 60% and up for vocabulary similarity and from 70% and up for intelligibility. No figures are given for 95% and up; at that level there are not enough instances in any of the data sets to produce a correlation. An additional row and column have been added in order to include the full range of levels in the table.

The lower left corner cell (0,0) of (3) gives the correlation for all the data. In effect, it measures the entire scatter away from the line of regression that the figure in (2) shows: 0.34.

If we look only at that part of the data where Simons expects to find a useful level of intelligibility, 60% lexical similarity and higher and 70% intelligibility and higher (the 60% column with the 70% row), the scatter is relatively greater and the correlation measure is closer to zero: 0.21.

The other figures in the table show what the scatter would be if we were to cover up the lower part of the figure in (2) so as to focus only on higher levels of intelligibility, and cover up the left hand part so as to focus on higher levels of vocabulary similarity, calculate the equation for another regression line like the one that crosses (2) at an angle, but that covers only the points that have not been covered up, then calculate how much those data points deviate from what the new equation predicts they should be. For example, if we look only at those pairs of dialects whose similarity measures are 65% or more (the 65% column in (3)) and whose intelligibility measures are 80% or more (the 80% row in (3)), the correlation measure that indicates how well the equation predicts the facts is 0.31.

Perfect correlation would have been registered if all the data in a plot like (2) had fallen on a straight line instead of scattering out all over the plot as they have. In (3), perfect correlation would be expressed by a figure of 1.00.[6] Coefficients of correlation below 0.50 indicate considerable scatter of points off the regression line and are unimpressive in arguing for a consistent relationship between two variables; correlations below 0.30 indicate a still vaguer relationship, more scatter than the 0.34 that (2) demonstrates. The computer program that calculated the 1,953 correlations that went into (3) and (4) is reproduced in appendix D.[7]

The overall correlation improves in places as some higher levels of vocabulary similarity and intelligibility are considered, but at best it indicates a loose relation, not one with high predictive value. There is an exception in the 90% row of the table when similarity values lower than 75% are taken into account; suddenly everything appears closely correlated. The data in question are the top three rows of (2), those pairs with intelligibility 90% or higher, which can be seen to fall fairly close to a straight line. The dialects involved, those in the upper left of (2), are probably showing the effects of bilingualism. No analogous localized pocket of high correlation shows up in any of the ten other sets of data investigated in this way. It therefore appears to be a local fluke that can be attributed to a few situations where bilingual behavior was not recognized; it is probably not the manifestation of any principle.

Other language areas

Similar pairings of figures are available in Simons's monograph (1979) for ten other areas of the world. As with the Philippines, in most cases it is not possible to know whether the vocabulary similarity figures are based on counts of genetically demonstrable cognates or on apparent phonetic similarity alone, nor can we be sure that the intelligibility tests were not applied to bilingual behavior by mistake. Some of the intelligibility testing was done before the internal safeguards described by Casad (1974) were developed: the Iroquois tests, for example, were the first ever given.

[6] Positive correlations go with regression lines that rise from left to right like the one in (2). If the line fell, indicating inverse correlation (the more of this, the less of that), the correlation would be negative, and perfect negative correlation would be expressed as -1.0.

[7] Only 1,448 correlations are actually given because the rows and columns that correspond to 95% were dropped from all the tables after they were calculated, since none contained enough data to give a valid correlation. Each correlation involved from six pairs of numbers for Biliau to twenty-nine pairs for Polynesian.

The data are taken from Simons (1979), who also discussed the sources and their quality. In order to show how vocabulary similarity and intelligibility correspond or fail to correspond in general, I have analyzed his data both with and without various adjustments he proposes. With one exception, the adjustment factors influence the overall picture hardly at all.

One of his adjustments is for discrepancies greater than 10% in intelligibility measurements made in two directions, village A tested on village B and village B tested on village A. Intelligibility scores are almost always asymmetric in this respect. Simons suggests that discrepancies greater than 10% are due to social factors rather than linguistic factors. Since the social factors he refers to are more or less equivalent to the bilingual learning I said sometimes takes place between closely related dialects, a threshold on the order of the 10% he suggests is one way to recognize those factors tentatively, though its magnitude needs to be validated in areas like Spanish vs. Portuguese and in Chinese dialects, where difficulties in intelligibility can be traced to specific areas of phonology. (In the data from the Philippines given in (3), none of the intelligibility data report tests given in two directions. It is possible that only the higher score for a pair of tests was reported.)

The adjustment Simons makes for asymmetry is to exclude the pair of dialects with the higher intelligibility score, reasoning that the lower score is less likely to reflect a bilingual learning factor. In the data marked "exclude" I follow his practice.

A major difference between Simons's correlations and mine, however, is that he includes the measure of a dialect on itself—often called the "home-town" measure—among the data to be correlated, and I do not. Discrepancies between home-town scores actually measured and the 100% we might expect are part of the information needed in order to calibrate the test itself, but they are not part of the statement of the problem I am addressing. Even where the test results average below 100%, the effect of including the home-town scores is a considerable increase in the correlations. I have therefore left all home-town scores out of the calculations.

Because the earlier test designs (including the ones Casad reports) had not eliminated the sources of low home-town scores,[8] it was thought

[8] The problems introduced artificially by adjusting intelligibility scores that fall below 100% for the subjects' home towns have been greatly reduced by stipulating that the only questions considered admissible for a test be those on which the panel of home-town speakers that Casad (1974) uses have a score of 100%. At the time he published his monograph it was still not clear that this would work. On pp. 61–62 he speaks of throwing out questions that half the speakers had difficulty with; the improvement has come from following through on this concept by throwing out all questions (out of a very large initial pool of possible ones) that the panel cannot

necessary to adjust average scores by applying a correction factor based on the low home-town score. Two kinds of corrections were applied. If it was assumed that speakers learned to take the test by the time they had finished the home-town part of it, the home-town score could be safely taken as not distinct from 100%, and the other scores could be left alone. Simons applies this kind of correction to his data for Biliau, Ethiopia, Mazatec, Siouan, and Trique. In (4) the results of adjustment for these are the same as for the basic data.

If, on the other hand, it was assumed that the difficulty carried over from one test into another, then the home-town score was treated as equivalent to 100% and others were adjusted up in proportion to it. It is corrections of this form that Simons applied to Buang, Iroquois, Polynesia, Uganda, and Yuman. In (4) the results of adjustment for these are too small to notice.

Adjustments of this kind ought not to be applied willy-nilly to data from other parts of the world without independent proof that conditions sufficient to justify them hold there; often they do not. Improved criteria for test construction have for practical purposes eliminated the need for adjusting scores.

I therefore give in (4) three kinds of summaries of the correlation ranges in Simons's data as he gives them in his appendix 1. The first, tagged as DATA, has no special adjusting factors applied. In that form the summaries are typical of the information that becomes accessible during the course of many language surveys at a stage before some of the possible adjusting factors can be estimated.

Second, for tests where other scores are adjusted proportionally to the change made in the home-town score, data from Simons's "adjusted intelligibility" column are used in tables tagged as ADJUST.

Finally, in the tables tagged with EXCLUDE, I exclude from the computation the scores he marks with "X" because they are more than 10% higher in intelligibility than the score going the other way. The small tables that make up (4) are the (b) or condensed form of the tables that were explained in connection with (3). Each adds in parentheses the total number of dialect pairs available when all levels are taken into account.

answer well. In earlier testing, inability to answer questions in one's mother tongue had been thought to be due mainly to the unfamiliarity of the test situation, so the corrections proposed treated it as an error factor in learning that would diminish with time and experience. For the argument I am making, I have to assume that its noise effect is randomly distributed throughout these data, because we have no record of the order in which the test tapes used were presented to different subjects, and hence of how their responses might be weighted for learning behavior.

The negative *r* correlations in Buang, Polynesian, and Siouan indicate a reverse relation: the higher the similarity, the worse the intelligibility, within the thresholds given. Negative correlations are shown as minus signs in (4).

Of the eleven sets of scores presented in (3) and (4), most of the data that contribute to high correlations come in the area of low vocabulary similarity predicting low intelligibility. To the extent that adjustments based on hypotheses about how home-town scores work and exclusions of large asymmetries make any difference at all, they also make it mostly in this same area.

Higher on the scale, where marginal intelligibility is involved, two sets of scores out of the eleven, those for Mazatec and Trique, show a strong overall correlation (taking the pocket of high correlations with adequate intelligibility in the Philippines as something with no parallel anywhere else). Yuman has a stronger correlation than Mazatec in the area of marginal intelligibility, but only when asymmetries are excluded.

It may be significant that Mazatec and Trique comparative linguistics is fairly well advanced. Gudschinsky, who studied Mazatecan, and Longacre, who studied Trique, had both been in vigorous debate with Morris Swadesh about the validity of lexical similarity measures in general. They and Paul Kirk, who did the Mazatec survey with Casad, commanded both a working knowledge and a comparativist's broad grasp of the dialects that went far beyond what might turn up by the luck of the draw on a survey word list; they knew about cognates that might never be noticed without the detailed comparative work they had already done.

Discussion

To the extent that the areas for which we already have data on vocabulary similarity and intelligibility represent dialect areas in the world in general, chances are 4.5 to 1 against any survey that attempts to assess intelligibility solely on the basis of vocabulary similarity being able to do so with any confidence, on the basis of 9 areas of low correlation against 2 of higher correlation.[9] One has to do the intelligibility testing anyway, not only because it is the best indicator we have of areas of high communication potential, but in a secondary sense in order to validate whether the dialect area might indeed be one of the less likely ones in which vocabulary similarity bears a significant relation to comprehension.

[9] The score is 8 to 3 if the Yuman data after exclusions are taken into account as well.

(4) Correlation ranges in ten language areas. (Data from Simons 1979, Appendix 1.)

Tables are in the condensed format of (3b) .

* indicates fewer than 5 pairs, insufficient to correlate.

– indicates negative correlation

DATA: unadjusted

ADJUST: adjustment based on home town score

EXCLUDE: exclusion of asymmetries over 10%

```
BILIAU.DATA (6)
Intelligibility> =I%
90% * * * * * * * *
85% 2 2 2 2 2 2 * *
80% 3 3 3 3 3 3 * *
75% 3 3 3 3 3 3 * *
70% 3 3 3 3 3 3 * *
 0% 3 3 3 3 3 3 * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
    Similarity> =S%
    0 to 9 R=.0 to .9
```

```
BILIAU.ADJUST (6)
Intelligibility> =I%
90% * * * * * * * *
85% 2 2 2 2 2 2 * *
80% 3 3 3 3 3 3 * *
75% 3 3 3 3 3 3 * *
70% 3 3 3 3 3 3 * *
 0% 3 3 3 3 3 3 * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
    Similarity> =S%
    0 to 9 R=.0 to .9
```

```
BILIAU.EXCLUDE (3)
Intelligibility> =I%
90% * * * * * * * *
85% * * * * * * * *
80% * * * * * * * *
75% * * * * * * * *
70% * * * * * * * *
 0% * * * * * * * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
    Similarity> =S%
    0 to 9 R=.0 to .9
```

```
BUANG.DATA (18)
90% * * * * * * * *
85% * * * * * * * *
80% – – – – – – * *
75% 4 4 4 – – – * *
70% 5 5 4 – – – * *
 0% 5 5 5 3 3 0 * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
```

```
BUANG.ADJUST (18)
90% * * * * * * * *
85% * * * * * * * *
80% – – – – – – * *
75% 4 4 4 – – – * *
70% 5 5 4 – – – * *
 0% 5 5 5 3 3 0 * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
```

```
BUANG.EXCLUDE (12)
90% * * * * * * * *
85% * * * * * * * *
80% * * * * * * * *
75% 5 5 5 * * * * *
70% 5 5 5 * * * * *
 0% 7 7 6 7 7 4 * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
```

```
ETHIOPIA.DATA (25)
90% * * * * * * * *
85% * * * * * * * *
80% * * * * * * * *
75% * * * * * * * *
70% * * * * * * * *
 0% 6 9 * * * * * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
```

```
ETHIOPIA.ADJUST (25)
90% * * * * * * * *
85% * * * * * * * *
80% * * * * * * * *
75% * * * * * * * *
70% * * * * * * * *
 0% 6 9 * * * * * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
```

```
ETHIOPIA.EXCLUDE (18)
90% * * * * * * * *
85% * * * * * * * *
80% * * * * * * * *
75% * * * * * * * *
70% * * * * * * * *
 0% 5 9 * * * * * *
    0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%
```

```
IROQUOIS.DATA (10)          IROQUOIS.ADJUST (10)        IROQUOIS.EXCLUDE (8)
90% * * * * * * * *         90% * * * * * * * *         90% * * * * * * * *
85% * * * * * * * *         85% * * * * * * * *         85% * * * * * * * *
80% * * * * * * * *         80% * * * * * * * *         80% * * * * * * * *
75% * * * * * * * *         75% * * * * * * * *         75% * * * * * * * *
70% * * * * * * * *         70% * * * * * * * *         70% * * * * * * * *
 0% 6 1 1 * * * * *          0% 6 1 1 * * * * *          0% 6 * * * * * * *
    0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%

MAZATEC.DATA (13)           MAZATEC.ADJUST (13)         MAZATEC.EXCLUDE (11)
90% * * * * * * * *         90% * * * * * * * *         90% * * * * * * * *
85% * * * * * * * *         85% * * * * * * * *         85% * * * * * * * *
80% * * * * * * * *         80% * * * * * * * *         80% * * * * * * * *
75% * * * * * * * *         75% * * * * * * * *         75% * * * * * * * *
70% 8 8 8 8 8 9 * *         70% 8 8 8 8 8 9 * *         70% 8 8 8 8 8 9 * *
 0% 6 6 6 6 6 7 * *          0% 6 6 6 6 6 7 * *          0% 7 7 7 7 6 7 * *
    0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%

POLYNESIA.DATA (69)         POLYNESIA.ADJUST (69)       POLYNESIA.EXCLUDE (59)
90% * * * * * * * *         90% * * * * * * * *         90% * * * * * * * *
85% * * * * * * * *         85% * * * * * * * *         85% * * * * * * * *
80% * * * * * * * *         80% * * * * * * * *         80% * * * * * * * *
75% – – – – * * * *         75% – – – – * * * *         75% * * * * * * * *
70% – – – – * * * *         70% – – – – * * * *         70% * * * * * * * *
 0% 6 5 4 3 5 * * *          0% 6 5 4 3 5 * * *          0% 7 6 5 4 5 * * *
    0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%

SIOUAN.DATA (20)            SIOUAN.ADJUST (20)          SIOUAN.EXCLUDE (15)
90% * * * * * * * *         90% * * * * * * * *         90% * * * * * * * *
85% * * * * * * * *         85% * * * * * * * *         85% * * * * * * * *
80% – – – – – – – –         80% – – – – – – – –         80% * * * * * * * *
75% – – – – – – – 0         75% – – – – – – – 0         75% 5 5 5 5 5 5 5 5
70% – – – – – – – 0         70% – – – – – – – 0         70% 5 5 5 5 5 5 5 5
 0% 7 7 7 7 7 7 6 0          0% 7 7 7 7 7 7 6 0          0% 8 8 8 8 8 8 7 0
    0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%

TRIQUE.DATA (11)            TRIQUE.ADJUST (11)          TRIQUE.EXCLUDE (7)
90% * * * * * * * *         90% * * * * * * * *         90% * * * * * * * *
85% * * * * * * * *         85% * * * * * * * *         85% * * * * * * * *
80% 6 6 6 6 6 5 * *         80% 6 6 6 6 6 5 * *         80% * * * * * * * *
75% 6 6 6 6 6 5 * *         75% 6 6 6 6 6 5 * *         75% * * * * * * * *
70% 7 7 7 7 7 5 * *         70% 7 7 7 7 7 5 * *         70% * * * * * * * *
 0% 6 6 6 6 6 5 * *          0% 6 6 6 6 6 5 * *          0% 8 8 8 8 8 8 * *
    0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%
```

```
UGANDA.DATA (8)             UGANDA.ADJUST (8)           UGANDA.EXCLUDE (7)
90% * * * * * * * *         90% * * * * * * * *         90% * * * * * * * *
85% * * * * * * * *         85% * * * * * * * *         85% * * * * * * * *
80% * * * * * * * *         80% * * * * * * * *         80% * * * * * * * *
75% * * * * * * * *         75% * * * * * * * *         75% * * * * * * * *
70% * * * * * * * *         70% * * * * * * * *         70% * * * * * * * *
 0% 8 7 * * * * * *          0% 8 7 * * * * * *          0% 9 9 * * * * * *
    0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%

YUMAN.DATA (20)             YUMAN.ADJUST (20)           YUMAN.EXCLUDE (16)
90% 4 4 4 4 4 4 4 *         90% 4 4 4 4 4 4 4 *         90% * * * * * * * *
85% 3 3 3 3 3 3 3 3         85% 3 3 3 3 3 3 3 3         85% * * * * * * * *
80% 2 2 2 2 2 2 2 4         80% 2 2 2 2 2 2 2 4         80% * * * * * * * *
75% 2 2 2 2 2 2 2 4         75% 2 2 2 2 2 2 2 4         75% * * * * * * * *
70% 5 5 5 5 5 5 5 4         70% 5 5 5 5 5 5 5 4         70% 9 9 9 9 9 9 9 *
 0% 9 9 5 5 5 5 5 4          0% 9 9 5 5 5 5 5 4          0% 9 9 9 9 9 9 9 *
    0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9             0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%            0 0 5 0 5 0 5 0%
```

This does not mean that we abolish the use of word lists in language surveys. It means instead that we no longer try to squeeze out of them information they are inherently incapable of giving. They do show up areas where intelligibility is unlikely, the ones where similarity is below 60%. Above that, counts based on them are helpful mainly to point up the need for intelligibility testing, but they are not a substitute for it.

Word lists should be used instead—especially now that we have rapid methods for establishing consistencies in sound correspondence—to give an initial picture of language groupings based on shared innovations in sound change, and to show the specific sound changes that result in those groupings. For such groupings, based on demonstrable genetic divergence, we can if we like quantify the conclusions reached by the comparative method, whether through phonostatistical indices of divergence,[10] groupings based on shared rules,[11] or computations of vocabulary similarity based on the retention of proven cognates under various conditions that encourage or discourage borrowing.

But even quantifications of full-fledged comparisons do not measure the other factors that are known to influence intelligibility. We have no comparable measures yet for calibrating morphological differences, syntactic

[10] See Grimes and Agard 1959 and Grimes 1964 for phonostatistical quantification.

[11] Suitable methods are described in Romesburg 1984, in sections on qualitative resemblance matrices.

differences,[12] or shifts of meaning, or for the social and geographic factors; nor if we had them would they necessarily combine meaningfully with a similarity index into a single composite figure that we could validate against measures of comprehension. For the present, however, we do have an effective strategy for arriving at decisions about language programs[13] that gives reasonable results unless we try to cut corners with it.

1. Inspect word lists.
2. If similarity is below 60%, assume separate programs.
3. If similarity is 60% or better, test for intelligibility after screening subjects for possible bilingual learning of the other dialect.
4. If intelligibility scores are uniform (standard deviation below 15%) and average scores are 85% and above, combined programs may be possible.
5. If combined programs are linguistically possible, test social attitudes to make sure a combined program is feasible. The sampling requirements and the testing strategy for questionable cases are very different from the ones appropriate for testing for inherent intelligibility.
6. If intelligibility scores are spread out (standard deviation 15% or greater), the problem becomes one of assessing what proportion of the population is at each level of bilingual proficiency. The sampling requirements and the testing strategy for determining this are very different from those appropriate for inherent intelligibility.

[12] Andrew S. Noetzel and Stanley M. Selkow, and David Sankoff and Roger J. Cedergren, in two chapters in Sankoff and Kruskal 1983, lay the groundwork for tree-to-tree comparison measures.

[13] Barbara F. Grimes (1985a) explains the rationale for this strategy.

Inherent Intelligibility, Bilingualism, or Both?

Randy Kamp

Until quite recently there has been a tendency to view situations in terms of either intelligibility or bilingualism. For example, in the Sociolinguistic Survey Conference held in November of 1982, John Bendor-Samuel presented a paper in which he said, "Probably most, if not all, of the situations we encounter fall into one of the following situations: intelligibility or bilingualism" (1982:13).

The emergence of this dichotomy, at least within SIL, is not without historical explanation. Although the methodology developed by Casad (1984) and others did not explicitly claim to be designed to measure INHERENT intelligibility, as opposed to LEARNED intelligibility, its primary focus was to discover the centers of dialects in order to know the best location for vernacular literacy programs. Also, intelligibility testing was first done in areas where the language of wider communication, Spanish, is not related to the Amerindian languages. It is not surprising, then, that eventually, language surveyors would come to realize that different testing techniques are needed for these two typically different situations.

However, it is becoming more and more apparent that many situations are considerably more complex. In her article entitled *Evaluating Bilingual Proficiency in Language Groups for Cross-Cultural Communication*, Barbara Grimes (1986a:19) states, "In situations where the second language is not related to the other language, it is easy to distinguish between bilingualism and intelligibility from linguistic closeness. However, when the second language is related to the first, it is sometimes difficult to distinguish what understanding is inherent because of linguistic closeness, and what is learned through contact." Although she calls this "a bilingual overlay on intelligibility," I prefer to call it LEARNING-MODIFIED INHERENT INTELLIGIBILITY.

This paper will attempt to summarize the results of a survey among the Karao people of the Philippines, a mixed situation such as is described above. First, a brief description of the geographical and linguistic setting will be offered. Secondly, an overview of the survey will be presented. Thirdly, each of four survey tools used will be discussed and, where appropriate, correlations between them will be noted. Finally, any insights which might be of benefit to surveyors in similar situations will be recapped.

The Karao people and language

The Karao people, numbering about 1300, live in eastern Benguet in the municipality of Bokod, about 56 kilometers from Baguio City. They live in two barangays, Karao and Ekip, the former being only 3.4 kilometers by road from Bokod Poblacion. All of the other barangays of Bokod are Ibaloi-speaking; therefore, their neighbors consider this little linguistic pocket to be something of an oddity.

Most researchers have considered both Karao and Ibaloi to be southern Cordilleran languages (Walton 1977:21; McFarland 1980:76). Although there is now little doubt that Ibaloi and Karao are more closely related to each other than to any other languages, it is important to know just how similar they really are. Unfortunately, linguistic similarity is not an easy concept to quantify due to the complexity of the systems involved. In fact, no model has yet been developed which allows an investigator to be able to summarize linguistic similarity in one number. Certain aspects of linguistic similarity have received most of the attention of investigators. Lexicostatistics has been widely used for many years, although not without its detractors. Phonostatistics has also been attempted, but it has not gained the acceptance that lexicostatistics has. Almost nothing has been done to try to quantify linguistic similarity.

A comparison of Karao to Ibaloi reveals the following results. An apparent cognate count yields a figure of 70%. Phonologically, we notice a great deal of similarity in the inventory of phonemes. Also, a quick phonostatistical analysis reveals .95 average number of features different per cognate word, indicating that cognate words or borrowed words do not undergo major phonological changes. Therefore, if it were possible to quantify phonological similarity, perhaps we would discover that there is greater than 70% similarity between the two systems. Grammatically, the similarities significantly outweigh the differences. The personal pronouns and deictics are almost all the same and the systems of verb morphology are very close. It is only in the set of case-marking particles where one finds significant differences. Considering these results as a whole, it would seem safe to conclude that there is at least 75% linguistic similarity between Karao and Ibaloi.

Overview of the 1986 Karao survey

The survey which is the subject of this paper was conducted while my family and I lived in Karao from November 1985 until May 1986. It is not the first which has been conducted there. In 1973 a survey team visited the area and concluded that Ibaloi would probably adequately meet their needs. More recently, in 1984, a survey which involved three days of testing led to the recommendation that literature in the Karao language is necessary. These somewhat contradictory results are partly due to the fact that the 1973 surveyor viewed the situation as one of inherent intelligibility, while the 1984 surveyor interpreted the results as if it were bilingualism.

Aware that both linguistic similarity and learning are contributing to intelligibility of Ibaloi, and interested in seeing what correlation there is between the survey methods, we set out to use several survey tools. We knew also that we needed to assess both intelligibility and acceptability. The table in (1), based upon an article by George Huttar (1977:48), provides a summary of our survey methodology. For the purposes in this paper, only the four methods which have to do with intelligibility or proficiency are discussed below.

Obviously it is possible to make inferences about the linguistic ability of the Karao people only if the sample that was tested is truly representative of the entire population. And, if factors such as education or age might affect their ability, it is important that the size of the sample reflects these parameters within the community. To do this we used a quota sampling method. In this type of sampling the proportions of the various subgroups of the population are determined, and the sample is drawn to have the same percentages in it (Downie and Heath 1974:153).

We chose to limit the number of parameters to three: age, sex, and education; each was broken into two subgroups. For our purposes, age was divided into young—ages 13 to 39, and old—ages 40 to 65. Similarly, because any schooling beyond elementary is done in the Ibaloi area, we divided education into less educated—less than high school, and more educated—one year of high school or more. Obviously sex was broken into male and female.

(1) Survey goals

	Controlled testing	"Ask the informant"	Ask others	Observation
INTELL	Proficiency interview Self-test questionnaire	Self-score on sociolinguistic questionnaire	Interview Ibalois about the ability of Karaos to speak and understand Ibaloi	Observe Karaos in situations where Ibaloi is being spoken
ACCEPT	Two tape tests	Sociolinguistic questionnaire	Interview Ibalois about their relationships with Karaos	Observe interaction of Karaos with Ibalois

After gathering information from census data and from the estimates of knowledgeable people, we came up with the percentage breakdown of the Karao population as shown in (2).

(2) Karao population breakdown

	Young—67%		Old—33%	
	More Ed.—70%	Less Ed.—30%	More Ed.—45%	Less Ed.—55%
Male 50%	24%	10%	7%	9%
Female 50%	24%	10%	7%	9%

The second step in quota sampling is deciding the minimum number for each subgroup. We decided that we needed at least five. The table in (3) displays the number of people that we wanted to have in our sample.

(3) Sample size = 70

	Young—67%		Old—33%	
	More Ed.—70%	Less Ed.—30%	More Ed.—45%	Less Ed.—55%
Male 50%	17	7	5	6
Female 50%	17	7	5	6
Total	34	14	10	12

This method has also been called STRATIFIED RANDOM SAMPLING, although language survey testing is almost never random in the strict sense. To be completely random we would have had to find out the names of everyone in all of the subgroups and use a table of random numbers or some other technique to choose the people to test. In this case, as in many field surveys, we have what is sometimes called a PRESENTING SAMPLE (Langley 1968:49). That is, we accepted whoever presented themselves to be tested until each subgroup was completed. That did not seem to introduce a great deal of bias and the sample is random in the sense that no one was turned away because of ability or lack of ability in Ibaloi.

We also did our best to ensure that we tested people from the remote sitios as well as from the central sitios. We estimated that about half of the Karao speakers live in the central sitios which have easy access to the Ibaloi area and about half do not. The sample reflects that fact, with about half of the sample composed of residents of the more remote sitios.

A comparison of four survey tools

Several weeks were spent in formal preparation for the time of formal testing. We decided to use the following tests:

1. A sociolinguistic questionnaire to determine their language use and attitudes toward Karao and Ibaloi.
2. A self-score test as part of the questionnaire where they would estimate their own ability in the languages that they claim to know.
3. A self-test questionnaire similar to the one that Stephen Quakenbush (1986) used in the Agutaynen survey.
4. An interview conducted by two testers and scored by a third from a tape to be able to determine the subject's proficiency according to the levels[FSI levels]FSI levels.
5. A second method of scoring the interview based upon weighted factors.
6. Three tape tests; one in Karao for screening purposes, a second in Ibaloi with Karao questions, and a third more complex one in Ibaloi with questions also in Ibaloi.

Although the results of the sociolinguistic questionnaire and the weighted method of scoring the interview are interesting, they will not be discussed in this paper.

We concluded that our ability in Karao was too limited for us to do the testing ourselves and we also needed at least one native speaker of Ibaloi and two more Karao speakers who were fully bilingual in Ibaloi. Therefore we decided to hire the kind of individuals that we needed and to train them to do the testing for us. One full week was spent training these three in the various testing techniques as well as preparing the tapes for the tape tests.

The proficiency interview. Because one of our goals is to see how the results of the various methods correlate, we will begin with the proficiency interview. What we would really like to know, of course, is how each of the tests correlates with capital *P* proficiency. That is, which really comes closest to being an accurate index of reality? Unfortunately, we can never know what actual proficiency is. We can only hope that our tests of performance are coming close to revealing true competence. Most would probably agree, though, that the interview method comes closer than most others.

Most of the interviews lasted fifteen minutes, although some were a little longer. At the end of the interview the testee would take a short break while the testers filled out the score sheets (see appendix A). The form has places for scoring the interview in two different ways. The first, called a global score, is simply assigning a level of from zero to five (see Grimes 1986a for a description of the levels) with the possibility of plus (+) levels when the subject is considered better than a given level but not quite at the next. The second is called a weighting procedure. The scorers would simply choose which level from A to F best describes the testee for each of the areas of accent, grammar, vocabulary, fluency, and comprehension, based upon the criteria printed on the form.

This latter method has been developed by the Educational Testing Service. It is not recommended as the only method that should be used and was designed to be a check on the global score. That was its primary use in the Karao survey as well. It served to show up any glaring irregularities, although there were very few.

The data sheets in appendix B show the final global scores which the three testers assigned to each of the seventy interviewees. Some adjustments were made to the original scores before the scores became final, although for the most part there was a good deal of consistency. For fourteen interviews all three testers had given the same score. For thirty-two others, two scores were the same and the other was only a half level different. For nine more, all three scores were different but there was only one full level between the three. So, fifty-five out of seventy of the sets were completely acceptable for our purposes and easy to average. Of the fifteen unacceptable sets, five were where two scores were the same and the other was a full level away. These scores also could have been easily averaged if we had chosen to. For one other, two scores were the same and the other was a level-and-a-half away. For seven more, all three scores were different and there was a level-and-a-half between them. For the last two, the three scores were different and there were two levels between the three.

All fifteen of the unacceptably scored interviews were listened to again, discussed, and rescored until the scores fell into one of the acceptable categories. These final scores, then, were averaged in the following way. Where all three were the same or where two were the same and the other a half-level different, the majority score was taken. Where all three scores were different with one level between the three, the middle score was taken as the average. The breakdown of the scores is displayed in (4).

(4) Averaged interview scores

Level	Number	Percent
0+	4	5.7
1	14	20.0
1+	8	11.4
2	5	7.1
2+	7	10.0
3	11	15.7
3+	13	18.6
4	6	8.6
4+	2	2.9

Barbara Grimes (1986a:20) recommends that when testers are unsure about whether they are testing intelligibility or bilingualism, they should follow those procedures that are used for sampling bilingualism. Although other methods have been discussed, the interview method is usually perceived as being the best. It is true that the interview method has a good track record for measuring bilingual proficiency since it was developed in the early 1950's by the Foreign Service Institute. There are, however, some significant differences between the situations where it has been most effective and situations such as this one.

First of all, it has usually been used in a situation where the individual is fully aware of what is going on. Initially it was used to test prospective Foreign Service Institute (FSI) employees. Since then it has been used in a variety of other testing situations (Peace Corps, etc.), but in most cases the testees are relatively sophisticated and know that it is a test. In fact, one of the strengths of the method is that it resembles a more or less natural conversation. However, if the individual cannot really comprehend that it is a test of production ability in a second language, one can expect that normal communication patterns will emerge. If one normal pattern for some Karaos is to use their own language with Ibalois, or to switch from one to the other depending on the domain, then we can expect some of that to be taking place in the interviews. And that is exactly what happened. Obviously this makes scoring extremely difficult.

Secondly, and related to the first, the FSI interview has usually been used to assess proficiency in a second language that is not closely related to the first. In such cases the testee demonstrates the ability that he has based upon what he has learned. Things are generally quite clear-cut; to the extent that he has learned the language he is able to carry on a conversation. In the Karao situation, however, which we have called LEARNING-MODIFIED INHERENT INTELLIGIBILITY, there is a good deal more fuzziness. Everyone understands a certain amount and is able to produce somewhat similar speech due to linguistic similarity. Others have considerably improved their ability to comprehend the second language but have not been motivated to develop verbal skills. Still others have built upon the linguistic similarity and have become fully bilingual in all of the language skills. I am not convinced that the proficiency interview provides a measure which takes into account these factors.

Thirdly, the FSI interview assesses oral proficiency. It is true that comprehension is considered as one of the components, but comprehension is measured by the testee's ability to respond verbally. Under normal circumstances, the proficiency interview does not provide a completely adequate picture of intelligibility or comprehension, especially in situations where there is more intelligibility than verbal ability.

In the Karao survey we were able to recognize some of these difficulties because we lived in the area for several months before beginning the testing. Because of the relationships that we had developed, we were able to review the performance of the sample with one of the interviewers and eventually remove ten from the sample because they had demonstrated less ability than they actually possessed. No doubt the sample would have remained skewed if we had done the testing while unfamiliar with the area.

The self-score question. Little needs to be said about this method. We were interested in knowing whether people would be able to estimate their own ability in a second language. So, while administering the sociolinguistic questionnaire, we would get them to look at a simple scale marked in half points from zero to five. The tester would say something like, "Look at this scale. At one end are those who know no Ibaloi. At the other end are those who can speak it just like a native speaker. Where would you place yourself?"

Actually, we were a little surprised that people found it a relatively easy matter to point to a place on the scale that they thought represented their own ability. The results are displayed in (5).

(5) Self-score scores

Level	Number	Percent
0+	0	0.0
1	3	4.3
1+	3	4.3
2	7	10.0
2+	16	22.9
3	16	22.9
3+	18	25.6
4	7	10.0
4+	0	0.0

It would appear that people felt most comfortable with choosing a middle spot for themselves. However, the correlation between these scores and the interview scores is not entirely discouraging. Using the Pearson product-moment correlation, the result was a positive relationship (r= .74). If it were possible consistently to achieve such a close correlation, this extremely economical tool might be worth using. In a more recent survey, however, the correlation was only .43. This is probably closer to what can be expected.

The self-test questionnaire. This questionnaire (see appendix C) was modelled after the short-form questionnaire that Quakenbush (1986) used, although some of the questions were changed. A few of the questions proved to be problematic. Question 3-D, “Are you sometimes unable to finish a sentence because you don’t know how to say something in Ibaloi?”, would almost always be answered “Yes” until we explained that we were talking about leaving a sentence unfinished, not just a pause. Then they always answered “No.” Question 4-D, “Do you sometimes make mistakes when you speak Ibaloi?”, was almost always answered “Yes.”

The biggest problem, though, was that most people thought they could do just about everything in Ibaloi—even those who later showed that they could do almost nothing.

Often the tester would continue with the questions even though a person answered “No” early in the questionnaire. This revealed several anomalies. For example, on more than one occasion one of the level two questions was answered negatively but a level five question such as 5-A, "Can you use as many words in Ibaloi as in Karao?", was then answered positively.

We need to confess, though, that we were not optimistic from the beginning that this method would offer us any definitive results. For that reason we did not 'test the test' by using a longer questionnaire and finding the questions which proved to be the most effective. That probably should have been done. The results that we did obtain are displayed in (6).

(6) Self-test scores

Level	Number	Percent
0+	4	5.7
1	1	1.5
1+	3	4.3
2	1	1.4
2+	10	14.3
3	7	10.0
3+	39	55.7
4	1	1.4
4+	4	5.7

The correlation between these scores and the interview scores is a weak positive one (r= .54). When Quakenbush used a similar self-test questionnaire he noted a correlation of .56 with the interview scores. The Educational Testing Service has also experimented with a similar questionnaire and has found a correlation of about .60 with the interview (Frith 1980:20). Perhaps that is the best that can be hoped for.

All of the three tests discussed above were designed to measure bilingual proficiency. There is still some doubt whether the last two effectively do that. However, if the surveyor in a mixed situation begins with testing for bilingual proficiency, especially with tests that measure verbal ability, it is possible that the results would reveal a greater need than actually exists. The proficiency results do not show how much unmodified inherent intelligibility there is. For example, several whom we tested were unable or unwilling to use very much Ibaloi. They would have to be scored at level 0+ or 1. However, that score really does not reflect the fact that these people replied (although in Karao) to even the most complex Ibaloi questions.

The tape tests. It has become fashionable to maintain that the usual "Casad tests" are not sensitive enough to provide an adequate measure of learning-based intelligibility. One apparent limitation is that a single narrative text is likely to refer to only one domain of language use. Related to this, Grimes (1986a:16) notes:

> Merely testing for comprehension with a narrative text tests only for Level 3 bilingual comprehension or below. If all speakers chosen happen to be at Level 3 proficiency, they could all get perfect scores on the test, but that would not show that any of them could understand the more complex or abstract discourse that would be understood by a person with Level 4 proficiency. To test for Level 4, at least one more text should be included which includes discussions of problems, feelings, or fears, and the reasons behind them. It needs to be longer and grammatically more complex than the usual short text used for testing intelligibility so that the complexity of the situation can be described and understood.

It was our goal in the Karao survey to use a typical intelligibility test to see what the correlation really is to the proficiency results, and to use a second, more complex text to see what the prospects are for developing graded tape tests. However, the first thing that we found is that it is not an easy matter to obtain the kinds of texts which Grimes describes. Even in the relatively controlled situation in which we worked we were unable to get a text which was anything other than a simple narrative. Eventually we wrote a story in English which we thought met some of the requirements and had our assistants translate it into Ibaloi. Obviously, this introduced the possibility of it being unnatural Ibaloi. The point is that in many survey situations it will probably be very difficult to get the kind of texts which distinguish between level 3 and level 4.

However, is it true that the usual narrative text tests only for level 3 bilingual comprehension or below? The test results are in (7).

(7) Tape test scores

	Hometown Test	Ibaloi Test 1	Ibaloi Test 2
Number	70	70	70
Mean	98%	93%	90%
Standard deviation	4.1%	10.2%	10.6%

At first glance the difference between the means looks statistically insignificant. However, a matched pairs t-test was run on the scores which showed that the difference in the means is statistically significant at the 98% confidence level. A breakdown of the scores according to proficiency levels, as in (8), permits some interesting observations.

(8) Tape test means at each proficiency level

	Ibaloi tape 1			Ibaloi tape 2		
Level	Mean	S.D.	N	Mean	S.D.	N
0+	83.3	14.4	3	88.7	5.1	3
1	85.1	13.6	9	83.0	15.3	9
1+	85.7	8.5	4	78.5	15.8	4
2	90.0	13.6	5	85.4	12.3	5
2+	97.4	3.2	7	88.9	12.8	7
3	90.2	12.4	11	92.1	9.7	11
3+	96.9	4.3	13	93.5	5.6	13
4	97.0	3.3	6	97.3	2.6	6
4+	100.0	0.0	2	98.5	2.1	2

We notice first of all that the tape test scores of the level 0+ and level 1 people averaged almost the same for both Ibaloi tapes. In fact, when the twelve levels 0+ and 1 scores are averaged the result is 84.66 on the first Ibaloi tape and 84.42 on the second. (This might well be a close approximation of what inherent intelligibility is. It is at least a measure of minimal intelligibility.) However, the mean scores are not the same at all levels. Those with very little proficiency in Ibaloi scored about the same on both tapes and so did those with considerable proficiency, but levels 1+ to 2+ show an interesting difference. The results suggest that on the first Ibaloi tape, even a little bilingual proficiency increases the tape test scores. By level 1+ the mean score is already in the nineties. On the second tape, however, the means do not reach the nineties until level 3. From level 3 on the correlations are very close to being the same.

What conclusions, then, can we reach? First of all, it would seem to support the hypothesis that most narrative tape tests do not distinguish higher than level 3 ability. Secondly, it would suggest that our attempt to produce a more complex tape was not entirely unsuccessful. Perhaps there is hope for the development of graded tape tests.

Conclusions

Assessing learning-modified inherent intelligibility is a difficult task. No one method seems to be a panacea for the surveyor. Simple narrative tape tests, even administered to a large, stratified sample, do not provide a foolproof measure of the situation if bilingual ability is a significant factor. The usefulness of methods such as the self-test or self-score has not been adequately proven even in situations where there is a clear distinction between first and second languages. The added element of complexity when the two languages are closely related increases the likelihood that the test results would not be reliable. Nor is the proficiency interview the answer for every situation. Quite apart from the practical difficulties of finding and training testers, it needs to be used with a great deal of care if within the language group there is a variety of communication strategies: those who operate on linguistic similarity alone, passive bilinguals, and fluent bilinguals.

Perhaps we can develop graded tape tests. However, is it possible to obtain a level 4 text? How do we know that it distinguishes level 4 proficiency? Do we then look at the individual scores on the test or do we look at the mean and the spread? If we consider the individual scores, what score is adequate to show level 4 proficiency? 100%? If we pay attention to the mean, do we try to set a threshold for the most complex text? Is it possible to say that if there is a mean of 85%, for example, on the most difficult of the graded texts that there is adequate proficiency within the group? That would seem to obscure the fact that a certain number may not have adequate proficiency.

Until a single test is developed which proves effectively to measure learning-modified inherent intelligibility, perhaps the safest approach is to administer two tests: one which is designed for inherent intelligibility, and one designed for bilingualism. Not only will it provide a clearer picture of the specific survey need, it will very likely give us data which will help us to understand better the relationship of intelligibility to proficiency.

[blank]

Part III: Windows on Bilingualism

[blank]

Sociolinguistic Community Profiles

Calvin R. Rensch

Some sociolinguistic factors which we study are not evenly distributed throughout a community. Levels of proficiency in a second language, for example, are usually much higher in some sociolinguistic groups in the community than in others. This is sometimes also true of patterns of language use and language attitudes.

In many communities the more educated people and those who travel more show higher levels of second-language proficiency. It is of interest then to study the distribution of such factors in the society in order to learn which factors correlate with the higher levels of second-language proficiency. One implication is that a change in the frequency of that factor is likely to affect the progress of bilingualism.

Furthermore, the number of people associated with the different factors varies from factor to factor. Men and women are likely to be equal-sized groups. On the other hand, younger educated men often constitute a much larger group than older educated men. Men who are educated may be much more numerous than women who are educated.

When selecting subjects for studying multilingualism it is important to draw them from various sociolinguistic groups even though members of the different groups may not be equally available. Educated young men, for example, are frequently quite willing to serve as test subjects whereas older women, usually uneducated, are not. So, if it proves impossible to draw subjects from the various sociolinguistic groups in numbers proportionate to the size of their group, a sociolinguistic profile permits the scores of each group to be weighted in proportion to the size of that group. Let us assume, for example, that fifty percent of the subjects were drawn from the group of young educated males. Let us further assume that they performed quite well on a bilingualism test. These data need to be placed in the perspective that the sociolinguistic group of young educated males constitute only, say, twelve percent of the population. It needs to be recognized that that high-performing group constitutes an influential, but modest-sized, segment of the population.

So, it is very useful to conduct a census which leads to a sociolinguistic profile of the community for at least three reasons: (1) It is helpful in studying the association of certain factors in the population with high levels of second-language proficiency; (2) it provides information about the relative numbers of people in the various sociolinguistic groups of the community; and (3) it provides guidance in the selection of test subjects so that the sample selected will be as representative as possible.

In such a census a representative of each household is interviewed to gather information about the members of that household. Categories suggested by Frank Blair (1990:40) are as follows:

1. number/name of interviewee
2. date of interview
3. location of interview
4. name
5. age
6. sex
7. education
8. occupation
9. previous occupation(s)
10. religion
11. place of current residence
12. place(s) of previous residence
13. caste or social class
14. clan or moiety
15. marital status
16. number of children
17. number of people in household
18. mother tongue
19. other tongue(s)
20. literate and in which script(s)

If the community is not large, it is wise to collect such information concerning each household of the community. If the community is very large, it may be necessary to select simply a large sampling of the households of the community. Since factors such as availability of educational opportunity vary from one community to another, it is helpful to conduct a census in more than one village of the language area in order to avoid the skewing that might come from information from a single village. For example, in the Hinko study in northern Pakistan, census information was collected in four quite different villages. The four villages were found to vary with respect to access to education, especially for girls, and the particular language(s) of wider communication used in the vicinity. Several of these community variables were found to be related to differing patterns of language use and differing levels of second language proficiency in those communities.

On the basis of the census data collected, a sociolinguistic profile of the community can be prepared, in which the various sociolinguistic variables thought to be significant form the categories of the display.

A sociolinguistic profile of men in a Torwali-speaking community of northern Pakistan is provided as an·illustration. In this profile age ranges and levels of education in the male population are the factors which define the sociolinguistic groups.

(1)

	more educated		less educated		both education groups	
	n	% of men	n	% of men	n	% of men
15–24	45	24.9%	28	15.5%	73	40.3%
25–39	27	14.9	33	18.2	60	33.1
40+	7	3.9	41	22.7	48	26.5
all age groups	79	43.7%	102	56.4%	181	100.0%

Some of the significance of the sociolinguistic groups of the Torwali-speaking community can be seen from the fact that the performance of subjects from the various groups varied considerably. The average scores of twenty more-educated and twenty less-educated Torwali male subjects in three age groups are shown in (2), which gives results on the recorded-text test in Urdu.

(2)

	more educated	less educated	both groups
15–24	86.9 %	56.3%	75.2%
25–39	90.0	64.3	75.0
40+	100.0	88.0	91.4
all age groups	89.0%	67.0%	78.0%

From the scores partitioned in this way it can be seen that education is a very powerful factor in learning Urdu. This is not surprising since school is the primary context in which Urdu is learned in the Torwali-speaking area. This is in contrast to Pashto, which is learned more informally. Performance on tests in Pashto does not show any correlation with levels of education.

However, it can also be observed that progressing age leads to increased proficiency in Urdu. The older men scored noticeably higher than the younger or middle-aged men, suggesting that later in life men use the Urdu that they learn in school and thereby become more proficient. If the performance of the subjects had not been compared to the sociolinguistic profile of the Torwali-speaking community, the factors which lead to proficiency in Urdu (and Pashto) among Torwali men would not have been evident.

Notes on Oral Proficiency Testing (SLOPE)

Barbara F. Grimes

A modification of the second language oral proficiency testing procedure developed by the U.S. Foreign Service Institute (FSI) has been made for use in preliterate societies and where the linguistic investigator may not speak the first or second language of the group being tested. The method, Second Language Oral Proficiency' Evaluation (SLOPE), is described in detail in a syllabus in Notes on Linguistics 40:24–54 (Summer Institute of Linguistics 1987). Additional considerations concerning testing and sampling for bilingual proficiency in a community are discussed in a paper in the same volume (Grimes, 1987b).

A trial testing period of about two years (April 1987 to May 1989) was agreed on, after which the method would be evaluated by those who used it, and further modifications and recommendations made. In the meantime, the method has been tested in Senegal and South Asia, and certain aspects of the method have turned out to need highlighting or further clarification. Other questions have arisen about the feasibility of using SLOPE versus other methods. It is the purpose of this paper to discuss some of those aspects.

The method: A brief summary

The SLOPE testing team includes (1) a tester, who is a native speaker of the subjects' second language; (2) a linguist, who is trained in the procedure, knowledgeable about the second language, in charge of the test, responsible to see that an adequate sample of the subject's speech is obtained in each part of the test, and the one who evaluates each subject's proficiency;

and (3) a first language assistant, who is a native speaker of the subject's first language, who assists in giving instructions to the subjects, and passes on the subject's understanding of the second language in the second part of the test. The addition of the assistant to the regular FSI testing team enables testing to be done in preliterate situations, and where the linguist does not know the first language of the subject.

The test includes the same three parts as the FSI<D>[FSI]FSI prototype: (1) a warm-up conversation in which the subject is put at ease by conversing informally in his second language with the tester on social and personal matters, (2) the subject uses the second language to get information from the tester on a different topic than was used in the warm-up, and then passes on what was said to the first language assistant in his first language, and (3) the subject gives information to the tester on a different topic than was used in the earlier parts of the test, and the tester questions him about what he said.

The SLOPE syllabus includes detailed information about preliminary research, training the tester and the first language assistant, suggested topics for each part of the test, procedures for each part, scoring, and other considerations.

Standardization and certification. FSI maintains careful control over its testing so that different linguists (whom they call 'Examiners') evaluate the proficiency levels of subjects consistently in the same second language, and so that the levels are applied consistently between different second languages. FSI keeps an archive of tapes representing different proficiency levels for each second language they are concerned with.

FSI also has recertification checks for linguists at regular intervals to be sure they maintain standards, by checking their evaluations against the archival tapes mentioned above. Without this recertification procedure, there is a tendency for evaluations to become inconsistent over a period of time by the same linguist, between different linguists, and with respect to different second languages.

Preliminary research. Presumably some sociolinguistic knowledge of a situation already exists in order for the decision to have been made to do SLOPE testing. The more a linguist knows about the two languages and sociolinguistic situations, the more accurate the evaluations will be. If he can speak one or both languages with some degree of proficiency, so much the better. If bilingualism seems to be high and widespread, it is worth putting in the preliminary time needed to get accurate results. It may even be necessary to delay testing until the linguist has more knowledge of the language.

If there is a standard dialect of the language, it is important for that dialect to be identified ahead of time, and that the tester speak that dialect as mother tongue. For situations in which there are divergent nonstandard dialects, different ones may function as second language for different language groups (1986b), and each should be tested separately as second language.

Adequate training of participants. In order to carry out SLOPE testing, a linguist should be trained adequately in the procedures, have enough guided practice in actual testing in the linguist role, and follow the instructions in the SLOPE syllabus carefully. He may need to function in the linguist role in five or more interviews in each new language combination, or until his evaluations agree with those of archival tapes, before he can be confident of giving accurate evaluations. This is especially true if he has only minimal knowledge of the languages involved.

The tester and first language assistant need to be adequately trained in their roles ahead of time so they are aware of what they need to do and can feel comfortable doing it. They too may need to participate in at least five interviews before they are able to do valid testing.

When the linguist does not know the second language, it is especially important to have a well-trained tester and assistant. If possible, it is best to work with the same team of tester and assistant for all tests in a given first- and second-language combination, because they will have learned what is needed and how to get that information, as well as what characteristics of each subject's speech should be noted and reported back to the linguist.

The tester and assistant need to be fluent enough in the language they use for communicating with the linguist that meanings can be conveyed accurately.

Role of the assistant. The assistant explains instructions to the subject in the first language. He uses only the first language with the subject. From one point of view, it is better to work with an assistant who is more fluent than the subject in the second language. Then he can more adequately report lack of agreement between what the tester said in the second language and what the subject reported back to him in the first language during the getting-information part of the test (James, Masland, and Rand 1987:B2). If the assistant is well trained, he will not substitute his understanding of the second part for the subject's lack of understanding. If he is inadequately trained, however, it is difficult to tell which information went through the subject and which through the assistant without checking back carefully the recording of that portion of the test with the assistant and tester. For that reason, some investigators think it would be preferable to have an assistant who does not know the second language at all. It might not be possible, however, to locate such a person, one who would otherwise be able to function well as assistant.

The assistant can help the subject by suggesting questions to ask during the getting-information part of the test, if the subject needs or wants help. Some subjects need or want more help than others. It is probably best for the linguist or tester not to write down suggested questions even if the assistant is literate, because that may take away from the naturalness of the interview. The naturalness of a conversational format was mentioned both by James, Masland, and Rand (1987:12–13) and by Quakenbush (1986 and 1988:7) as being an advantage of the interview method. The assistant and tester, however, may wish to take notes during the interview.

Pauses while getting information. The tester needs to pause frequently enough while the subject is getting information from him during the second part of the interview that the subject can explain the meaning of what he said to the assistant in the first language. If the tester does not pause often enough, the subject may forget part of what was said, and the test then turns into one of memory rather than of comprehension. If that happens, the linguist (or subject) should interrupt after two or three sentences, to ask permission for the subject to explain what was said to the assistant.

Topics for each part of the interview. The topics given in the syllabus are intended to be suggestions only. Some might be used in other parts of the test instead, or be substituted by more appropriate topics. One consideration in choice of topic is that the topics in the three parts of the test be from different sociolinguistic domains. Another is that the subject be adequately tested in his understanding of new information by having the topic in the getting information, or second part of the test, be one not normally discussed in the first language, and the topic in the giving information, or third part of the test, be one not normally discussed in the second language.

The topic should not be so complex and tightly packed with new information, however, that the tester has difficulty handling it. If possible, it is good to choose topics that will "challenge but not overwhelm the subject," but that is not always possible to gauge ahead of time (James, Masland, and Rand 1987:B2).

Checking back with the assistant. It is important for the linguist to check back on the getting information part of the test with the assistant, after the test is finished. This is time-consuming, but necessary when the linguist does not know the first language of the subject. By playing the recording back between pauses, and asking the assistant to explain what was said, he can get a fairly accurate idea as to how well the subject understood what the tester said, and what the subject omitted. If the linguist does not know the second language, the tester should also be in on this review, to better compare the meanings given in the two languages. It may not be necessary to check in detail through the entire getting information part if a clear picture emerges sooner.

Checking back with the tester. If the linguist does not know the subject's second language, it is also important for him to check back with the tester on the first and third parts after the test. He should ask for specific examples of correct and incorrect language usage, and should take notes on what the tester says. He needs to find out where the subject understood or misunderstood the tester, where he did or did not express himself clearly, where his presentation in the third part was well organized or where ideas were strung together without signalling the relationships, where his speech was natural or awkward, what word choices were natural or imprecise, what mistakes or correct combinations he made in grammar, what time sequences were correct oF unnatural, what unfamiliar information was handled well or poorly. The linguist should refer to the descriptions of the five factors (grammatical structure, lexicalization, discourse competence, comprehension, fluency) given in the syllabus in his questioning of the tester. His evaluation will be more accurate and less time-consuming if he becomes familiar with, and even memorizes, those descriptions.

Time required for testing. James, Masland, and Rand (1987:13) found that they were able to do three or four tests a day before fatigue sets in. Each test required at least one hour.

In the beginning, tests may take more time than that, but it is reasonable that with more experience in working with the same testing team and familiarity with the languages, the time needed for checking back with tester and assistant can be reduced.

Validity and reliability of SLOPE

For a test to be valid it needs to actually test what it claims to test. For it to be reliable it needs to give comparable results when repeated by the same or different investigators in the same or different situations (Fasold 1984:90). That means that tests that test as directly as possible the characteristic in question are likely to be more valid than those that try to test indirectly; that is, those that test some other characteristic that is supposed to have some relationship to the factor in question. If a test is evaluating a different factor, as in an indirect test, that factor should be shown by independent means to have some psycholinguistic or logical relationship to the factor in question. Extensive, repeated testing should demonstrate the validity of the indirect test in different situations. It needs to be calibrated against a more direct test and to agree with independent knowledge of the situation.

SLOPE needs to be validated by field testing just like any other method. It does have several advantages to start with, however. It is based closely on the procedures developed and validated by FSI for over forty years in many languages with thousands of subjects and by many different investigators all over the world. SLOPE is modified from the FSI method only as much as is necessary to enable it to be administered in preliterate societies, and in situations in which the linguist may not know the first or second languages of the subjects. The modifications were made under the guidance of Dr. Thea Bruhn, head of testing for FSI. Dr. Bruhn has training in linguistics, sociolinguistics, language teaching, testing, and experience in several societies, including minority language situations.

Validity and reliability of other methods

Other methods have been proposed or tried for evaluating bilingual proficiency in minority language situations. They still need extensive testing in many different societies and with many speakers before they can be shown to be valid and reliable. All of them known to this author test something other than bilingual proficiency, are assumed to have a definite relationship to bilingual proficiency, or they test only a small part of the different kinds of factors bilingual proficiency involves, or they are largely based on subjective opinions.

Sentence repetition tests assume that there is a consistent relationship between the ability to repeat something and the degree of bilingual understanding one has. It is, however, incapable of evaluating lexical discrimination, discourse comprehension, productive use, and situational appropriateness, at the least. All are important factors in bilingual proficiency that are covered explicitly in SLOPE.

Testing understanding of recorded texts assumes that understanding alone is a predictable indication of total bilingual competence, and that one or a few short narrative texts are an adequate sample of bilingual comprehension and degree of proficiency (Grimes 1987b:14–15). A valuable Senegal survey (James, Masland, and Rand 1987:11–14) explored this.

In that survey, recorded text testing turned out to have no significant correlation to SLOPE results; those who scored a level 2 on SLOPE ranged from 30% to 95% on recorded text, 2+ on SLOPE from 45% to 95% on recorded text, 3 on SLOPE from 65% to 100% on recorded text, 3+ on SLOPE from 45% to 100% on recorded text, 4 on SLOPE from 55% to 65% on recorded text, and those with 4+ on SLOPE from 75% to 95% on recorded text. The correlation of recorded text testing with SLOPE testing was only 0.15, too low to take seriously.

Reading tests assume that there is a predictable relationship between reading comprehension and bilingual proficiency. FSI has found no regular correlation.

Translation tests assume that there is a regular relationship between the ability to translate and bilingual proficiency. However, the two are different abilities (Grimes 1987b:17). Most bilinguals in the world learn their second language outside the classroom, and find it hard to translate well without the training to do so, though a few people who translate well spontaneously turn up from time to time.

Questionnaires administered either orally or in writing often ask for opinions or self-reports, and are widely recognized by sociolinguists to be subjective and unreliable. Fasold says, "As we know, self-reported data are often of questionable validity" (1984:147). Quakenbush found that among the Agutaynen of the central Philippines (1988:7–8) the self-report questionnaire was not reliable when compared with the results of an FSI type oral proficiency interview. FSI found their own earlier self-check list to be unreliable when compared with the results of oral proficiency interviews (Grimes 1987b:16).

Observation as a method of evaluating bilingualism in a community depends on the training, experience, and sensitivity of the observer (Grimes 1987b:16–17). However, it cannot be used to determine degree of proficiency unless the observer is fluent in the second language of the community, and unless he or she investigates bilingual behavior with the same attention to detail that SLOPE requires; impressions can be misleading.

If indirect methods such as sentence repetition testing or recorded text testing are proposed as a screening device to find language groups that are not highly bilingual, those tests would still need to be validated. They would need to be calibrated against SLOPE testing in enough new combinations of first and second languages and cultural settings, because we cannot know that the dynamic relationships among the various linguistic, psycholinguistic, and sociolinguistic factors are regular from situation to situation.

Shortcuts in the SLOPE method or any of the other methods mentioned above make that test less reliable unless the modified test is also validated.

Sampling and accuracy of results

All of the methods discussed above, including SLOPE, require adequate sampling of the population both as regards probing the relevant factors for each situation, and selecting adequate numbers within each relevant sample group (Grimes 1987b:8–12). Bilingualism is rarely uniform within a society, but is ordinarily differentially distributed by age, sex, region, education, and possibly other factors. Adequate sampling is needed even if one of the indirect methods is proposed only as a device to screen out less bilingual situations.

Our goal is to arrive at an evaluation of a situation that is as close to the reality of the actual situation as possible. Bilingual proficiency in a minority language situation is complex. Accuracy of results can best be assured by using methods that have been validated and have been shown to be reliable. Indirect methods which have not been validated should not be substituted for direct methods solely because they take less time. Although time is important, accuracy, validity, and reliability should never be compromised by time considerations.

Surveying Language Proficiency

John Stephen Quakenbush[1]

Prior to discussing methods of surveying language proficiency it is helpful to define two key terms: survey and proficiency. The first term, survey, is the more easily defined. Cooper characterizes survey research as "research carried out with respect to an entire population, whether as small as a hundred neighboring households... or as large as a nation... " (Cooper 1975:29). A survey of language proficiency fits into the larger category of sociolinguistic survey, which Cooper also defines succinctly as an endeavor which "gather(s) information about the social organization of language behavior and behavior toward language in specified populations" (29). Language behavior includes such phenomena as proficiency, acquisition, and usage. Behavior toward language includes both attitudinal and implementational behavior, the latter being observable, the former only inferable. Sociolinguistic surveys have largely been motivated by the need for information of language policy makers and program planners. They can also be justified on the basis of their contribution to more theoretical concerns involving the interaction of language and society.

The second key term here, proficiency, is the more difficult term to define. A working definition for the purposes of this study has been "the

[1]This paper was originally presented at the Fifth International Congress on Austronesian Linguistics, January 1988, Auckland, New Zealand, and is to appear in *VICAL2 Western Austronesian and Contact Languages*, Ray Harlow, editor. The volume itself is to be a *Te Reo* Special Publication of the Linguistic Society of New Zealand. I would like to express my sincere thanks to Ray Harlow, the editor of that volume, and to the Linguistic Society of New Zealand for their permission to reprint this paper here [editor's note].

degree to which a language can be used successfully in face to face interaction." Proficiency in this sense involves primarily the skills of listening and speaking. Although it is possible to consider degrees of proficiency in the other two major skill areas of reading and writing, there are many situations for which these literacy skills are not relevant.

How have language surveyors traditionally measured language proficiency? In the absence of any standard method, a variety of techniques have been employed. The most obvious distinction between survey techniques has been between those methods that gather reports of estimated proficiency versus those that actually administer some type of performance test. The report method is by far the most common, with reports usually gathered by means of a written questionnaire. A further distinction can be made in the report method between those that gather a respondent's estimate of the proficiency of others versus those that collect self-report data. Self-report data are the kind most commonly collected, but in some instances it is helpful to gain an individual's estimate of the second language proficiency of an overall community.[2]

Self-report techniques usually ask respondents to rate their proficiency according to predetermined levels, such as 'fluent, fairly well, a little' or 'very proficient, adequately proficient, hardly proficient, not proficient'. The number of levels distinguished, as well as the descriptions of these levels, varies from survey to survey.[3] One large scale survey employed a slightly different method by asking respondents if they could handle a particular language successfully in a series of thirteen situations, each new situation supposedly more difficult than the preceding one (Polomé and Hill 1980:116). This yielded an oral proficiency score for an individual of anywhere from 0 to 13.

The second major kind of technique employed by language surveyors to measure language proficiency entails some sort of direct testing. Some surveys in effect test only comprehension, or listening proficiency, as they ask the respondent for some sort of response to a verbal stimulus.[4] The most fully developed survey procedure for testing comprehension is the dialect intelligibility test, described in Casad 1974. This type of test consists of a short, tape recorded story, followed by a series of simple content

[2] For an example of a survey where informants were asked to estimate the language abilities of a surrounding community, see Ladefoged, Glick, and Criper (1968:53).

[3] For various scales of proficiency employed in self-report surveys, see Aguilana 1978, Bautista et al. 1977, Cooper and King 1976, Whiteley 1974, and Olonan 1978.

[4] See Serpell 1978 and Barcelona 1977 for elicited nonverbal responses. Casad 1974 and Kashoki 1978 describe techniques requiring oral and written responses in the mother tongue to aural texts in another language variety.

questions to be answered orally. The questions are sometimes interspersed in the body of the text, and are usually asked in the respondent's mother tongue. The resulting scores are taken as an indication of the extent to which dialect B is comprehensible to speakers of dialect A. This type of testing has proved very useful in situations where the key factor is inherent intelligibility due to linguistic similarity, as opposed to learned bilingualism gained through social contact.[5]

Other surveys of language proficiency have concentrated on productive capacity.[6] Most of these tests have depended on single word or otherwise minimal responses. Apparently no large-scale survey has ever tested actual conversational ability, presumably because of the many difficulties and indeterminacies involved. There is, however, a recognized method for measuring overall oral proficiency, developed by the United States Foreign Service Institute and related agencies.[7]

The oral proficiency interview developed by the FSI consists of a trained interviewer conducting a more or less natural, yet highly structured, conversation with a respondent in an attempt to discover that respondent's overall strengths and weaknesses in a given language. Factors explicitly evaluated include accent, comprehension, fluency, grammar and vocabulary. Different factors assume prominence at different levels of proficiency. Possible levels range from 0–5, with 0 being no knowledge of the language and 5 being educated, native-speaker proficiency. Levels 0–4 may be further refined by the addition of a '+' or half point. The FSI interview procedure is a complex one. The logistics of training interviewers and the time required for conducting the interviews, quite apart from the task of convincing the general public to submit to being evaluated, will probably preclude its use in any large scale survey. The main point of this paper, however, is that the FSI method can be adapted successfully for use on the community level, and can be further adapted as a useful tool for gathering self-report data from larger groups of respondents.

[5] See Grimes 1986a for more complete consideration of the differences between inherent intelligibility and learned bilingualism.

[6] Serpell 1978, De Gaay Fortman 1978, and Bautista et al. 1977 utilized visual stimuli to elicit linguistic responses.

[7] See Adams and Frith 1979 for a detailed explanation of this method. The oral proficiency interview has since been adapted by the Educational Testing Service (ETS) for use by the Peace Corps in evaluating the language proficiency of volunteers, and more recently by the American Council of Teachers of Foreign Languages (ACTFL) for use in an academic setting.

A sample study

The Agutaynen sociolinguistic survey was conducted in 1984–85 in Palawan, Philippines under the auspices of the Summer Institute of Linguistics.[8] One of its main purposes was to investigate the extent to which mother-tongue Agutaynen speakers could use the second languages of Cuyonon, Tagalog and English. Over 200 Agutaynens were interviewed in three municipalities of northern Palawan province. All interviews were conducted by the present researcher exclusively in the Agutaynen language. Responses were tape recorded on a small, hand-held audio recorder.

The section of the interview concentrating on language proficiency consisted of a set of seventeen yes-no questions (see appendix E). These questions involve specific language skills, each one associated with a particular level of proficiency, as defined by FSI. A level 1 question, for example, is "Can you understand and respond correctly to questions about where you are from, if you are married, your work, date and place of birth?" A level 5 question is "Do you know as many words in X as you do in Agutaynen?" This particular set of questions was derived from a longer set adapted from FSI materials by Barbara F. Grimes, and then further adapted to fit local circumstances. The original, longer set contained 37 questions. It was found during pilot testing, however, that this was far too many to maintain the interest of the interviewee. Many of the questions also seemed singularly unreliable in that respondents invariably answered them positively. For these reasons, the number of questions was reduced to 17. One of the criteria for selecting a question for the shorter version was that it be answered negatively at some time during the pilot testing. Another criterion was clarity. The shortened set worked smoothly.

In order for an interviewee to rate a certain level of proficiency, he or she had to answer positively all questions for that level. If a respondent could also answer positively two of the questions at the next level, a '+' was assigned.[9] During the actual interviews an attempt was made to maintain the atmosphere of a naturally occurring conversation. Many questions were asked from memory. Nonverbal or paralinguistic cues were also taken into consideration. For example, an elderly woman obviously uncomfortable in discussing her Tagalog proficiency was not asked more than the most basic questions for that language. On the other hand, statements by the respondent to the effect that his or her proficiency was very high in a given language precluded asking the most basic questions for that language. No attempt was made to administer the questions in

[8] See Quakenbush 1986.

[9] 'No' was considered a 'positive' answer for Questions 4-D and 5-D.

exactly the same order for each respondent, although questions did generally progress from easier to more difficult. The end result of this flexibility in the order and number of questions was a more comfortable interview for the respondent, and it would be hoped, a rating of proficiency that was correspondingly more valid.

This self-report method for measuring language proficiency was employed for the Agutaynen survey as a whole. It was considered superior to previous self-report methods for two reasons. First, it asked about specific language skills, rather than for an abstract appraisal of global language ability. This allowed respondents to give simple yes-no answers while focusing on specific behaviors, rather than forcing them to give self-evaluations in terms that could be more directly linked to core values and self-esteem. Secondly, this particular set of questions was based on a scale of proficiency that is still being found useful after years of sustained, careful scrutiny by professional language testers and those they evaluate. Still, the Agutaynen survey was on new ground. This particular method had never been tried in a community-wide survey. Therefore, some sort of test of the method's validity was desirable. It was for this purpose that a separate test was carried out in another Agutaynen community in Brooke's Point, Palawan, subsequent to the main survey.

The Brooke's Point test consisted of assigning proficiency ratings for a sample of 40 individuals by two methods—the self-report method utilized in the larger Agutaynen survey, and an actual oral proficiency interview conducted in the field. Assuming that the two methods were language independent, only the Cuyonon language was used for the Brooke's Point test. Two language evaluators were trained in the technique of the oral interview, specifically for this purpose. One interviewer was a 49-year-old woman, the other a 35-year-old man. Both were native Cuyonon speakers, college-educated, and elementary school teachers by profession. The combination of both sexes and different ages was part of a deliberate effort to insure that a broad range of respondents would be comfortable in being interviewed by this team.

The general sequence followed for an interview was for the present researcher to first interview the respondent according to the self-report method, out of hearing of the other two language evaluators. Certain biographical information was also collected at this time. One of the language evaluators would then interview the same respondent, with the other evaluator an active observer. Afterwards, the evaluators individually assigned proficiency ratings without discussing the respondent's performance. All interviews were taped so that any serious differences in ratings could be discussed later. In the end, none of the evaluations differed more than one level. In these cases, an average score was taken as the final rating. In instances where the evaluators disagreed by only a half point, the lower score was chosen.

For the 40 interviews conducted, 8 of the ratings varied by one point, 5 by a half point, and 27 were exactly the same (see appendix F). The evaluators' ratings came more in line with each other as time went on. Had the interviewers been more experienced at the start, perhaps their scores would have been in even closer agreement. That the evaluations of novice interviewers agreed as much as they did is strong evidence for the reliability of the oral proficiency interview.

When the direct test ratings are compared with the self-report ratings, the results are favorable, if not overwhelmingly so. The table in (1) illustrates the two kinds of scores compared. Only 4 scores were exactly the same for the two methods. An additional 23 were within a half point. Four more were one point apart. In all, 31 out of the 40 self-report scores could be considered reasonably accurate (one point or less difference).

The mean and standard deviation of the two sets of scores are quite similar. The positive correlation between the two sets of scores, however, is only moderate.[10] The table in (2) lists the mean, standard deviation and Pearson product-moment correlation for the two sets of scores.

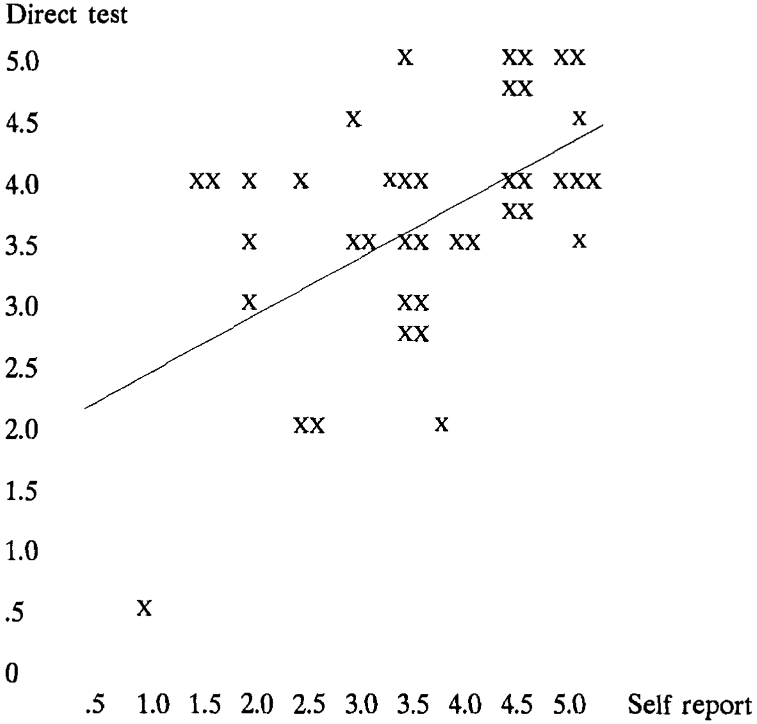
The moderate correlation may be spuriously low due to the restricted range of proficiency scores represented. A well-distributed sample of 40 proficiency scores would contain approximately 20 scores of 3.0 and above and 20 scores of 2.5 and below. As can be seen from (1), however, 31 of the 40 self-report scores are 3.0 and above, while 36 of the direct test scores are in the same category. Had the proficiency scores been distributed more evenly along the continuum, the positive correlation between the two methods of evaluation may have proved to be stronger.

The self-report method did not consistently yield higher or lower scores than the direct test method. To what extent, then, could the effects of the unreasonably low and unreasonably high self-report scores cancel each other out? The table in (3) shows that the differences are almost evenly split between scores that are too low and scores that are too high. This would minimize the importance of individual differences among a larger sample.

[10] Guilford (1956:145) gives the following interpretation system for measurement of correlation:

0.01–0.20	slight, almost negligible relationship
0.20–0.40	low correlation; definite but small relationship
0.40–0.70	moderate correlation; substantial relationship
0.70–0.90	high correlation; marked relationship
0.90–0.99	very high correlation; very dependable relationship

(1) Scatter diagram of Brooke's Point test scores. Regression line plotted.



(2) Descriptive statistics on Brooke's Point test scores.

	Self report	Direct test
Mean	3.60	3.74
Standard deviation	1.10	.95

Pearson product-moment correlation r = .56

A total of 9 out of 40 self-report ratings were off by 1.5 points or more. These 9 respondents represent a variety of ages and a range of educational and occupational backgrounds. There is, however, one striking similarity—7 of these 9 respondents were women. It would seem then, that Agutaynen women tend to understate their language proficiency under certain circumstances (6 of the 7 were understatements). The fact that they were being interviewed by an American male researcher may have been enough to produce this effect. At any rate, these women did not understate their actual language performance in conversation with the native Cuyonon speakers.

(3) Summary of differences between ratings obtained by self-report and direct-test methods.

Degree of difference	Self report lower	Self report higher	Total
0 points	—	—	4
.5	9	14	23
1.0	1	3	4
1.5	4	2	6
2.0	1	—	1
2.5	2	—	2
	17	19	40

Evaluation of case study

It is helpful to evaluate the Brooke's Point test from two perspectives. First, why did the direct-test method work as well as it did? Second, why did the self-report method not work any better than it did? We will then be ready to consider the implications of this test for future surveys concerned with the measurement of language proficiency.

The oral proficiency interview based on the FSI procedure worked well in the Brooke's Point test. This is somewhat surprising considering it was never intended for evaluating entire communities of sometimes marginally literate speakers. The FSI method was developed as a test for highly-educated individuals in the context of an intensive foreign language study program. Being interviewed is not optional for these individuals, but mandatory. Achieving a certain minimal rating is important to their careers. In Brooke's Point, in contrast, respondents had little obligation to submit to being interviewed. They were not involved in a formal language study program. Indeed, some had little experience in formal educational settings of any kind. Why, then, did the technique work? The answers to this question lie in the technique itself, and in the nature of the Cuyonon language evaluators and the Agutaynen community of Brooke's Point, Palawan.

The success of the oral proficiency interview can be attributed primarily to its natural and adaptable format. Although it may serve as a test for proficiency, on the surface it appears to be a natural communication event where information is conveyed between speakers in a socially meaningful way. It does not require a recitation of facts about language, a list of forms in a language, or answers to a series of multiple choice questions. The conversational format of the interview was such a strong factor that it apparently overshadowed any resemblance to a test situation. The actual content of the interview may vary greatly from individual to individual, or from context to context. For example, whereas a foreign service officer may be asked to describe a political process, an Agutaynen farmer may be asked to describe a rice harvest.[11]

The success of the direct-test method in Brooke's Point also was due in great part to the personal characteristics of both the evaluators and the respondents. The Cuyonon language evaluators were willing to help, well educated, quick learners, good conversationalists, and friendly and unintimidating individuals. The Agutaynen respondents, on the other hand, were open to talking to outsiders—especially when one of those outsiders embodied the fascinating composite of an Agutaynen-speaking American. As a whole, they were also familiar with the idea of 'survey' and 'school project' (terms used to describe the present research) and were very willing to cooperate. Another factor which possibly contributed to the success of the Brooke's Point test was that the survey team had numerous personal acquaintances either in the Brooke's Point community or with relatives of community members. The surveyors' presence was further legitimized by two local guides who had been appointed by the chief political leader of the community to accompany the survey team.

With all of these positive aspects of the Brooke's Point situation, why did the self-report method not work any better than it did? Most likely, the main reason is in the very nature of self-report data. In reporting one's own abilities, concern for presentation of self may override concern for accuracy in either direction. That is, a respondent may overstate or understate an ability. In Brooke's Point, the majority of those who gave seriously misleading responses were understating their abilities, presumably in the interest of humility, but perhaps also in fear of being 'put to the test' and found wanting. It may also be the case that the correlation between self-report and direct-test methods would have been stronger had the sample not been skewed toward the higher ratings. In any case, 31 of the 40 self-report scores were accurate within one level, assuming the direct-test method yielded an 'accurate' standard for comparison.

Conclusions

What conclusions can be drawn from the above comparison of two methods for surveying language proficiency? First of all, it is evident that

[11] See Quakenbush (1986:277–287) for the materials used in training the Cuyonon language evaluators. Some minor adaptations in the procedure were made in the interest of cultural relevance.

assigning proficiency levels to individuals is not an exact science, no matter how these ratings are obtained. Proficiency data, therefore, and especially self-report data, must not be interpreted rigidly. Rather, they must be seen as indications of general trends. To the extent that a measure of proficiency is simply imprecise, it may reasonably be hoped that those scores which are slightly high will offset those scores which are slightly low, at least in part. Scores that are seriously off, however, will less likely cancel each other out. In the Agutaynen survey, for example, there was an apparent tendency for a proportion of women's self-report scores to be seriously underestimated. This leads to the second point, that self-report data on proficiency ideally will be interpreted in light of at least a subsample of direct testing measures.

The purpose and extent of a language survey will ultimately determine whether it is more beneficial to rely on a self report or direct measure of language proficiency. In the Brooke's Point test, the purpose was to assign proficiency ratings for 40 individuals in one second language. The self-report method took five minutes or less per respondent to gain the necessary information. The direct method generally took a manageable, but much longer, fifteen to twenty minutes per respondent. The overall Agutaynen survey, in contrast to the Brooke's Point test, examined proficiency for over 200 respondents in three languages. It would have been impractical, to say the least, to attempt direct testing as the sole method in such a survey. Depending on the purpose and extent of a survey, it may be advisable to sacrifice some precision in the interest of time, effort and expense. Nevertheless, future language surveyors who are concerned with the more precise measurement of language proficiency would be best advised to at least attempt the more time-consuming direct interview method when this is possible. Many circumstances can work against its successful utilization in the community, but as the Brooke's Point test demonstrates, it can also work surprisingly well.

Regardless of the particular instrument used in a language survey, the Agutaynen example demonstrates that the FSI levels of proficiency can provide a meaningful, standard framework for eliciting and interpreting degrees of oral language proficiency. The use of the FSI scale of proficiency should be encouraged in future language surveys, not only to ensure more comparable, and comprehensible, results, but also to test the usefulness of this scale in a broad range of speech communities.[12]

[12] Barbara F. Grimes, editor of Ethnologue, is compiling proficiency profiles on minority language communities using the FSI scale (see Grimes 1984b). Frank Blair (personal communication) has expressed reservations about the applicability of such a scale to nonliterate societies.

Language Use and Second-Language Proficiency

Calvin R. Rensch

In most of Asia multilingualism is a very significant factor in the sociolinguistic picture. In multilingual areas in order to make sound decisions we must have reliable information about levels of proficiency in the one or more languages of wider communication current in the area. Yet, collecting such information can be a time-consuming process since usually data must be collected from a wide variety of subjects.

Therefore, the South Asia survey team has been trying to identify situations in which extensive multilingualism testing will not be necessary because of patterns of language use. In recent meetings of the team a hypothesis was developed which we will be testing in fieldwork. We present it here for discussion and so that others can test it in their fieldwork.

The hypothesis specifies two language-use situations. If either of these is present in an area, it will NOT be necessary to undertake extensive multilingualism testing. The hypothesis is as follows:

Extensive second-language proficiency testing is not necessary in an area if:

1. The second language is not used in the home/family domain and children are not being raised as mother-tongue speakers of that second language; or
2. The second language has a standard dialect usually acquired through education and 80% of the population is not educated beyond four (successful) years of school.

[blank]

Calibrating Sentence Repetition Tests

Joseph E. Grimes

Sentence repetition tests are potentially useful for surveying the distribution of different levels of bilingual proficiency in areas where a second language is part of the sociolinguistic picture. They are patterned on tests that have been used monolingually to investigate speech development and speech disorders.

In order to validate the use of such tests, it is necessary to give them to many subjects and compare the results with those of fine-grained qualitative tests such as the U.S. Foreign Service Institute's well-tried interview test or its spinoff for nonliterate subjects, the Second Language Oral Proficiency Evaluation.

A procedure described by Radloff (1991) for setting up a sentence repetition test for validation corresponds to a type of implicational scaling: test probes are chosen to maximize the likelihood that the rank order of correct responses will correlate well with the rank order of proficiency test outcomes.

This paper discusses a computer program that implements a procedure equivalent to Radloff's that may expedite the calibration of possible test probes. It also calls attention to questions that remain concerning the suitability of sentence repetition tests for assessing all aspects of bilingualism.

Variation in bilingual proficiency

Minority languages go out of use when their speakers become fluent in another language, provided they come to use nothing but that second language in all walks of life.[1] The sociolinguistic study of minority languages, then, often requires the study of how a second (or third, or fourth) language [2] is used by different segments of a community.

Good second-language research goes beyond simply listing what other languages are used. Three requirements for such research are widely recognized:

1. Different individuals in a community learn their second language with different degrees of proficiency, depending mainly on their personal need to communicate in it and their opportunity to learn it. If bilingualism is widespread, it is the task of the survey to characterize in social terms who is and who is not proficient.
2. Different sectors of a society have different modal patterns of bilingual proficiency. The proficiency of men often differs from that of women; different geographical or subcultural regions may differ widely in proficiency; differences in age, education, and travel experience usually go with differences in proficiency. The effect of these social divisions, of their combinations, and possibly of other distinctions in some situations, is what has to be investigated.
3. In a field situation, it is invariably the more bilingual people that an investigator meets first. He or she may therefore have to go to great lengths to find a sample that reflects the variability of proficiency throughout the society as a whole. This usually implies a fairly large sample that is stratified according to the likely lines of division mentioned in point two.

The size of the sample (typically 25 to 100 subjects in each community, depending on how segmented the society is) sometimes becomes a barrier to teams that need to survey bilingualism, because the higher precision tests that are available (Bruhn 1989, Summer Institute of Linguistics 1987) take a lot of time. So it is natural and legitimate to look for less time-consuming ways to get an accurate reading on bilingual proficiency.

A test known as the sentence repetition test is used routinely in speech pathology and speech development studies. It is based on the observation that once a sentence reaches a certain level of complexity, you can't repeat

[1] Bilingualism as such does not lead inevitably to language extinction. The second language has to take over all speech domains. Many bilingual groups settle into a stable pattern of using one language for certain functions and the other for other functions.

[2] I employ "bilingual" here to refer to the use of any language other than the speaker's mother tongue; "multilingual" would be more appropriate for many societies.

it if you can't understand it. Language is such a tightly knit web that if one strand gets loose it takes others with it.

Many people can do a fair job of mimicking a short utterance in a language they don't understand—two or three words is common in the teaching of phonetics. In a language they know only imperfectly they can sometimes mimic six or seven words successfully. But by the time a sentence gets longer than that—and especially if it is structurally complex like 'A loose nut was what we were advised to watch out for'—they have to control the syntax and the lexical valences rather well or they trip all over themselves. A native speaker or a fluent second language speaker, on the other hand, has little difficulty repeating that kind of thing.

A sentence repetition test is simply a list of natural sentences recorded on tape. Test subjects are asked to repeat each sentence. The list is ordered from less difficult to more difficult. It is usually preceded by warmup or training sentences that nearly anyone who knows the language at all can handle, but that are not taken into account as part of the score.

Each subject hears each sentence only once.[3] His or her task is to repeat the sentence exactly as it was uttered. There is a little leeway in the scoring that is explained later, but basically the subject either gets it right or misses it; scoring does not need to make fine discriminations that require expert judgments in order to quantify degrees of error. This makes it relatively easy to train people who do not know linguistics as testers.

The advantage of such a test is that it can be administered to a large number of people over a wide area in a relatively short time: twenty minutes or less per subject.

The disadvantages are twofold. As long as the surveyor is not tempted to draw conclusions that the test is incapable of giving by its very nature, neither disadvantage need stand in the way of using it. The disadvantages are:

1. Before the sentence repetition test for a particular second language[4] can be interpreted meaningfully, it must be calibrated against an independently validated test of bilingual proficiency. This calibration is a major effort in itself; to be convincing it involves administering

[3] Earphones can keep bystanders who are also potential test subjects from having the opportunity to practice ahead of time.

[4] It is an open question whether a test that is effective for native speakers of one language who also speak the test language is automatically effective for native speakers of another language who also speak the same test language—for example, whether a French test given to Arabic speakers who also speak French gives comparable results when administered to Italian speakers who also know French. The question merits investigation in its own right.

fine-grained interview tests to a large number of subjects representing all proficiency levels, preferably in each of several communities.

2. The content of the test is such that it does not probe all the areas of language that have proven critical in assessing higher levels of second language proficiency. It probes mainly sentence structure, some comprehension, and phonology. It does not probe command of the lexicon very extensively (the ability to choose words or phrases appropriate to the interaction), nor does it give much information about fluency or the ability to manage connected discourse or overall comprehension. It therefore permits only an a fortiori argument about proficiency: If a subject can't handle sentence structure and phonology and basic comprehension very well, or doesn't know even the 150 or so words used in the test, it stands to reason that he or she should have trouble with complex comprehension, the lexicon, discourse, and fluency as well, therefore is not very proficient. But the opposite argument cannot be made: A perfect score on a sentence repetition test still tells nothing about lexical command or discourse, and gives only a sentence-size reading on fluency and comprehension.[5]

Nevertheless, if a sentence repetition test is properly validated, and if it is used as a screening test to sort out subjects on the lower end of the proficiency scale without attempting to use it for discriminations at high levels of proficiency, it might form the backbone of the bilingualism phase of a number of surveys.

Calibrating a sentence repetition test

Calibrating a sentence repetition test begins by finding a pool of fifty or so bilingual people—preferably in each of several culturally important communities—whose proficiency in their second language has already been

[5] Thea Bruhn (personal communication) reports that lexicon, discourse, fluency, comprehension, and structural command are evaluated independently of each other by the Foreign Service Institute because they correlate poorly enough with each other that none of them has turned out to be a statistically impressive predictor of any of the others.

assessed by an external test such as SLOPE (SIL 1987).[6] Complete calibration requires a pool of people whose abilities range from practically nothing to practically perfect bilingualism.

The sentences used for testing can be drawn from any unconstrained text—newspapers, conversations, everyday life, stories. They should be long enough that mere mimicry breaks down. Only a few should be as short as five words; most should be nine words or longer. They should be sentences that can stand on their own two feet, clear without context. They should not represent nonstandard or low class speech; on the other hand, neither should they be literary or fancy. The more grammatical complications they include naturally, the better.

Arrange fifty or so sentences of this kind from short to long, from apparently simple to apparently complex, for the initial recording. Have several native speakers of the language make high quality test tapes of the whole list, preferably in a recording studio. If they stumble when they try to utter some of the sentences, edit those sentences out—if native speakers can't manage them, what will the others do?

Choose the best recording as a calibration tape. Run it past the pool of fifty or so bilinguals whose proficiency levels you already know. Have them try to repeat every one of the fifty or so calibration sentences with no replays. Score 3 for each sentence they get absolutely perfect, 2 if they make one mistake, 1 if they make two, and 0 if they make any more—Radloff (1991) tells how to set up testing sheets and what counts as a mistake.

Fifteen or twenty of the test sentences are to be chosen out of the initial list to form the actual field test. The ones that are retained show three characteristics:

[6] Where bilingualism is supported through schooling in a major language, the classroom-oriented interview tests of the Educational Testing Service and the American Council of Teachers of Foreign Languages may be useful. They are less precise than SLOPE when it comes to discriminating the higher proficiency levels. Another method for guessing intelligently at proficiency levels without really observing speech behavior has also been used for calibration, though it is more properly taken as another screening method in its own right. In the Reported Proficiency Evaluation described by Radloff (1991), a number of mother tongue speakers of the language in question are interviewed about their impressions of people they know who speak that language as a second language—for example, native speakers of standard Spanish might be interviewed about how well their Quechua-speaking acquaintances handle Spanish. The interviewer guides them through a set of criteria similar to the ones used in SLOPE, but there is no way to verify the interviewees' impressions against specific speech samples. The level of precision that can be attained is therefore more suitable for general screening than for calibrating another test.

1. They can be ordered in such a way that any subject who repeats a later sentence correctly is almost certain to repeat all the sentences before it correctly as well. In the other direction, however, repeating an earlier sentence is no guarantee that any of the later sentences can be repeated. This kind of relationship among sentences is generally known as an implicational scale (Hatch and Farhady 1982) or a "Guttman scale" after Louis Guttman, who first developed the idea.
2. A subject's ranking on the implicational scale gives a clear indication of his or her ranking on the independent scale established via the use of a higher precision test of bilingual proficiency.
3. The sentences chosen give the clearest achievable discrimination across all levels of proficiency.[7] At this point a computer program called SENREP[8] simplifies accumulating the calibration test results, ranking the sentences by difficulty and the subjects by proficiency, and sorting out those sentences that fit the model of an implicational scale from those that don't. In addition, the same program can be used for administering the test widely (as opposed to calibrating it, which is merely the startup phase) by matching the response patterns of test subjects with those of the original calibration subjects.

The computer program goes through several steps:[9]

1. You give it an identification of the language being tested and the survey that the test is part of.
2. You give it the text and the order of the fifty or so candidate sentences used to establish the calibration, from which the actual working set is drawn.

[7] Because of the inherent inability of sentence repetition tests to probe discourse competence, lexical choice, range of comprehension, or fluency, my own recommendation would be to lump together all levels of proficiency from 3 to 5 on the Foreign Service Institute scale into a single range of "3 and above" in order to reduce the temptation to read high level discriminations out of a low level test. The reason for that cutoff level is that above level 2+ it is precisely the factors that the sentence repetition test cannot probe that are most important in making discriminations.

[8] Machine readable copies of SENREP compiled for PC-compatible computers are available for the cost of reproduction and mailing from the publisher of this volume. There is no guarantee that they will work. The function of a manual is fulfilled by a help file included with the program that can be called up at any point by pressing function key 1.

[9] You can stop at any point and store the data you have entered up to then in a file. Later on you can retrieve that file and continue from where you left off. A typical calibration for one community involves a minimum of fifty subjects and fifty candidate sentences, which is too much to type into a computer accurately at one sitting.

3. You give it the calibration test results for each subject: a coded identification of the subject (to safeguard anonymity), the proficiency level established for that subject by the independent test (FSI or SLOPE), and the tester's rating of that subject's response to each of the fifty or so candidate sentences in 2.
4. It gives you its analysis of each candidate sentence in the form of a chart. The chart is stored in a separate computer file that can be examined and printed by importing it into an ordinary spreadsheet program.[10] The columns in the chart correspond to the calibration subjects. Each column is headed with the subject's identification code, his or her proficiency level as established independently, and the cumulative score over all the test sentences. The subjects are arranged from those who got the fewest answers correct to those who got the most. The rows in the chart correspond to the candidate sentences themselves. They are ordered from easiest to hardest: a sentence that everybody repeated correctly would have a difficulty index of 0 associated with it, and one that nobody could handle would have a difficulty index of 1. In addition, each sentence carries a deviation index (Radloff's "discrimination index") that tells how closely that sentence fits an ideal implicational scale, with 0 deviation for a sentence that scales perfectly relative to the rest.
5. It gives you its summary of the scalability of the whole chart. There is an overall index of reproducibility or scalability, which in the calibration phase is likely to be unacceptably low because none of the candidate sentences that don't scale well have been eliminated yet. Two other indexes sum up the overall difficulty and deviation.

The table in (1) illustrates the calibration process in miniature. It consists of only seven candidate sentences instead of fifty, taken from an Urdu example of Radloff's (with hypothetical Urdu data). The sentence repetition ratings are given for only twelve calibration subjects instead of fifty, in order to keep everything on a single page.[11]

[10] Commonly used spreadsheet programs include Lotus 1–2–3 (TM), Quattro Pro (TM), Microsoft Excel (TM), or AsEasyAs (TM). The spreadsheet controls the printer and lets you look at the parts of the chart you want to see. You can remove sections of the chart that are superfluous. Other uses of a spreadsheet program in linguistics include putting discourse information into a Thurman chart (Grimes 1975) or into Longacre's (1989) band structure, or manipulating charts of tone patterns, affixes, or anything else that involves rows and columns. There are also the standard project management uses of spreadsheets: doing accounts, budgets, tax calculations, mailing lists, statistics, itineraries, bibliographies, progress records, and graphs of all the above.

[11] "Probe" is used instead of "calibration sentence" in keeping with general testing terminology, and to save space on the chart.

(1) Sentence Repetition Test calibration data

Calibrate probes by sum of subject scores

Subject→			E	A	I	B	F	J	C	G	K	H	D	L	
Proficiency			0+	0+	0+	1	1+	1+	2	2+	2+	3	3	3	
Sum of scores			1	1	2	3	6	7	9	13	14	16	17	20	
Probe	Diff	Dev													
1 (P6)	0.31	0	1	1	1	2	2	2	2	2	3	3	3	3	*talim hasɪl kʌrke . . .*
2 (P7)	0.56	2	0	0	1	0	1	1	2	2	2	2	2	3	*kʌl tʃʌlte wʌqt . . .*
3 (P3)	0.56	2	0	0	0	0	1	1	2	2	2	3	2	3	*ɪs dʒʌdid dɔr . . .*
4 (P4)	0.58	4	0	0	0	0	0	1	0	3	2	3	3	3	*mæn ʌb mʌzid ta . . .*
5 (P5)	0.64	0	0	0	0	1	1	1	1	1	1	2	2	3	*ʊn ki talimat ke . . .*
6 (P1)	0.64	4	0	0	0	0	0	0	1	2	3	1	3	3	*tʊm ne hath pʌr . . .*
7 (P2)	0.69	0	0	0	0	0	1	1	1	1	1	2	2	2	*kjunkɑ talim hʌm . . .*

Reproducibility (scalability)	0.95, should be > =	0.9
Minimal marginal reproducibility	0.62, should be <	0.9
Proportional reduction in error	0.87, should be >	0.6

The subjects are identified by capital letters. Each of them has been assigned a proficiency level by an independent test of known accuracy. The scores in each column (3 for a perfect repetition, 2 or 1 for minor mishaps, and 0 for any more serious flaw) are added up, and this sum is the basis for the left-to-right order in which the subjects appear.[12] The full text of each calibration sentence is given.

In (1) the sentences are no longer in the order in which they were presented to the calibration subjects (P1, P2, . . . , Pn). That tentative order was based on their length and an informal guess about their relative complexity. In (1), however, they are ordered by their calculated difficulty index and renumbered accordingly, carrying along the original presentation ordering (P) in order to make it easy to find them on the audio tape.

[12] The program also calibrates the probe deviations by the ranking of proficiency judgments. Superficially this appears less neat than the ranking of sentence repetition scores, but it is probably more realistic because the proficiency scores come from a nominal scale established by a set of definitions, not an ordinal scale based on higher or lower scores. Furthermore, the proficiency scale is nonlinear: the relative amount of effort required to get from level 0 to 0+ is only a small fraction of the amount of effort required to get from 3 to 3+, for example. The time projections for attaining proficiency given by Brewster and Brewster (1976:377) are very much like a Fibonacci series (0, 1, and the sum of the two preceding numbers), which is nonlinear. The sentence repetition scoring, on the other hand, is linear. In the long run I will probably recommend using the ranking of proficiency judgments for calibration.

From a printout of the calibration chart you can choose which candidate sentences give you the most information for the least work. Choose fifteen to twenty sentences out of the original fifty by these three rules:

1. Ignore sentences with a difficulty index less than .05 or over .95—the reasons why everybody might get everything right, or nothing right, are extremely complex but tell us almost nothing.
2. Choose a spread of difficulties that spans the spectrum evenly.
3. Prefer low deviation indexes.[13]

Bad sentences are the ones with high deviation indexes: they do not fit the implicational scale, therefore you cannot make accurate judgments from them. You don't want to choose sentences with difficulty indexes that are too close together to discriminate; you want your difficulties spread out evenly from .05 to .95, not bunched up. You may have to flip a coin to decide which of two equivalent sentences to use.

Once you decide which sentences form the best test, you invoke the computer again to record your choices. It shows you each calibration sentence from the easiest to the hardest, giving you its original number, text, difficulty index, and deviation index. Then you register whatever you decided: to use that sentence or drop it from the list. (The computer never really erases a sentence; you can get back to the original set at any time by going back into calibration mode.)

After you have registered your judgments about everything on the list, the computer recalculates everything using only the sentences you have decided to keep for the actual test. It gives you a new sum of scores for each subject, taking into account only what you found to be the most consistently useful sentences. It reorders the subjects by their new scores. It recalculates difficulty and deviation indexes and the three summary indexes and puts everything out in a new chart file. At this point you save the new working data base so that you will see only the sentences you want to use next time you retrieve the data.

The table in (2) is the recalculation for the mini-test given in (1). It is limited to three sentences instead of the usual fifteen to twenty. The sentences selected were the three with the least deviation from ideal scalability. Choosing those sentences and ignoring the rest improves the reproducibility scores noticeably. The test sentences in (2), however, cover the difficulty range only from .31 to .69, and cover it unevenly. This would

[13] Radloff (1991) reports that in a full scale calibration effort there are hardly ever any deviation indexes as low as zero. In any full scale testing, we have to settle for minimum distortion, and may never achieve zero distortion.

not be an adequate test in practice, but it illustrates how the real one is put together.

(2) Sentence Repetition Test calibration results

Calibrate probes by sum of subject scores

Subject→				I	E	A	B	J	G	F	C	K	H	D	L	
Proficiency				0+	0+	0+	1	1+	2+	1+	2	2+	3	3	3	
Sum of Scores				1	1	1	3	4	4	4	4	5	7	7	8	
Probe		Diff	Dev													
1	(P6)	0.31	0	1	1	1	2	2	2	2	2	3	3	3	3	*talim hasıl kʌrke...*
2	(P5)	0.64	0	0	0	0	1	1	1	1	1	1	2	2	3	*ʊn ki talimat...*
3	(P2)	0.69	0	0	0	0	0	1	1	1	1	1	2	2	2	*kjunkɑ talim hʌm...*

Reproducibility (scalability)	1.00, should be	> =	0.9
Minimal marginal reproducibility	0.68, should be	<	0.9
Proportional reduction in error	1.00, should be	>	0.6

Rating test subjects against calibration subjects

Radloff takes a straightforward approach to evaluating the performance of test subjects once the calibration is finished. She uses a linear regression model: that is, she observes that the scores of the calibration subjects plot out fairly close to a straight line when their sentence repetition scores are put across the bottom and their independently established bilingual proficiency scores down the side. A formula that generates that straight line can be calculated from the scores, together with a correlation coefficient that expresses how close on the average they come to the hypothetical line. From the regression formula it is possible to match any test score with a proposed bilingual proficiency score.

There are, however, two reasons to think that working from a regression formula could be an oversimplification. The first is that the test sentences usually do not fit the ideal implicational scale perfectly; their average deviation index for a full scale test is in practice greater than zero. This means that the error that is due to lack of scalability compounds with the error expressed by a correlation coefficient less than one, which indicates the extent to which the calibration data fail to fit the regression formula.

If the compound error were to turn out unacceptably large, something other than the regression equation would be desirable.[14]

The second reason for caution about simple regression has to do with the nonlinearity of the bilingual proficiency scale mentioned in footnote twelve. A difference in sentence repetition test ratings at the low end should correspond to a greater difference in bilingual proficiency ratings than a difference of the same magnitude at the high end, because of the way the proficiency scale is defined. A lot of work on test designs will have to be done before this statement can be either affirmed or denied: the point for now is that it is desirable to have an alternative to the regression formula as a way to interpret the results.

The SENREP program gives another way of matching test subjects (whose bilingual proficiency you are trying to estimate) with calibration subjects (whose bilingual proficiency you established by giving them a fine-grained, independent test). It compares each` new subject with all the calibration subjects in order to be able to assign them a proficiency level indirectly.

In test mode you identify each test subject to the computer. The computer then takes you through the probes that have been selected—no longer the full list—in their order of difficulty and has you put in their scores as you did for the calibration.

The new chart file, when you bring it up in your spreadsheet program, begins with a copy of the current calibration chart for you to refer to. Then it shows you a comparison of each test subject with up to ten calibration subjects.

It chooses the subjects to compare by looking for sums of scores that are as close as possible to the test subject's sum. Then it picks the five calibration subjects lower than that point and the five above it to make the comparison (fewer, of course, at the ends of the list of calibration subjects).

The subjects who are listed first are the ones most like the test subject. If they are all at a single rank, say 2+, then you can safely guess that your test subject may be at 2+ too. If the closest calibration subjects are at several levels, or if they really aren't very close, then you might be more tentative about classifying that test subject.

The measure of closeness has two parts. The first looks at the test subject's answers to individual questions one by one, and compares them with the answers each calibration subject gave to the same questions. It

[14] There are other association measures that might be equally appropriate. The cosine of the angle in k-dimensional space between the vectors that represent the k test and calibration answers is widely used in information retrieval for a similar function, and is easily calculated.

adds up the differences and reports them as "d"; d=0 indicates identical responses.

The other part measures differences in the shapes of the response patterns. It is the

From (3) it can be seen that there might be considerable leeway in deducing proficiency levels from sentence repetition responses—though a full set of fifteen or more probes would undoubtedly give less helter-skelter responses than the ones in the example. One would, for example, feel confident about assigning subject M to level 0+ (or possibly level 1); but subject N might reasonably match anywhere from level 1 to level 2+, though the match is more likely with 1+ or 2.

(3) Sentence Repetition Test matches on calibration subjects

```
Test subject M is closest to calibrator I, level 0+
  M's scores are 2 0 0 Total 2
d=1  r= 1.000: Subject I,  total 1, assigned level 0+   1 0 0
d=1  r= 1.000: Subject E,  total 1, assigned level 0+   1 0 0
d=1  r= 1.000: Subject A,  total 1, assigned level 0+   1 0 0
d=1  r= 0.866: Subject B,  total 3, assigned level 1    2 1 0
d=2  r= 1.000: Subject J,  total 4, assigned level 1+   2 1 1
d=2  r= 1.000: Subject G,  total 4, assigned level 2+   2 1 1
d=2  r= 1.000: Subject F,  total 4, assigned level 1+   2 1 1
d=2  r= 1.000: Subject C,  total 4, assigned level 2    2 1 1

Test subject N is closest to calibrator J, level 1+
  N's scores are 2 1 1 Total 4
d=0  r= 1.000: Subject J,  total 4, assigned level 1+   2 1 1
d=0  r= 1.000: Subject G,  total 4, assigned level 2+   2 1 1
d=0  r= 1.000  Subject F,  total 4, assigned level 1+   2 1 1
d=0  r= 1.000: Subject C,  total 4, assigned level 2    2 1 1
d=1  r= 1.000: Subject K,  total 5, assigned level 2+   3 1 1
d=1  r= 0.866: Subject B,  total 3, assigned level 1    2 1 0
d=3  r= 1.000: Subject I,  total 1, assigned level 0+   1 0 0
d=3  r= 1.000: Subject E,  total 1, assigned level 0+   1 0 0
d=3  r= 1.000: Subject A,  total 1, assigned level 0+   1 0 0
```

```
Test subject O is closest to calibrator K, level 2+
  O's scores are 3 1 2 Total 6
d=1  r= 0.866: Subject K,  total 5, assigned level 2+   3  1  1
d=1  r= 0.866: Subject H,  total 7, assigned level 3    3  2  2
d=1  r= 0.866: Subject D,  total 7, assigned level 3    3  2  2
d=2  r= 0.866: Subject J,  total 4, assigned level 1+   2  1  1
d=2  r= 0.866: Subject G,  total 4, assigned level 2+   2  1  1
d=2  r= 0.866: Subject F,  total 4, assigned level 1+   2  1  1
d=2  r= 0.866: Subject C,  total 4, assigned level 2    2  1  1
d=2  r= 0.000: Subject L,  total 8, assigned level 3    3  3  2
```

Closing

The text of the SENREP program, as well as a disk containing it, which can be obtained from the publisher of this volume, is written in PDC Prolog (formerly Borland Turbo Prolog), a computer language that makes heavy use of recursively defined functions. As a consequence it carries no inherent limitations on the number of probes, calibration subjects, or test subjects that it can handle, nor on the length of specific pieces of data. It is, however, limited by the memory capacity of the computer it runs on. A slower version with much greater capacity could be developed by putting the larger data base components on a hard disk.

[blank]

Part IV: Technical Review

A Review of Walker's Research

Dale Savage[1]

For more than two decades there has been an awareness within SIL of the crucial role that language attitudes may play in the implementation of vernacular translation and literacy programs. Casad (1974) indicates an SIL interest in language attitudes related to bilingualism as early as the mid to late 1960s in Mexico. A recognition of the vital role of language attitudes in minority language planning spread widely in SIL during the 1970s and is reflected in the 1980s in three important conferences[2] which focussed significant attention on the issue of language attitudes in assessing the

[1] I wish to thank Bob Bodwell, a professionally trained statistician, with whom I discussed a number of the issues presented here before the actual writing was begun. I also wish to express appreciation to Eugene Casad, the editor, for his patient and helpful suggestions, and special credit is due my wife, Lorraine, who read drafts for coherence and who heard me rehearsing the central ideas more often than I care to think about.

[2] The first of these, the Sociolinguistic Survey Conference at Stanford University, November 27–28, 1982, (Huttar 1982) was organized and hosted by Charles Ferguson, Shirley Brice Heath, and John Rickford of Stanford, and brought together several participants from professional academia and SIL. Several participants discussed the difficulties of language attitude assessment during the relatively brief encounters associated with a sociolinguistic survey.

A second conference, the Stanford Conference on Vernacular Literacy, July 24–25, 1987, (Shell 1988) with a major attitudinal component was also hosted by Shirley Brice Heath and Charles Ferguson. Participants in this conference raised the issues of language attitude assessment and the implications of language attitudes for literacy program planning.

The third conference, the International Language Assessment Conference at Horsleys Green, England, May 24–31, 1989, (SIL 1989) included a large section of papers dealing with language attitude assessment.

prospects for vernacular literacy and translation programs among minority language groups.

It is within this context that Roland Walker has considered the problems of assessing language attitudes and vernacular literacy acceptance. A contributor to all three of the conferences mentioned above (Walker 1982, 1988, 1991), many of Walker's ideas about assessing attitudes and predicting the acceptance of vernacular literacy grew out of his field experience in Irian Jaya, Indonesia. His graduate studies at the University of California, Los Angeles, culminated in his 1987 dissertation *Towards a Model for Predicting the Acceptance of Vernacular Literacy by Minority-Language Groups.*

A general theme running through the works cited has been that the assessment of language attitudes is important for making good decisions for language programs, but that the direct assessment of attitudes is, for numerous reasons, too difficult to accomplish in a field survey setting. Walker's approach, then, has been to assess other factors and assume that language attitudes could be inferred from them.

Walker's Horsleys Green paper (Walker 1991) indicates that his research and thinking have had a significant impact on language survey design and interpretation in SIL Indonesia's Irian Jaya Program Committee. Through his publications, Walker's ideas about assessing language attitudes and vernacular literacy acceptance have the potential for gaining considerable currency within SIL beyond Irian Jaya. This is partly due to dual drives within SIL to assess its remaining task of providing vernacular translations and literacy programs as quickly as possible, and also to a desire in some quarters to limit the remaining task to a manageable size. Walker's approach, therefore, is attractive to some because it promises speed, and limits the task somewhat by identifying those language communities which have a low probability of accepting vernacular literacy.

Given its potential for wide influence, it is advisable to subject Walker's approach to careful scrutiny to ascertain if it is adequate in its present form to play a major role in decisions regarding initiating particular language programs. Several questions need to be asked: (1) Are the claims made about language attitudes and other underlying assumptions valid? (2) Is the sample drawn in a manner designed to render valid generalizations about potential literacy programs worldwide as it purports? (3) What is the nature of the data generated by the questionnaires? Is it reliable? (4) If the study meets the normal standards of reliability and validity, how good are the results as a decision-making tool?

Complexities of assessing language attitudes

Perhaps the critical questions to be addressed should be: (1) Does the model presented by Walker provide an adequate basis for assessing attitudes? (2) Are the claims made by Walker warranted that the results of this study have confirmed the approach of assessing language attitudes by evaluating the forces that shape them?

The importance of these questions becomes readily apparent when we consider the great shift that has occurred in thinking within SIL about how language attitudes (and other sociolinguistic factors) should be incorporated into program planning. John Bendor-Samuel (1982), participating in the 1982 Stanford Sociolinguistic Survey Conference, clearly enunciated the then generally accepted position that if comprehension (whether intelligibility or bilingualism) were low then work should be done. If, on the other hand, comprehension were high, then language attitudes should be assessed to determine whether a project should be undertaken anyway because of the negative attitudes toward the second language or its speakers.

By contrast, in recent years there has been a strong push toward the opposite position that even when there is clearly inadequate comprehension of a second language, putative negative attitudes toward mother-tongue literacy or other negative social factors may deter us from projects where translation might have been desirable 'in principle'. See David Bendor-Samuel (1991) for an exposition of this view.[3] Thus, in the current climate, questions of language attitude become crucial when we consider that reports of negative social pressures (including attitudes) may result in curtailed programs even for those minority language groups with clear comprehension needs. It is in this setting that we must address Walker's claim that this approach presents us with an effective means of assessing attitudes.

If we look at Walker's presuppositions and plan of research, I believe the answer to our second question above becomes evident. First, Walker assumes that attitudes toward language are shaped by sociolinguistic forces from within and without the community, and that these sociolinguistic variables [forces] can be observed and measured. On the other hand, Walker assumes that the difficulties of trying to assess attitudes, in their own right, in the field are so great that, "it would seem best... to set aside the study of language attitudes, and rather, to evaluate the sociolinguistic forces that shape language attitudes which are observable and measurable." (Walker 1987:49–50).

Walker then turns his attention to devising a questionnaire based on some of the factors implicated in the literature on language shift and death

[3] But note also Early (1991) and Hollenbach (1989) for two alternative viewpoints.

(Walker 1987:51). Rather than assessing attitudes, he has actually placed their assessment outside the scope of his study, focussing instead on establishing linkage between some of the factors related to language shift and the acceptance of vernacular literacy.[4] This is a perfectly legitimate research goal and strategy, and a very worthwhile investigation. However, Walker never returns, neither in his dissertation nor in any subsequent research, to validating the linkage between these factors and the assumed relation to language attitudes—the relationship of these factors to language attitudes remains merely an assumption just as it was at the beginning.[5]

Walker's insistence that this approach is a means of gathering attitudinal data is unfortunate for two reasons. First it draws attention away from the genuine contribution of this line of thought. There really does appear to be a relationship between the general factors Walker is considering and literacy behaviors in minority language groups. Its value is in no way enhanced by claiming it assesses language attitudes. The second consequence of claiming this is an effective (and adequate) means of assessing language attitudes is that such a claim may potentially undermine efforts to carefully assess the types of attitudinal data necessary to supplement this type of research for literacy program implementation.

There is no denying that trying to assess attitudes in emerging communities is full of difficulties. It is not, however, impossible, and there are numerous indications that static approaches which focus solely on observable social phenomena do not adequately account for the observed response set of social behaviors and the attitudes they imply.

In a long term research program which bears a great deal of topical similarity to Walker's research, Howard Giles and a number of associates articulated a theory of language in intergroup relations and ethnolinguistic vitality (Giles, Bourhis, and Taylor 1977). From the inception of the model, vitality was assessed by considering three classes of objective factors: status factors, demographic factors, and institutional support and control factors. The early operational procedures involved consulting "demographic, economic, sociological, and historical documents to arrive at as 'objective'

[4] Like Walker, I assume that there is a connection between language attitudes and the factors involved in language maintenance and shift. The operational aspects of his model, however, assume an isomorphic relationship between these factors and attitudes to which I do not subscribe.

[5] If Walker wishes to confirm this approach as a method of assessing attitudes, then he should actually measure the attitudes with a well validated, standard measure of attitudes and then do regression analysis of that result against the observable social factors. Lambert et al. (1960) provides a good description of the type of process involved in validating a new methodology (the matched guise technique) for attitude assessment. See also Huff (1954:74) for what he calls the semi-attached figure.

an assessment of a group's vitality as possible" (Bourhis, Giles, and Rosenthal 1981:146–47).

While the objectively assessed vitality appeared to provide "a useful tool for comparing ethnolinguistic groups in cross-cultural research," as early as 1979, objective factors alone were not considered sufficient to "account for group member's intergroup attitudes, skills and motivations for second language learning, attitudes toward language usage and use of code switching strategies" (Bourhis, Giles, and Rosenthal 1981:147). Instead in 1981, a parallel track of research was initiated to investigate ethnolinguistic group member's "subjective perceptions" to the same set of vitality factors in order to supplement the objective data.

> A combination of objective and subjective data may be extremely valuable in assessing the likelihood that ethnic minorities will survive as distinctive cultural and/or political entities in majority cultures. 'Subjective' vitality data may provide advance indication that a particular minority group is to mobilize in an ethnic revival phase not otherwise forseeable solely on the basis of 'objective' vitality information. (Bourhis, Giles, and Rosenthal 1981:147)

There are many other instances in the sociolinguistic literature which indicate that something beyond the observable social milieu is necessary to account especially for the behavior of subordinated or socially disfavored linguistic groups. As Ryan (1979) notes, once the legitimization of a dominant language,

> has resulted in universal recognition of the standard, one might expect the other varieties to disappear over a generation or two. However, many regional, ethnic, and social class varieties . . . have tended to persist for centuries, surviving strong pressures to succumb in favor of the standard dialects. (Ryan 1979:145)

This notion of persistence in the face of dominant languages (and their speakers) implies that even though there may be "universal recognition" of the legitimacy of the standard, in many cases there are also other forces including attitudes at work which conspire to preserve group unity and identity in the face of pressures for assimilation.[6]

Susan Gal (1989) and especially Kathryn Woolard (1985; 1989) have pointed out, that in the case of linguistically dominated groups, quite different complexes of attitudes may lie beneath the surface of outwardly similar 'objective' circumstances. Woolard who did extended research on the politics of language and ethnicity (including language attitudes) during the period in which Catalonia achieved autonomy argues,

> We cannot read hegemonysaturation of consciousness directly from the institutional domination of a language variety. Just as nonstandard practices may accompany standard consciousness, so it is logically possible that standard linguistic practices may accompany or conceal resistant consciousness, as a form of accommodation to coercion rather than the complicity essential to the notion of cultural hegemony. The distinction is important, because accommodative behavior may be more easily dislodged and does not present the same problem for social change as does collaborative consciousness. (Woolard 1985:741)

Woolard, an anthropologist, gathered the data for her analysis through a variety of means including participant observation and interviews, but referring to the measurement of attitudes and the analysis outlined above,

[6] Unfortunately there are relatively few empirical or ethnographic studies of persistence. As Fishman (1990:5–10) reports in an important new thrust on "reversing language shift," as a result of "several societal and social biases," there has been definite skewing of research in the direction of shift rather than persistence. Sociolinguists (and other social scientists) have generally focussed most of their attention on processes of change within language groups. Accordingly we have a very refined taxonomy,

> with respect to the 'minus' side of the ledger (we speak of language attrition—shift—endangerment—loss—death and can itemise many studies of each way-station along this increasingly negative progression), while the 'plus' side remains rather gross and undifferentiated and studies of revival, restoration, revitalization and restabilisation remain proportionately few and far between. (Fishman 1990:6)

Fishman attributes the lack of attention to persistence to "our modern fascination with the dynamics of change *per se*," and points out that, "the forces and processes of change coexist, *in a single process*, with the forces and processes of persistence, and what most social scientists mistakenly call 'change' is really the by-product of the *interaction* of persistence and change" [italics in the original] (Fishman 1990:11).

If Fishman's call for researchers and "change-agents on behalf of persistence" is successful in attracting attention to the study of reversing language shift, we may begin to see more data and analysis into how and why linguistic groups successfully resist structural coercion and linguistic dominance.

writes, "this finding comes not from data on language use, but from what are called 'subjective reaction' tests. This is a form of empirical evidence on the social evaluation of language use, as important as evidence on language use itself" (Woolard 1985:741).

It is, in part, the social and psychological complexity of attitude structures that leads to the low correspondence "between attitudes and actual behavior" that is discussed in Agheyisi and Fishman (1970). We may agree with the notion that attitudes are "agendas to action," but there appear to be important situational constraints that mediate between various observable stimuli and attitude/behavioral responses.[7]

Thus, as Walker himself notes, a respondent may tell you he has one attitude with respect to an object of affect and then perform an act which runs counter to his verbal report. This does not mean that the verbal report was necessarily deceptive or inaccurate. Rather it is a function of the fact that attitudes are associated with a wide range of values, beliefs, and intentions with respect to objects of affect (including other languages and their speakers), and different social situational contexts necessitate differential normative patterns of behavioral response.[8]

This is in fact, the situation Labov (1972:292–96) describes in which nonstandard speakers endorsed the norms of the dominant group in the test situation, but did not wish to adopt those norms. In coming to grips with this apparent anomaly, Labov posited the existence of "covert norms" in support of the vernacular. In his own words,

> Why don't all people speak in the way they obviously believe they should? ... Careful consideration of this difficult problem has led

[7] The whole notion of the seeming inexact match between attitudes and behaviors is one of the classic discussions in the attitude literature. Two major review articles of language attitude research, Agheyisi and Fishman (1970) and Giles et al. (1987), deal with the topic as either "intervening" or "mediating" variables between language attitudes and behavior. Brudner and White (1979) present research which highlights the language attitude/behavior problem with respect to Irish Gaelic. Ehrlich (1969), Fishbein and Ajzen (1975), and Wicker (1969) present discussions and theories about attitude/behavior problems from the perspective of mainline attitude research. An entirely new and promising approach to the problem of attitudes and behavior is being developed by several scholars using Catastrophe Theory, the recently articulated qualitative mathematical theory of Rení Thom (Anderson 1985, Ball, Giles, and Hewstone 1984, Flay 1978, Tesser 1980). This latter may eventually serve to make the notion of intervening variables obsolete.

[8] Given that this discussion is about social norms and attitudes, individual dispositions and deviance, though relevant to the study of attitudes and normative behaviors, will not be addressed here. Tesser (1980) provides a short, but interesting, entrée into the conflict between individual dispositions and social norms that may be useful to those involved in intensive participant observation.

> us to posit the existence of an opposing set of covert norms, which attribute positive values tọ the vernacular . . . We have therefore some empirical support in positing the opposition between two sets of values as the normative correlate of stable sociolinguistic markers . . . We agree with Homans (1955) that *the proper object of study should not be behavior alone, or norms alone.* (Labov 1972:295–96) (emphasis added)

Trudgill (1984) notes a very similar situation in Norwich, England where a definite language change is underway in a nonstandard direction related to the concept of covert prestige. In Norwich, Trudgill found that there were expressed values about language that were consonant with conferring prestige based on standard language norms. But he also uncovered data which demonstrate that for certain sex and class combinations, "nonstandard speech is in a very real sense highly valued and prestigious," and the working class dialect is gaining speakers from the middle class. That is to say that in the face of a socioeconomic structure which confers prestige on more standard speech varieties, a language shift is taking place in which a stigmatized variety of speech is gaining speakers at the expense of the dominant variety.[9]

Trudgill's Norwich studies are highly instructive as we consider assessment. If we were to take a rather simplified look at the existing social, economic, and educational structures, we would find considerable negative pressure exerted on the non-standard working class urban dialect. Regarding attitudes as well, at the level of conscious awareness, there were overt expressions of dissatisfaction by subjects with their own speech and stated desires to "speak properly." If the assessment of social factors and attitudes stopped at that, we might be tempted to assume that the working class dialect would be doomed to disappear under the weight of outside social pressure harnessed with negative attitudes toward the vernacular. Trudgill, however, did find that there were "deeper motivations for their actual linguistic behavior than these overtly expressed notions of their own 'bad speech' " (Trudgill 1984:57). So, in this case at least, observable sociolinguistic forces and a cursory examination of language attitudes

[9] In a review of a large number of language attitude studies, Ryan (1979:152) found a general trend that, "both evaluative reaction and questionnaire studies have revealed that nonstandard speech varieties may have low prestige but are associated with other values of importance for an ethnic group."

would not be sufficient for us to predict the type of covert prestige-driven language shift that is occurring.[10]

In numerous studies focussed specifically on language attitudes, important differences in attitudes within a group have been discovered based on locally relevant social categories such as age, gender, social class, occupational groupings, etc. (Ryan 1979). Walker's approach does not take into account the fine-grained attitudinal differences which have been demonstrated to exist in various parts of the world.

The general drift of virtually all the attitude research literature in the mentalist tradition (to which Walker subscribes) indicates that an attempt to assess language attitudes without more internally-focussed discovery procedures (such as interviews or subjective evaluation methods) and careful, in-depth observation is not likely to yield the quality of attitudinal data which can be most useful in the planning and implementation of literacy programs.

At this point, we answer the first of our questions: a model which gathers only the more easily assessed "objective sociolinguistic factors" is too impoverished in itself to adequately account for the complex attitude structures which interact within dynamic social systems. A research model of this type does not assess the target population's subjective perceptions vis-à-vis the objective factors, nor is it equipped to probe for resistant consciousness to linguistic domination. The model can reveal neither the complexities of attitude structure and conflicting norms in various social contexts, nor the potentially meaningful differences in attitudes accruing to socially relevant subcategories within a linguistic group.

It is imperative, therefore, that we realize that the assessment of observable social phenomena (including patterns of language use) and the assessment of language attitudes are not fungible; rather they are complementary. When we have one without the other, we are unable to properly interpret the significance of either.

The criterion variables

Turning from questions about whether or not Walkers' research regimen assesses attitudes, we now look into the actual content of the research question itself. Walker first built indices of vernacular literacy acceptance by asking fieldworkers to assess community literacy acceptance in their

[10] This accords well with Ryan's (1979:154) observations that both "direct and indirect measures of language attitude appear to be critical," and "direct questions may not reflect the whole picture."

locales[11] according to four criteria: (1) the sale of vernacular literature, (2) reading ability, (3) the amount of informal reading, and (4) the usage of vernacular Scriptures in churches. Pearson correlations were performed to correlate the four criterion variables with 19 predictor variables drawn largely from the language shift literature.[12] Finally, multiple regression analysis was performed on those criterion variables which garnered enough significant simple correlations to entertain this analytical technique.

In implementing this regimen, only the first criterion variable, the percentage of the population purchasing vernacular literature and the fourth criterion variable, usage of vernacular Scriptures in churches[13] exhibited enough significant simple correlations to perform multiple regression analysis.

Within the test group then, a relationship is indicated between the predictor variables and the sale of vernacular publications (criterion variable 1) on the one hand, and between the predictors and the frequency of public reading of vernacular Scriptures in church (criterion variable 4) on the other. The more interesting criteria, reading ability (criterion variable 2) and informal usage of vernacular literature (criterion variable 3), which represent individual literate behaviors, washed out of the model.

There are legitimate questions as to whether the two remaining criterion variables in themselves provide a very useful way of characterizing vernacular

[11] Walker's study was limited to the single communities which the respondents knew best, not the entire language groups. His stated purpose in confining the study to single communities was to restrict the variability in the data because in a 1986 study "using the entire language group as the unit of analysis" tended "to average out the variation that could be explained by the predictor variables" (1987:71). While this is a legitimate restriction for the purpose of the study, the very variability he has chosen to restrict suggests that within many of the communities where literacy has not been accepted (as measured by Walker's criterion variables), alternate allocation strategies or program approaches might have produced more desirable effects assuming, of course, that the general thrust of Walker's model is valid.

[12] The actual questions used to assess the criterion and predictor variables from Walker's questionnaire are included in appendix G. A copy of the entire original questionnaire may be found in Walker (1987:238–45); a shortened, revised version appears in Walker (1988:41–45). Question numbers in this paper correspond to the numbering of the predictor variables found in appendix G.

[13] This criterion variable actually consisted of a weighted average for each church based on the percentage of the community attending each church (Walker 1987:78). In discussing this variable with a number of fieldworkers, many thought that effect of even a single small church using vernacular Scriptures could have a disproportionate influence in the further penetration of vernacular literacy into the community. If their views are correct, then an unspecified number of communities may be underscored on this variable. This point could, of course, be verified or disproven in the course of further field research.

literacy acceptance. That is, the sale of vernacular literature does not necessarily tell us much about the use of that literature; indeed, if we array literate behaviors along a continuum, simple sales would fall on the low end of the scale. On the other hand, the public reading of vernacular Scripture in church probably indicates a great deal of community acceptance of vernacular literature; it may, or may not, however, be accompanied by individual literate behaviors in the vernacular. So we are left with a multiple regression model based on two criteria separated by a broad behavioral gulf, and presenting little information about individual literate behaviors.

Since it is doubtful that these two criterion variables alone give an accurate index of vernacular literacy acceptance, more work needs to be done on the model either to develop additional indicators to reflect the mid-range literate behaviors which are lost through the failure of criterion variables two and three (reading ability and amount of informal reading), or perhaps other predictors could be found which would be effective with these criteria.

The sample

Randomness. One of the key elements in evaluating research which purports to draw inferences from a statistical base is an examination of the sample and how it was drawn. Statistical methods are built on probability theory and depend for validity upon each event[14] having an equal opportunity

[14] An event in this situation is a literacy program meeting the criteria for selection. The population from which the sample should be drawn consists of all the literacy programs worldwide which would meet the criteria for inclusion. In fact, however, all the places where literacy programs are in place only represent a sample of the possible allocation sites where literacy programs could have been or will be initiated.

If fieldworkers have been using similar (stated or unstated) criteria in choosing their allocations, then those allocations may be systematically unrepresentative of the remaining potential allocations in unforeseen ways. For example, if a field entity had a policy of initiating work in all its rural language groups prior to allocating teams among urban-based language groups, then a survey carried out in that entity before all the rural allocations were filled would be totally biased toward rural allocation. This is a patently concocted example, but, however subtle, the possibility exists that through common training, ideology, romanticism, etc., we may have exercised bias in choosing earlier allocations.

The previous paragraph discusses possible bias as reflected in choice of allocation sites between language groups. There is also "within group" bias that may be represented in the sample, i.e., bias guiding the choice of one community as an allocation site over others in language groups characterized by multiple communities. So we can see that the communities represented in the study comprise a sample of possible allocations for those language groups represented in the study sample.

to be selected in the sample. This notion is commonly referred to as randomness. In a simple random sample, every potential member of the study has the same probability of being selected, and the selection is independent, i.e., the choice of one item will not affect the choice of another item. There are a number of ways to randomize the sample, and most books on statistics for the social sciences contain discussions of the significance of simple random samples and how to draw them.

In Walker's study, the population about which he wishes to generalize is all potential minority literacy allocations. It is obvious, however, that only a subset of the population can qualify for inclusion in the sampling frame, i.e., those allocations which have been filled and which have had a qualifying minimum of literacy work conducted. This is a reasonable limitation in the study even though there may be some unidentified biases represented in these programs (see footnote 13). Given this limitation, the most reliable sample which could be drawn would be a simple random sample of all (SIL) minority literacy programs worldwide.[15]

As we examine the composition of Walker's sample (1987:xi–xii; 72–74) in (1), however, we see that the 54 cases in the sample are drawn from just eight countries: Mexico, Guatemala, Brazil, Cameroon, Ivory Coast, Philippines, Indonesia, and Papua New Guinea.

(1) Sample composition (Walker 1987)

Country	Number of cases in study	Percent of cases in study
Mexico	14	26
Guatemala	4	7
Brazil	13	24
Cameroon	4	7
Côte d'Ivoire	1	2
Philippines	5	9
Indonesia	3	6
Papua New Guinea	10	19
Total	54	100

At no point does Walker inform us how the countries that were included in the study were selected; the only information given in the dissertation is that letters were sent to the SIL literacy coordinators in the countries

[15] The reason we would want to draw the sample from all projects worldwide is that the generalization Walker wishes to make is to all potential literacy programs worldwide.

represented soliciting their cooperation (Walker 1987:74). Given that no description of a randomization process is presented, it does not seem likely that the countries were randomly selected from a larger pool. This means that the sampling frame is restricted to just those countries shown in the table. Whatever criteria (convenience of the researcher, willingness of branch administrations to participate, etc.) were used to select the countries included in the study, the exclusion from potential samples of programs not located in these countries introduces serious bias into the study.[16] In fact, "a sample can only be representative of the population included in the frame" Fowler (1984:19). If these countries were not chosen at random, then the practical effect of drawing the sample just from them is that any generalizations issuing from the data can only be valid for programs from those eight countries.

A further problem emerges as we examine Walker's description of the respondent selection process (1987:73–74). We can see clearly that within the countries included in the sampling frame, there is no attempt to choose a random sample. Instead, the questionnaires were distributed to as many people as possible who met the minimum criteria of two vernacular Scripture publications distributed and the ability to answer the questionnaire. The sample consisted of those who were simply willing to spend the "one to three hours to complete the questionnaire" and then return it. This is commonly referred to by terms such as an availability sample, a convenience sample, haphazard sample, etc. The problems inherent in availability samples such as this one are discussed in Fowler (1984:20),

[16] Perhaps the best known case of problems arising from a poorly selected sampling frame is the infamous *Literary Digest* poll of 1936. In that instance, *Literary Digest* sent out ten million questionnaires to prospective voters asking their preference in the upcoming presidential election. With a response rate of about 2.4 million, *Literary Digest* predicted a landslide victory for Alf Landon (Republican) over Franklin Delano Roosevelt (Democrat) by a margin of 57% to 43%. When the actual vote was counted, however, Roosevelt had won by a margin of 62% for Roosevelt to 38% for Landon.

How could the results from such a large sample be so wrong? Most of the error has been attributed to a biased sampling frame. The sample was drawn from sources such as automobile registration lists, club membership rolls, telephone directories, and magazine subscription lists. While this might be a reasonable sampling frame today, in 1936 the country was polarized politically along economic lines. Republicans tended to be much wealthier than the more numerous Democrats, and the sampling frame (which was biased toward those with larger disposable incomes) was loaded with Republicans far beyond their proportion of the actual voting population (McClave and Benson 1985:918).

This example underlines a profound, but obvious, point. Large samples and statistically significant results are meaningless if the sample does not represent the relevant population.

Bernard (1988:97–98), and numerous statistics texts for the social sciences. One of the more succinct statements, however, can be found in de Vaus:

> Availability samples . . . are the least likely of any technique to produce representative samples . . . Using this approach anyone who will respond will do . . . This type of sample can be useful for pilot testing questionnaires or exploratory research to obtain the range of views and develop typologies, but *must not be used to make any claim to representing anything but the sample itself.* (de Vaus 1986:69) (emphasis added)

At least one other potentially serious source of bias remains. Walker, as far as I can ascertain, does not discuss how many people were sent the questionnaire but didn't return it. He implies, however, that some did not "take [the] one to three hours to complete the questionnaire." It is unfortunate that we do not know the rate of nonresponse. If the nonresponse rate is significant at all, then it is likely that there is nonresponse related bias reflected in the data.

Fowler (1984) reports with respect to mail surveys,

> that people who have a particular interest in the subject matter . . . are more likely to return mail questionnaires . . . This means that mail surveys with low response rates almost invariably will be biased significantly in ways that are related directly to the purposes of the research. (Fowler 1984:49)

Additionally, Fowler notes that although we cannot know much about the bias of nonresponders, "it is seldom a good assumption that nonresponse is unbiased" (Fowler 1984:52).

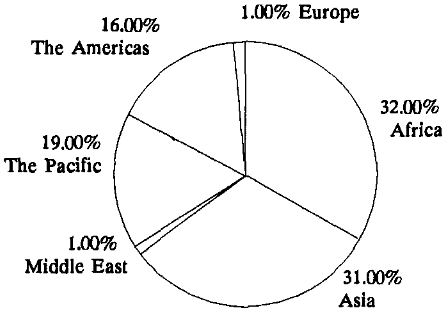
Representativeness. If the sample is not purely random, it should at least be representative of the population about which a generalization is to be made.[17] For a representative sample, the researcher attempts to predetermine (often through pilot studies or literature reviews) which natural categories may be relevant to the findings of the study, and randomly selects a portion of the sample from each of the categories.

As a very rough first estimate of how representative Walker's sample may or may not be, we can compare the geographical distribution of the world's living languages with the geographical distribution of the sample. Looking first to the figure in (2), we see the world's languages are partially

[17] In many cases it is advantageous to use a representative (or stratified) random sample rather than a simple random sample. See Fowler (1984:24–26), de Vaus (1986:57–59) or, especially, Babbie (1975:156–57) for discussions of the mechanics and advantages of stratified random samples.

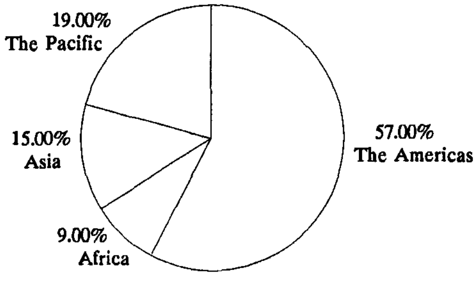
distributed in the following manner: Africa (32%), Asia (31%), the Pacific (19%), and the Americas (16%).

(2) Geographical distribution of living languages (source: Grimes (1984a:xvi))



When we compare (2) with (3), we see that the American continent with only 16% of the world's living languages is highly overrepresented, comprising 57% of Walker's sample. The Pacific is statistically represented just right if we only look at the percentages. If, however, we look beyond the numbers, we see that all the Pacific cases are from Papua New Guinea. It is an open question just how representative these language groups are of those found in other parts of the Pacific such as Polynesia or Micronesia. Looking to Africa we see an even more striking contrast. Africa which contains 32% of the world's living languages is represented by just 9% of Walker's sample, and that translates to only five languages from two countries in West Africa. East Africa is not represented nor is North Africa.

(3) Composition of Walker's sample (geographically distributed)



There are other factors besides geography—affiliation with which major religion (Islam, Hinduism, Buddhism, Christianity.), for instance, that we could use to examine the sample to see how representative it is, but it is unlikely that this sample could be construed as representative.

What this sample means for us is that we really have no basis for drawing statistical inferences from the dissertation study. More precisely, the results cannot be used to make generalizations about anything other than the programs it was drawn from. The statements of statistical significance are meaningless. It is not, however, a total loss; as a pilot test or exploratory research, a good deal can be learned from the data Walker gathered and analyzed. For instance, having to rely solely on criterion variables one and four as an index of vernacular literacy acceptance is not very satisfying (and perhaps only marginally valid). What has already been done can serve as a springboard for developing replacement variables for criterion variables two and three.

Statistical tests and the level of measurement

Given that the study may be repeated at a later date with a valid sample, it should be pointed out that some of the correlations are artificially high. This is due to the use of Pearson product moment correlations with dichotomous and ordinal variables (Walker 1987:94). In short, the Pearson's *r* makes strict assumptions about the data which ordinal and dichotomous data do not meet.

The way in which indicators are defined operationally in research also defines the level of measurement we attain in our data. This is important because the higher the level of measurement we attain, the more powerful statistics we may employ. Nominal variables are those in which the data are categorized into exclusive and exhaustive lists such as group membership, race, nationality, etc.[18] Dichotomous variables are a special subset of nominal variables which only have two categories, yes/no, gender, vernacular language/other language, etc.

Ordinal variables also produce data in categories that are exclusive and exhaustive, but the data are ranked as well. While the data are ranked, the distance between the ranks either has no meaning or cannot be ascertained. Scales such as low/medium/high, bilingual proficiency ratings, etc. produce ordinal data.

[18] The fairly ubiquitous nominal category, "other," is often used in social science research to fulfill the requirement that the categories be exhaustive. So, for example, a religion variable might have the categories: Christian, Jewish, Moslem, other.

Interval variables have all the properties of nominal and ordinal variables plus the characteristic that the intervals between values are equal.[19] For example, the ten degree interval from 60° to 70° is the same as the interval from 70° to 80°. Examples of interval scales are Fahrenheit and Celsius temperatures as noted, as well as population percentages, age, and weight.

Knowing the level of measurement allows us to choose the appropriate correlation coefficient for any two variables.[20] The figure in (4) illustrates the appropriate coefficients in tabular form. The different measures of association may themselves be ranked from those which make the least assumptions about the level of measurement of the variables to those which make the most assumptions about the variables.

(4) Appropriate coefficients based on data type (adapted from Fitz-Gibbon and Morris (1978:91))

		Variable 1		
		Dichotomous	Ordinal	Interval
	Dichotomous	Phi coefficient ϕ		
Variable 2	Ordinal	rank biserial r_{rb}	Spearman's rank order r_s	
	Interval	point biserial r_{pb}	Spearman's rank order r_s	Pearson's product moment r_{xy}

Phi incorporates the least assumptions about the data; it is accordingly the weakest of the measures of association. Pearson's r, on the other hand, incorporates the most stringent assumptions about the level of measurement; it requires interval data in both variables, and it is the strongest measure of association. If a coefficient is chosen which is based on weaker assumptions than the measures warrant, for example performing Spearman's r with two interval measures, then the result is still valid, but

[19] Ratio variables have all the properties of interval variables with the addition of an origin at absolute zero, but since there are no correlation coefficients which assume this higher level of measurement, they will not be discussed here. See Stevens (1946) for what is considered the modern "classic" treatment of the differences between levels of measurement.

[20] This discussion of the level of measurement is also relevant to the next step in Walker's analytical procedure, multiple regression on the criterion variables, since one of the basic validity assumptions of multiple regression analysis is that, "all variables are interval-level variables" (Loether and McTavish 1974:308).

it is understated. If, on the other hand, a coefficient is chosen that is too powerful for the measures being tested for association, for example using Pearson's *r* with two ordinal measures, then the result will be artificially high, and the reported correlation coefficient is not completely valid.

The table in (5) offers a categorization of Walker's predictor variables by their level of measurement and the appropriate coefficient for each when correlated with an interval level criterion variable. If Pearson's *r* were used on the dichotomous and ordinal level predictor variables, several of the correlation coefficients would be higher than they should be.

(5) Level of measurement and appropriate coefficients for predictor questions (scoring information from Walker (1987:87–94); table assumes the criterion variable is an interval measure)

Level of measurement	Predictor variables	Correlation coefficients
Dichotomous	11, 12, 13, 14	point biserial
Ordinal	4, 7, 8, 10, 17, 18, 19	Spearman's rank order
Interval	1, 2, 3, 5, 6, 9, 15, 16, 22	Pearson's product moment

Predictor variables 4 and 9 are of special interest. They appear to be interval data on the surface, but each is partly computed by multiplying a weighting factor, the bilingual level, which is clearly ordinal. The end result is a scale that appears to be distributed in even increments, but probably is not.[21] This is because the bilingual proficiency scale ranges from 0 to 5, and the increments are not generally considered equal. That is, the increment from 0 to 1 is not equivalent to the increments from 1 to 2, 2 to 3, etc.[22] The effect of multiplying a population percentage (which is interval data) by the bilingual level (which is ordinal data) is to create a composite number which serves to rank the communities after a fashion, but with respect to which the interval distance between scores is not interpretable in the same way as, for example, temperature, weight, age, and percentage.

[21] Essentially the same argument may be made concerning criterion variable 4, the usage of vernacular language Scriptures in churches. The scoring procedure for criterion variable 4 is similar to that of predictor variables 4 and 9 (Walker 1987:77). If one accepts that criterion variable 4 is essentially an ordinal variable, then (5) may be expanded using the information in (4) to create a column of coefficients for ordinal criterion variables.

[22] According to John Bordie of the University of Texas, Austin (personal communication), the increments implied by the FSI scale are probably more like the steps in a Fibonacci sequence. That is, rather than the increments being approximately equal, they are probably related in some (inexact or unspecified) geometric fashion.

In fact, the ranking of the composite number itself is not necessarily straightforward.

A hypothetical example illustrates the problem. For predictor 4, national language proficiency, the score consists of the average of the scores of each subgroup of the population (younger males, older males, younger females, older females). The score for each subgroup is determined by summing, "the products of the percentages of," the individual subgroups, "speaking the NL at a given level times that proficiency level," (Walker 1987:88). The table in (6) presents data for predictor 4 for three hypothetical communities.[23]

(6) Proficiency data and scores for three hypothetical communities for predictor 4.

	A		B		C	
Proficiency	%		%		%	
level	Population	Products	Population	Products	Population	Products
0	40	0				
1						
2					50	100
3			100	300	20	60
4					6	24
5	60	300			24	120
Score		300		300		304

As we can see from (6), the three communities have very similar scores, but their makeup is very different. Communities A and B each have a score of 300. In community B, bilingual proficiency is moderate and is spread evenly throughout the community. Community A, on the other hand, contains two groups. About 40% of the population is nearly monolingual, and 60% of the population has virtual native proficiency in the national language.[24] The majority of community C is characterized by moderate bilingual proficiency, and a significant minority have native-like proficiency. (Rensch's paper "Community language profiles" (this volume) provides a Pakistani example of such variation).

[23] The subgroups have been collapsed to simplify the example. A more complete example with data on all subgroups would actually reinforce the point made here that the measurement resulting from predictor 4 is not readily interpretable. That is, with the additional data, there is potential for considerably greater complexity.

[24] Though a community like A seems improbable at first glance, something similar could occur in a situation of extreme impermeable social stratification such as a caste society.

Three observations may be made based on the hypothetical data. First, the composite score thoroughly masks potentially significant social dynamics within the communities. Communities A, B, and C have nearly the same score, but are radically different in their bilingual behaviors. Second, the level of measurement is probably ordinal and not interval. That is, the distance between individual increments of the scale do not appear equal. Third, while the level of measurement is probably ordinal, it is not clear that all the rankings implied by the scale are meaningful or interpretable. There is a clear ranking across the extremes of the scale from 0–500 (no proficiency to perfect proficiency), but it is not apparent whether closely clustered scores as in (6) may be meaningfully ranked.

Reliability of the data

The data for Walker's study are primarily drawn from the responses to a questionnaire circulated to SIL fieldworkers. In evaluating the results of the study, it is important to (1) examine types of questions asked and (2) make some judgment about the level of reliability we can expect in the answers. The following discussion will not exhaustively address the individual questions from the questionnaire but, rather, will serve to indicate potential sources of error and lack of reliability.

The obvious key to evaluating most of the questions is to consider what the fieldworker is being asked to estimate and how reliable the estimate is likely to be. A number of the predictor questions are relatively straightforward and are matters which any two observers familiar with the community and its history should be able to agree upon independently. Examples of such questions are:

1. How many hours travel is it to a town where the national language (NL) is widely used?
16. List the number of symbols in the vernacular language (VL) orthography that are not found in the NL orthography or which have different phonemic values.

This type of question which involves direct measurement (with a watch or simply by counting discrete graphic symbols) should yield fairly reliable data. Most of the questions which have a bearing on the substantive issues of the study, however, call upon the respondent to make highly subjective estimates of things like the level of proficiency in the national language for different portions of the population (predictor 4) or the percentage of homes in the community where one spouse is not a mother-tongue speaker of the vernacular language (predictor 2). In some very small communities where the fieldworker intimately knows every individual, questions like predictor 2 may be answered with some reasonable degree of accuracy. In larger communities, however, there is little assurance that a fieldworker can estimate even relatively straightforward factors like the percentage of nonmother-tongue spouses accurately.

Part of the problem is that while the fieldworkers have the advantage of residing in or near the communities under study, they are not necessarily trained observers reporting the results of systematic observation or measurement. Some of them may actually be trained observers who have a prior interest in a portion of the data that Walker is asking for. The data from these fieldworkers on questions related to their prior interest will have a much greater probability of being accurate.

On the other hand, when we consider fairly complex behaviors such as estimating bilingual proficiencies of various subgroups of the village, as called for in predictors 4 and 9, it is not at all certain that any given fieldworker can make reliable estimates without conducting extensive testing.[25] In fact, following Huff's arguments, we should not even assume that the errors made by different fieldworkers will balance themselves out over the sample (1954:106).[26]

Looking a little more closely at predictor 4, we can see that there is a potential within some of the variables for compounded errors. First, the proficiency levels may be misjudged. That is, respondents may not be clear about what behaviors the categories imply. (As a variant of the problem, the categories are sufficiently vague that they may be conceptualized differently by the respondents). Second, estimates of the percentages of the populations at the various levels of proficiency may be wrong, possibly by large margins. (Contrary to daily experience, however, we might find some rare cases in which two wrongs do make a right). Predictors 9 and 15 and criterion 4 are subject to this potential of multiple errors of judgment, as well.

In evaluating the content of the questions to estimate the reliability of the data they return, we have to confront the fact that the attempt to "assess attitudes" without assessing attitudes is slightly fudged. Several of the predictor questions are, in fact, attitude questions. Note particularly that predictors 8, 10, 15, and 17 require the fieldworker to guess what the attitudes and motivations of others are with respect to diverse factors each of which are very likely to be sensitive to an array of social vectors.

[25] See SIL (1987) for an example of the complexities involved in obtaining a fairly accurate profile of community bilingualism based on the type of scale Walker proposes using.

[26] This is especially true in this case since we can't even make the assumptions that normally accompany a random sample.

Consider predictor 8, "how important do the people feel proficiency in the NL is to economic advancement?" No objective criteria are given by which the respondents may assess this "attitude" of the people. It is left, then, to the subjective estimate of the individual fieldworker. Respondent disposition such as personal optimism or pessimism will contribute to the outcome. The score for this item (and other attitude items) may be overly influenced by the respondent's more intimate acquaintances among the language community or by certain memorable events. The data, then, do not necessarily depend upon the range of response present in the community; the data depend upon the response raised in the fieldworker. When we consider the complexities of attitudes as presented above, it would seem that even a fieldworker familiar with a particular community would need to do specific investigation into attitudes to be able to provide reliable data on them. To answer these questions in the context of a field survey would undoubtedly require more than a few days of casual observation; it would require carefully crafted direct and indirect questions, interviews, and careful, systematic observation.

The point is that the data from the questionnaires should not be construed as reliably reflecting "objective" reality. Instead, for the most part, the data reflect the opinions of fieldworkers about factors which they have not measured. The only thing which these data may be said to reliably represent is the field worker's opinions about conditions in the village. In some cases these opinions may conform closely to what actual measurements of the real world phenomena under study would have been. In others they will not, but there is no way to judge the accuracy of the responses without careful, independent validation studies.[27]

When we realize the reliability of the data supplied by long term fieldworkers on many of the predictor variables is suspect, what then are the implications for surveyors who are conducting relatively brief and fleeting surveys? Only the very most obvious data, such as how long does it take to get to the next town, is likely to be accurate. Estimates, for example, of how many males between the ages of 10 and 25 speak the national language at an FSI level of three (part of predictor 4) are highly suspect from a local fieldworker if he has not indeed done a rigorous assessment of this group's bilingual abilities; as the subjective opinions of

[27] There is one attempt at validation mentioned in the dissertation (Walker, 1987:75), but since the validation consists of comparing questionnaires taken from three pairs of coworkers who presumably would have discussed many of the relevant issues during the course of program planning and implementation, we shouldn't assume the type of independence of response necessary to validate observer accuracy (or in this case, observer consistency). Validating observer accuracy, on the other hand, would require comparing observer response with actual measurements of the phenomena in question.

a surveyor, the estimates are virtually worthless. It is, in fact, a very complex and time-consuming process to accurately assess bilingual proficiency (and many of the other predictors, as well).

In planning (or evaluating) a survey or a study of this sort, there are several types of data which we might consider. The most reliable is that which is based on careful measurement of the population. For example, a thorough household census of the community should yield a fairly reliable picture of marriage patterns, professed language loyalty, etc. Likewise, thorough testing of bilingual proficiency throughout the community should give a reliable picture of the bilingual behaviors under scrutiny. The problem inherent in this type of data is that it is expensive and very time-consuming to gather.

The next most reliable type of survey data is that which is based on careful measurement of a random sample of the population. For larger communities, we can rarely perform measurements of entire populations. If the sample is random, however, we can at least state the statistical probability that our data are a reflection of the larger population. The point is, in our search for data, we are constantly confronted with the need to balance time and cost effectiveness against the degree of reliability necessary to inform our decision-making processes and program implementations. The figure in (7) is an attempt to graphically portray the trade-offs implicit in some of our choices.

There are many situations in which cheap data with a low reliability index (often referred to as "quick and dirty" data) are appropriate. Probably the key concern is what is the cost of making a wrong decision? If the cost of being wrong is low, "quick and dirty" data are probably the best choice. If, on the other hand, the cost of being wrong is high, such as bypassing a group that genuinely needs a program or filling an unnecessary allocation, then the cost of gathering more reliable data is justified.

(7) Cost-reliability ratio of various survey data sources[28]

+ Expense + Reliability			– Expense – Reliability
	measure population	measure random sample	measure representative sample
	opinions of insider populations	opinions of insider random sample	opinions of insider representative sample
	opinions of informed outsider population	opinions of informed outsider random sample	opinions of informed outsider representative sample
– Expense – Reliability			opinions of surveyors

Predicting vernacular literacy acceptance

As noted above, given the sampling problems connected with Walker's research, any generalizations, and, therefore predictions, based on this data are ill-conceived. However, this line of research appears to hold a great

[28] This chart is only meant to provide a rough rule of thumb. It is a relatively easy matter to think of exceptions. I believe the topology to be generally correct, however, and of course, there are other data sources which could be integrated. For example, availability samples could be included in an additional column on the right. A third dimension, complexity of data required, could be included in a three dimensional array to account for the differential reliability in answering questions like Walker's predictor one concerning travel time to nearest NL town versus the bilingual estimates called for in predictor four.

The general assumption of the table is that surveyors are collecting the data throughout. Thus in the final row, "opinions of surveyors" refers to instances where the surveyor provides opinions in contrast to data collected in higher rows or without reference to data from higher rows. Note also that this table begs the question of careful participant observation which is not generally within the purview of a relatively brief survey visit.

deal of promise and will probably be refined in the future.[29] Therefore the role of this model (and others like it) in prediction should be briefly addressed.

The theoretical underpinnings of Walker's research derive from the literature on language shift, and there is an overt desire to build a model with predictive power to use in assessing the potential for vernacular literacy acceptance. However, as Fasold writes concerning prediction of shift,

> Just as we saw in the case of language choice, however, where the same factors were cited independently by many scholars, there has been very little success in using any combination of them to predict when language shift will occur. In fact, there is considerable consensus that we do not know how to predict shift. (Fasold 1984:217)

It is unlikely that this attempt to predict vernacular literacy acceptance will prove successful either. Walker himself admits that his, "model... has not matured far enough to accurately predict how readily a community will accept VL literacy" (Walker 1987:202). Some crucial knowledge is surely lacking in our understanding of the processes of vernacular literacy acceptance, but the problem of prediction seems more fundamental than simply our temporary ignorance.

In a second paper in this volume, I address the mechanical aspects of correlation and prediction, noting that in correlational analysis prediction does not mean guessing right; it means guessing less wrong. It is always tempting to hope that we can progressively refine our predictive models until we get it right, but as Gregory Bateson points out in a major work on the epistemology of science, "the generic we can know, but the specific eludes us... There is a deep gulf between statements about an identified individual and statements about a class... and prediction from one to the other is always unsure" (Bateson 1980:45–46). Bateson's point is that the view that "a little more knowledge and, especially, a little more know-how will enable us to predict and control the wild variables... is wrong, not merely in detail, but in principle" (Bateson 1980:44).

Somewhat less eloquently, social forecasters, Richard Berk and Thomas Cooley, note,

> There is no disputing that forecasts of social phenomena will almost inevitably be wrong. Social phenomena are either inherently

[29] Unfortunately Walker's current program (1988:35, 41–45) for gathering more data by inviting whoever will to send additional data to him is doomed to perpetuate the sampling and response bias problems noted above. It will not produce a sample from which valid statistical inferences can be drawn.

> stochastic or as a practical matter, must be treated as such ... Most of the time forecasts will be wrong. (Berk and Cooley 1987:247, 263)

Walker's "model" is not a model in a technical sense. A technical model consists of an evaluation index derived from the regression analyses. That is, a model should contain a formula into which the predictor scores are fed to produce an index for vernacular literacy acceptance. A model is necessary to interpret the mass of data produced when a survey is undertaken to assess the prospects for vernacular literacy in a language community.

If we gather the data on the predictors, how do we know what it means? Some factors may appear positive; some may appear negative. How do we tell when the mixed positives and negatives mean that vernacular literacy acceptance is unlikely? Without a model and a way of interpreting it, there is no basis for using the material in decision-making. Walker writes concerning the use of this "model" in Irian Jaya, "at this point, we have not adopted a formula. Our predictions of VL literacy acceptance and decisions regarding priority language project status are still a product of subjective evaluation" (Walker 1991:86).

Essentially this means that there is no standard based on Walker's model for evaluating the scores a community receives as it is assessed. Language program decisions based on interpretation of survey data on these predictors are, then, *ad hoc*. This is because the "model" does not predict well enough for reliable decision-making. It also means that language groups which are subjected to an administrative evaluation using this data are not receiving a disposition based on objective criteria; rather, they are left to the vagaries of the subjective evaluation of whoever happens to be the evaluator at the time.

Conclusion

Walker's claim has been that this line of research provides an effective means of assessing attitudes. The discussion above has shown that the model in its present form is too impoverished to adequately reflect important intercommunity variation in attitude structures. Additionally, the sample was drawn in such a way that it has no external validity, and, therefore, generalizations and statistical inferences have no meaning for any cases other than those in the original study.

Even if there were no sampling problems or other potential problems with the reliability of the data from some of the questions, the model's inability "to accurately predict how readily a community will accept VL literacy" (Walker 1987:202), and the lack of a technical evaluation model to assess the data for future surveys severely restrict any effectiveness it might have as a decision-making tool.

The technical criticisms of Walker's research should not be taken to mean that the research is of little worth or that it shouldn't have been done. On the contrary, the research is original and likely to prove quite valuable. Though the model should probably not be used as a forecasting tool, the information it generates may provide some of the levers language planners and vernacular literacy workers need for improving the acceptance of vernacular literacy in many minority-language communities.

Of particular interest is Walker's finding that orthography design and the involvement of community leaders can significantly improve the chances for vernacular literacy acceptance. While this may seem intuitively correct, the regression analysis has confirmed that it is, in fact, correct for the 54 programs in the study. If this observation proves generalizable beyond Walker's study, then there is reason to hope that vernacular literacy may be successfully introduced into communities with clear comprehension needs but poor prospects for accepting vernacular literacy. Likewise, Walker's finding that the "percentage of the community... who aim at living their lives according to the Bible," (predictor 15) significantly and positively affected the acceptance of vernacular literacy, holds out hope that spiritual change outside the direct responsibility of the literacy worker can lead to greater acceptance of vernacular literacy.

Walker has done us a considerable service by probing beyond our previous level of knowledge about the processes of literacy penetration into minority language groups. Hopefully, we have been provided with tools which can help us improve the acceptance of vernacular literacy in those communities where translation and literacy programs are needed.

[blank]

Understanding Correlation[1]

Dale Savage

In recent years there has been an increasing use of correlation in research in SIL. For many without statistical training, however, the uses of correlation, the interpretation of correlation results, and the role of correlation in prediction are poorly understood. At least one earlier study by an SIL member, Kroeger (1986), explores several ways in which problems associated with the units of data and their abstraction can yield misleading correlation results.

Coefficients of correlation index the extent that two or more variables co-vary or co-relate, that is, they measure both the degree and direction that one variable changes concomitant to the change in some other variable. These coefficients allow us to answer questions such as, is there a relationship between changes in one variable and changes in the second variable, and do the two variables vary together in some systematic way? Where a relationship exists, correlational analysis often allows us to build predictive models which may be useful in decision making. The statistical function in correlational analysis provides a measure of the strength and direction of the relationship between the variables under study.

[1] I wish to express appreciation to all those who have in one way or another lent assistance in preparing this manuscript. My thanks to Bob Bodwell, a professionally trained statistician, who discussed many of the issues presented here with me before I began writing; thanks also to my wife, Lorraine, who read drafts for coherence. I wish also to thank Barrie Wetherill, Director of the Industrial Statistics Research Unit of the University of NewCastle upon Tyne, who kindly read a late draft of the manuscript, and finally, I wish to express my deep appreciation to the editor of this volume, Eugene Casad, for his constant and patient encouragement during the preparation of this manuscript. Any remaining errors or oddities are, of course, my own.

Correlation strength is measured on a scale ranging from zero to one or negative one. The higher the absolute value of the correlation, the stronger the relationship that exists between the variables. Therefore, a correlation of zero indicates no relationship, and a correlation of one or negative one indicates a perfect relationship.

When increases in one variable are paralleled by increases in the other, the relationship is said to be positive, and when increases in one variable are paralleled by decreases in the other, the relationship is said to be a negative, or inverse relationship.

A correlated set of variables that we take for granted will serve to illustrate these points. For a wage laborer, there is a perfect positive correlation between "earnings" and "hours worked." So, for example, if his wages are $4.00 per hour, he will earn $8.00 for two hours, $20.00 for five hours, $40.00 for 10 hours, etc. On the other hand, there is a perfect negative correlation between "hours worked" and "hours available for other pursuits." The more hours spent on the job, the less hours are available for other pursuits.

Correlation is an extremely useful tool in model building of all kinds. Theoretical hypotheses may be tested by choosing indicators for various components of the theory and analyzing their behavior to see if the correlations between them correspond to the theory's prediction. Similarly two theoretically related concepts may be measured in an attempt to understand the nature of the relationship between them. Simons (1977), Kroeger (1986), and Grimes (1988a and this volume) collectively represent an excellent example of the progressive exploration of the relationship between two variables, "lexical similarity" and "intelligibility" between languages.

Correlation also serves the pragmatic purpose of building predictive decision models. College entrance committees, for example, use correlation-based models to predict which candidates for acceptance are likely to succeed in their programs. These models use factors which can be, or have been, measured to predict possible outcomes of particular programs. Landin (1989) is an example of the type of correlational studies which may precede the formulation of decision rules. Grimes (1988a:29 and this volume) provides an example of a correlational analysis culminating in a decision model for testing to determine program need.

Yet another common use of correlation is to compare measurements of some phenomenon obtained by different instruments. Often in this case an instrument which is known to be valid and reliable is used to validate the results of a newer instrument which is perhaps less expensive or less time-consuming to administer. Kamp (this volume), James, Masland, and Rand (1989), and Quakenbush (1988 and this volume) use correlation to study different means of assessing bilingual proficiency. Radloff (1991) uses correlation as an integral part of the test calibration strategy for the Sentence Repetition Test.

At this point, we should clearly note that correlation is never proof of causation. Causal links are properly the domain of grounded and motivated theories and pragmatic, common sense knowledge of the everyday world. There may truly be a cause and effect relationship between variables such as that between rainfall amounts and crop yields, but the correlation itself can only indicate the degree and direction of covariance not causality. Cause and effect may only be inferred by appeal to factors external to correlation.

In some cases, two correlated variables may both be caused by common factors. For example, weight and height show a positive correlation among children, but few people would claim that weight causes height or the converse. It is likely, though, that both are at least partly caused by factors such as nutrition and heredity.

It is also the case that there are abundant spurious correlations involving variables which happen to change in tandem with no discernible causal links. Armchair analysis notwithstanding, no one has ever discovered a convincing causal link in the well-known correlation between "hemline length" and "market vitality on Wall Street" where rising hemlines in the fashion industry correlate well with cycles of vigorous market activity.

In this paper, I briefly review some of the research within SIL which uses correlation as a technique for data analysis. In addition, I attempt to walk the reader through a simple correlation calculation and discuss the meaning of the resulting numbers. The aim is not to teach the skill of calculating correlation,[2] rather it is to attempt to provide insights on how to think about correlation by examining some of the calculations underlying its results. Finally, I close with a few cautions regarding the need for the judicious interpretation of correlation figures obtained from the analysis of SIL language survey data.

Studies within SIL

There are a number of studies within SIL which have used correlation as a primary mode of data analysis. Simons (1979), Kroeger (1986), and Grimes (1988, and this volume), for example, examined the question to what degree can lexical similarity be used as a predictor of intelligibility.

[2] Given the number of adequate statistics programs for both IBM clones and Macintoshes available at reasonable prices, we should generally be doing the calculations by computer, but most common statistical functions can be calculated using a hand-held calculator with a square root function.

Simons (1979) used secondary data analysis of intelligibility scores and lexical similarity counts from numerous languages around the world to build a model which accounts for about 65% of the variance between lexical similarity and intelligibility in the sample.

Kroeger (1986) applied the same method to a large data set from Sabah, but found it only accounted for about 44% of the variance in that sample. It is important to note here that Kroeger's study is based on significantly more reliable data than that used by Simons. Where Simons used intelligibility figures that were derived in several distinct ways, Kroeger's figures come from a single large-scale survey which used a consistent method for arriving at an intelligibility index throughout. In other words, Kroeger's study is based on data which are strictly comparable, whereas the data from Simon's study are not.

Kroeger also pointed out that our SIL intelligibility and lexical similarity data sets are usually truncated[3] with some attendant skewing of the results. A very interesting set of correlations between geographic distance and intelligibility is presented as well. For some of the language groups in

[3] Kroeger observes that we normally do not test intelligibility when lexical similarity counts are below about 60% to 70%. Additionally, in the Sabah survey, intelligibility testing was not usually carried out when lexical similarity was above 90%. This exclusion of portions of the range of variation within the population from the sample has the effect of skewing the overall correlation results.

While it is true that considering only a truncated range of scores will produce a different correlation than when we consider the entire response range, the correlation produced is not internally skewed. It is only skewed in relation to the entire range of possible scores. If we are using lexical similarity figures as a portion of our program decision process, it is valid and actually more useful to consider truncated ranges of scores rather than the entire range of scores. For example, it is more helpful in decision-making to know the correlation between lexical similarity and intelligibility in the range from 60% to 80% similarity than it is to know the correlation for the entire range of lexical similarity from 0% to 100%. As Grimes (this volume) points out, prediction of intelligibility scores from relatively high lexical similarity counts is much less reliable than prediction from low lexical similarity counts. This leads to a simple decision rule: if lexical similarity is below a specified threshold, then intelligibility testing is not necessary (because intelligibility will likely be too low for our purposes); if lexical similarity is above a specified threshold, then further testing is necessary (because our` ability to predict intelligibility from upper range lexical similarity scores is not very reliable).

Kroeger's study, geographical proximity actually serves as a better predictor of intelligibility than lexical similarity.[4, 5]

A very significant feature of Kroeger's article is his discussion of some of the sources of error which may arise in correlational analyses of recorded text tests for intelligibility (1986:334–36). The sources of error outlined by Kroeger are: (1) "sampling bias" when bilinguals are present in the sample, (2) "non-normal distribution of the data" covering truncated data sets and the attendant skewing when bilinguals are included in the sample, and (3) "masked variation" caused by averaging the scores from individual test locations prior to performing the correlations. Additionally, Kroeger cautions us to peer behind the abstractions presented to us by the numbers, and to consider what they really represent. He comments as follows, "In order to interpret the statistics in any meaningful way, we must know quite a bit about the units of data, how they were collected, and what these measurements represent" (1986:310). These cautions are important because our necessary transformation of real world phenomena into numerical data is a process of abstraction compounded by our ability to perform arithmetic operations which further distance the numbers from their real world referents.

As an example, Simons' (1979) research reported above claims to have accounted for 65% of the variance between lexical similarity and intelligibility in his sample, but there are at least two problems any careful evaluator must consider concerning this conclusion. First, Kroeger's "masked variation" caused by averaging the scores from individual test locations is present. That is, the numbers used in the correlation calculations have already been stripped of a great deal of their variance by

[4] This probably represents a bilingual overlay on the inherent intelligibility of closely related languages, a notion which had not been widely discussed in SIL when Kroeger was developing the ideas presented in this paper but which Kroeger mentions briefly on page 315 and addresses more fully on page 324.

[5] On page 323, Kroeger points out the very interesting case of Kuijau in which the subjects (contrary to the trend among the other languages discussed) understood their neighbors better the farther away they lived. Kroeger points out some very plausible reasons for the phenomenon, and it stands as a caution against mindlessly using an oversimplified regression model for prediction. That is, the attempt to predict "intelligibility" of Kuijau with its neighbors based on the regression model for distance would lead to conclusions which are the opposite of the actual test results. Even regression models with very nice overall results may occasionally lead to very significantly wrong predictions.

averaging the variation away in each location.[6] Second, when we look at "the units of data, how they were collected, and what these measurements represent" (Kroeger 1986:310), we find that, although all the intelligibility scores (which are derived from several distinct studies) are reported on the same percentage scale and figure in the calculations as though they are the same thing, the intelligibility scores are actually the result of several different types of testing and scoring methods (Simons 1979:143–63 and 5–31). Thus the coefficient of determination in Simons (1979) rests upon scores which, in their abstract numerical form, are the same, but which represent very different real world measures. The extent to which the different measures are directly comparable is not a settled question. In short, Simons' results must be interpreted and applied with reservation.

Grimes (1988a and this volume) reexamined the question of how good a predictor of intelligibility is lexical similarity. By separately examining various ranges of the predictor variable (lexical similarity), Grimes argues that low lexical similarity (i.e., 60% and lower) is a good predictor of inadequate intelligibility, but that a high degree of lexical similarity is a poor predictor of adequate intelligibility.[7]

Another group of studies, Kamp (this volume), James, Masland, and Rand (1989), Quakenbush (1986, and 1988 and this volume), and Radloff (1991), use correlation to study the relationship between different means of assessing bilingual proficiency. The general aim of these studies is to find cheaper, easier-to-administer alternatives to recorded text tests and proficiency interviews.

Kamp (this volume) used correlation to test the relationship between measured bilingual proficiency (through proficiency interviews) and a "self-score" test which called upon a group of Karao speakers in the Philippines to rate themselves on bilingual ability in Ibaloi. He also used correlation to test the relationship between measured bilingual proficiency (through proficiency interviews) and a "self-test" questionnaire which called upon the respondents to rate themselves on bilingual ability using a range of behavioral criteria similar to those tested during the proficiency interview.

James, Masland, and Rand (1989) measured bilingual proficiencies in Wolof for a group of thirty-two Serer-Safen speakers in Senegal using three different methods: recorded text tests, questions on isolated Wolof

[6] Most of the averaging is not Simons' fault, but is an artifact of the reporting in the sources for his study. That is, not all of the investigators who carried out the studies that Simons cites arrived at their numerical results via identical methods, with the result that the percentage scores that they cite are not comparable from study to study.

[7] Preliminary data on twenty-seven pairs of test scores from the Dobel survey of Eastern Maluku begun by Jock Hughes and Eugene Casad appear to corroborate Grimes' figures from the Philippine surveys (Eugene H. Casad, personal communication).

sentences, and proficiency interviews (SLOPE). Of the three methods, SLOPE most directly measures bilingual proficiency and is probably the most reliable. It also requires the heaviest investment of personnel and time to administer. The thrust of the research then was to determine whether recorded text tests (normally used for measuring intelligibility rather than bilingualism) or questions on isolated sentences could be reliably substituted for full blown proficiency interviews. James, Masland, and Rand found that the recorded text test scores and the sentence test scores did not exhibit high enough correlations with the SLOPE scores to allow replacement of SLOPE with these methods, requiring less time and labor.

Quakenbush (1988 and this volume) describes one portion of a larger study conducted among the Agutaynen people in the Philippines in which self-reports of bilingual proficiency were calibrated using a correlational analysis against proficiency interviews based on Foreign Service Institute interview techniques (Quakenbush 1986). Self-reports were much more time efficient for the larger study, but the correlation with the more time-intensive proficiency interviews was necessary to provide a basis for interpreting the self-report data.

Radloff (1991) has actually integrated correlational analysis as an integral part of the methodology in developing and calibrating the scoring of a series of sentences to test the bilingual production capabilities of different linguistic groups. In their methodology, a group of subjects are pilot tested on their ability to produce a large number of sentences in a second language. These subjects are also tested on their second-language abilities using a discrete, descriptive scale. The results of the two tests are then correlated to extract a list of fifteen sentences which form the SRT<D>Sentence Repetition Test (SRT). The sentences of the SRT are chosen for their ability to discriminate the levels of bilingualism differentiated on the descriptive scale. Though time intensive in development, the SRT has the advantage of being easy and quick to administer and simple to interpret.[8]

Other studies such as Landin (1989) and Walker (1987) explore the relationships between various social factors and successful programs. Landin (1989) presents correlations between "acceptance of SIL's vernacular written materials" and "vernacular language usage and encouragement by missionaries" working in approximately 104 language groups in several countries. The findings Landin presents are part of a larger study of acceptance of vernacular-written materials. Correlations between acceptance

[8] Barbara Grimes (this volume) provides a brief critique of sentence repetition tests for evaluating bilingual proficiency, noting several features of proficiency that repetition tests cannot evaluate.

and some of the other factors from the broader study are presented in Table 2 (page 160) of his article.

Walker (1987) reports the multiple correlation results of an attempt to build a predictive model of vernacular literacy acceptance for minority language groups.[9] In this type of research, criteria for "vernacular literacy acceptance" are first defined. Then predictor variables are correlated against criterion variables. Multiple correlations, such as Walker uses in his analysis, are interpreted in much the same way as simple correlation, but include additional variables to increase the amount of variance accounted for and improve the predictive power of the correlation.

Exploring the meaning of correlation

By means of an example, we will explore the role and relationship of the correlation coefficient (r)[10] and the coefficient of determination (r^2) in prediction. In arriving at an understanding, we will consider the mean ($\bar{x}$), variance (s^2), and standard deviation (s) of a single variable. Then we will add a predictor variable and consider scattergrams, regression, proportional reduction of error, z-scores, and the correlation equation itself.[11]

Let us consider hypothetical attitude scores toward a second language for a ten-subject sample. For now we will compute and discuss the mean, variance, and standard deviation for the attitude scores alone. The necessary data is contained in (1) below.

[9] For a more thorough treatment see Savage (this volume), *A Review of Walker's Research.*

[10] There are actually several different types of correlation coefficient. The two most commonly used are the Spearman rank order correlation (rs) and the Pearson product moment correlation (r). The choice of which type of correlation a researcher uses depends on the nature of the data itself. The figure in (4) of Savage (this volume) *A Review of Walker's Research* displays the appropriate correlation coefficients based on data type. See especially Fitz-Gibbon and Morris (1978:90–92) for a discussion of how to choose the proper measure of association.

[11] It is not necessary to perform all the following calculations to find the correlation coefficient, but the purpose of this section is to illustrate the logic underpinning the results of a correlational analysis. In most cases the correlation coefficient is simply figured directly from raw scores.

(1) Data for calculation of standard deviation and variance of attitude scores

Case number	Attitude score x	Mean $\bar{x}$	$(x-\bar{x})^2$
1	30		256
2	40		36
3	35		121
4	40		36
5	45		1
6	35		121
7	40		36
8	55		81
9	60		196
10	80		1156
N = 10	Σ = 460	46	Σ = 2040

The mean is already computed, but for consistency's sake, the formula is included here.

$$\bar{x} = \frac{\Sigma x}{N} = \frac{460}{10} = 46$$

In prose the formula reads, the mean is equal to the sum[12] of the x values divided by the number of measured observations. If we wish to predict an attitude score for any particular member of the population, our best guess is the sample mean. There will be errors in our prediction, however. In the case of our attitude scores, in fact, none of the scores actually fall on the mean. To understand the scores we've obtained, it is necessary to know more than the mean; we also need to assess our error from the mean.

The most useful, though not directly interpretable, estimate of the error is the variance, s^2, which is based on using the mean as a best guess. Variance may be defined as the mean of the squared deviations from the

[12] Here and in the other formulas sigma, Σ, simply means "the summation of." For example Σx means to add up all the values of the variable x.

mean.[13] It is computed by subtracting the mean from each score and squaring the result, adding up the squared result for each score, and dividing by the number of subjects minus one.[14] The variance of the attitude scores is:

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{N-1} = \frac{2040}{10-1} = 226.67$$

In addition to leading us into the standard deviation of the scores, the variance gives us an index to the variation in the attitude scores which we will use in computing the proportional reduction of error in the attitude variable after regression analysis.

The standard deviation (s) is the square root of the variance:

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N-1}} = \sqrt{\frac{2040}{10-1}} = 15.06$$

The value of the standard deviation lies in its interpretation of the dispersion of cases in a normally distributed sample. More precisely, the distribution of the scores obtained from such a normally distributed sample can be approximated on a chart by a bell-shaped curve. This allows one to specify precisely particular quantities that correspond to distinct areas underneath the curve.

Thus the standard deviation divides the deviance from the mean into zones[15] wherein one standard deviation above and below the mean will encompass about 68% of the cases. A zone of two standard deviations above and below the mean will encompass about 96% of the cases, and three standard deviations above and below the mean will encompass about

[13] Squaring the deviations removes the negative signs from deviations below the mean. Otherwise the sum of the deviations would always be zero. The same thing can be accomplished by simply adding the absolute values of the deviations. This gives rise to the average or mean deviation. The average deviation is easy to interpret at a glance.

$$AD = \frac{\Sigma|x-\bar{x}|}{N-1} = 12.67$$

The average deviation, however, is of little use in computing more advanced statistics. The standard deviation and the variance, on the other hand, play an integral role in computing numerous stable statistics including correlation (Phillips 1988:31–35).

[14] Subtracting one from the number of subjects in a sample is a technique to compensate for a slight bias which arises in sampling.

[15] The standard deviation is expressed in the same scale units as the cases or scores being described. So then the standard deviation, about 15 in our case, may be added or subtracted directly from the mean, 46, to find the distributional zones around the mean. In this case, one standard deviation above and below the mean encompasses the range of scores from 31–61.

99% of the cases. Additionally, the standard deviation can be converted into a standardized measure, the *z*-score, which allows scores on different scales to be compared. We will encounter *z*-scores in the correlation equation as we add another variable to our table.

Trying to predict attitude scores based solely on our sample mean is a shot in the dark. If, however, we can identify another factor that covaries with attitudes toward the second language, we can improve our prediction. Correlational analysis provides us with an index of the direction and strength of the covariation and a measurement of the improvement of the prediction.

Let us assume that the number of years of education in the second language covaries with an individual's attitudes toward the second language. In (2) we have added education figures for each subject.

(2) Data for plotting regression

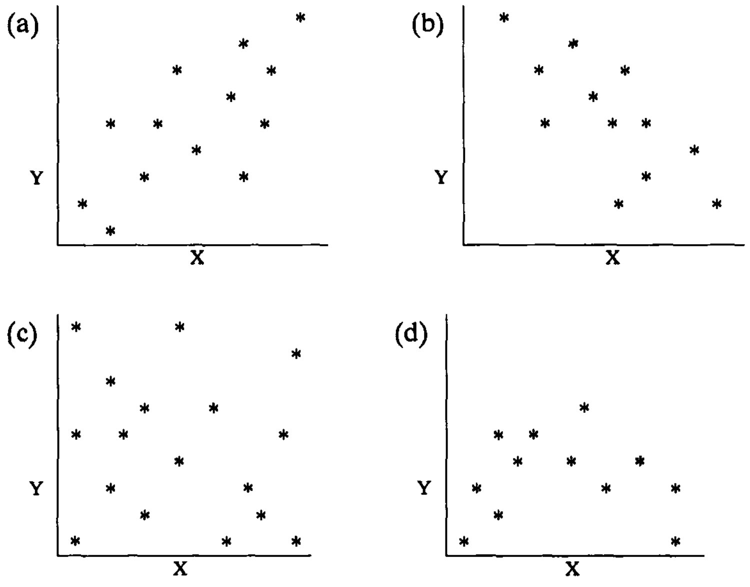
Case		Attitude				
number	Education x	score y	Mean $\bar{x}$	Mean $\bar{y}$	xy	x^2
1	0	30			0	0
2	0	40			0	0
3	3	35			105	9
4	4	40			160	16
5	6	45			270	36
6	6	35			210	36
7	6	40			240	36
8	8	55			440	64
9	12	60			720	144
10	12	80			960	144
N = 10	Σ = 57	Σ = 460	5.7	46	Σ = 3105	Σ = 485

The first step in any correlational analysis is to make a scattergram of the data. A scattergram is a graph where an *xy* coordinate is plotted for each case using the variable you are trying to predict as the *y* variable and the variable you are correlating with it as the *x* variable. The *x* variable is also known as the independent variable, and the *y* variable is known as the dependent variable. This is because we independently vary the *x* variable while looking for corresponding variation in the *y* variable. Thus our experimental design sets up a situation in which the *y* variable is dependent on the *x* variable. In our example, the attitude scores are the individual instances of the dependent variable and the educational level in the second language is the independent variable.

The common correlation measures assume that the data points lie along a single, generally linear path. If the linearity assumption is not met, i.e., the data points do not lie in a single direction,[16] then a different statistic, eta (*h*), which tests nonlinear covariation is called for. It is important, therefore, to examine scattergrams to insure that the proper statistic is applied to the data.

The figure in (3) provides us with a view of four different distributions: (3a) and (3b) demonstrate linear relationships between variables *X* and *Y*. That is, the data points lie in an approximately straight line path; (3c) demonstrates little or no relationship, and (3d) demonstrates a nonlinear, or curved, relationship. In addition, we may note that (3a) demonstrates a positive relationship, and (3b) demonstrates a negative relationship. That is, in (3a) when *X* increases, *Y* increases; in (3b) when *X* increases, *Y* decreases.

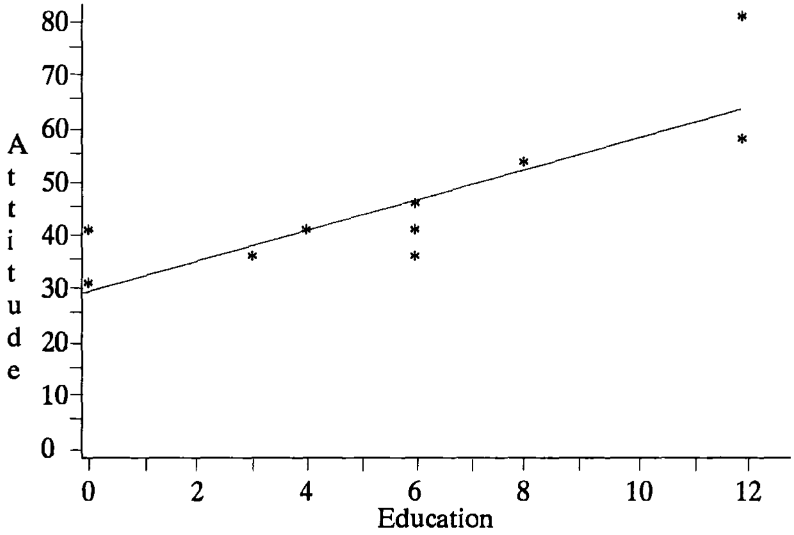
(3) Different association relationships illustrated by scattergrams



[16] One example of a nonlinear relationship is that between agricultural harvests and rainfall during the growing season. To a certain point, as rainfall increases, crop harvests also increase. For most crops, however, above a certain amount of rainfall, crop harvests will decline. Too little rain and too much rain both lead to poor harvests. This type of relationship will produce a scattergram roughly shaped like an inverted U.

The figure in (4) shows the scattergram of the education and attitude variables along with the regression line for the plot. An examination of the plot shows that the relationship is linear and positive.

(4) Scattergram with regression line



In (4) note also the regression line drawn among the data points. The regression line lies along a path which represents a best prediction of *y* values from *x* values. It is an idealized straight line lying along the least squared distance from all the data points. That is, the regression line lies along the mean of the smallest possible sum of the squared distances from each of the actual data points. (Recall that the variance in a single set of measurements is the mean of the squared deviations of each data point from the mean of the entire set). Thus the regression line, or "best fitting line," represents a line of prediction which minimizes the variance (estimate of error) in the data set by taking account of a second variable. Analogous to the mean as a best guess for a single factor, the regression line is the best guess for *y* values when given the *x* values of a particular case.

The formula for the regression of *y* on *x* is:

$$y = a + bx$$

where *a* and *b* are both constants. The constant, *b*, describes the slope of the regression line or how many units on the *Y* axis the line rises for each unitary increase on the X axis. It is computed using the formula below and data from (2) above.

$$b = \frac{N(\Sigma xy) - (\Sigma x)(\Sigma y)}{N(\Sigma x^2) - (\Sigma x)^2} = \frac{10(3105) - (57)(460)}{10(485) - (57)^2} = \frac{31050 - 26220}{4850 - 3249} = 3.02$$

The constant, a, is the Y intercept. It is the point at which the regression line crosses the Y axis ($X = 0$) and is computed using the means of both x and y from (2) and the value of b computed above.

$$a = \bar{y} - b(\bar{x}) = 46 - 3.02(5.7) = 46 - 17.21 = 28.81$$

Returning to the regression formula, we can now substitute the calculated constants, a and b.

$$y = a + bx = 28.81 + 3.02\ (x)$$

This is our model for predicting y values. In fact, the regression line falls on the predicted y values. By substituting each x value for x in the formula, we can calculate our predicted y values and calculate an error index which may be compared to the variance of the y scores alone. The variance of the attitude scores and the regression error will then be used to figure our proportional reduction of error.[17] The table in (5) contains the data needed to compute the proportional reduction of error.

(5) Data for proportional reduction of error

Education	Attitude score	Total variance	Predicted y	Regression error
x	y	$(y - \bar{y})^2$	$y = a + bx$	$(y - (a + bx))^2$
0	30	256	28.80	1.43
0	40	36	28.80	125.35
3	35	121	37.85	8.15
4	40	36	40.87	0.76
6	45	1	46.91	3.63
6	35	121	46.91	141.73
6	40	36	46.91	47.68
8	55	81	52.94	4.25
12	60	196	65.01	25.06
12	80	1156	65.01	224.81
	$\bar{y} = 46$	$\Sigma = 2040$		$\Sigma = 582.85$

[17] A general discussion of proportional reduction of error in measures of association is found in Mueller, et al. (1977:191–94).

There are several proportional reduction of error (PRE) measures all of which follow the general formula

$$\text{PRE} = \frac{E_1 - E_2}{E_1}$$

where E_1 is the old error and E_2 is the new error. In the case of correlation, the PRE formula incorporates the variance of y as E_1, and E_2 is equal to the regression error. In the following calculation we use the total variance and total regression error for convenience.

$$\text{PRE} = \frac{2040 - 582.85}{2040} = .714$$

We can multiply the PRE result by 100 to arrive at a percentage score. Using the regression formula we have reduced our error in predicting attitude scores by 71%. Note that this does not mean that we have improved our prediction by guessing the attitude score right 71% of the time; it means we have improved our prediction by reducing the size of our error (variance) by 71%. If we visualize our error based on just the mean of the attitude scores as an area covering a football field, then taking account of the number of years of education of our subjects confines our error to the area from the goal line to the thirty-yard line on one end of the field.

Having laid the groundwork for understanding the correlation coefficient (r) and the coefficient of determination (r^2), we will now calculate the Pearson Product Moment correlation for our data set using data from (6).

(6) Data for calculating Z-scores and the Pearson product moment correlation

Education	Attitude score					
x	y	$x - \bar{x}$	$y - \bar{y}$	z_x	z_y	$(z_x)\ (z_y)$
0	30	−5.7	−16	−1.35	−1.06	1.44
0	40	−5.7	−6	−1.35	−0.40	0.54
3	35	−2.7	−11	−0.64	−0.73	0.47
4	40	−1.7	−6	−0.40	−0.40	0.16
6	45	0.3	−1	0.07	−0.07	0.00
6	35	0.3	−11	0.07	−0.73	−0.05
6	40	0.3	−6	0.07	−0.40	−0.03
8	55	2.3	9	0.55	0.60	0.33
12	60	6.3	14	1.49	0.93	1.39
12	80	6.3	34	1.49	2.26	3.37
$\bar{x}=5.7$	$\bar{y}=46$			0.00	0.00	$\Sigma=7.61$
$s_x= 4.22$	$s_y=15.06$			$z_x{}^2=9$		

Although there are formulas for calculating Pearson's r from raw x and y scores,[18] we will use the basic formula which is built around z-scores since it highlights certain facts about r and r^2 which will be discussed later.

$$r = \frac{\Sigma z_x z_y}{N-1}$$

In prose, the formula reads as follows: the correlation coefficient is equal to the sum of the products of each z-score of x, times its corresponding z-score of y, divided by the number of cases minus one. This means that r is equal to the mean (adjusted for bias) of the products of the z-scores of x and y.

Z-scores, also known as standard scores, are the measure of the distance that each individual score on a variable differs from the mean. Z-scores are expressed as a multiple of the standard deviation.[19] We find the z-scores by first subtracting the mean from a score. Then we divide the result by the standard deviation. The formula for z is listed below.

$$z = \frac{x - \bar{x}}{s}$$

Using data from (6) and the formula to derive the z-score for the educational level (x) for the first member of our sample we find:

$$z = \frac{x - \bar{x}}{s} = \frac{0 - 5.7}{4.22} = \frac{-5.7}{4.22} = -1.35.$$

The same process is repeated for each x score to derive the scores listed in the z_x column of (6). We use the same formula to derive the z_y column. We illustrate with the attitude score (y), again, of the first member of the sample.

$$z = \frac{y - \bar{y}}{s} = \frac{30 - 46}{15.06} = \frac{-16}{15.06} = -1.06$$

[18] The following formula for calculating Pearson's r from raw scores is included for comparison. It is usually more convenient to calculate r directly from raw scores rather than converting them into z-scores.

$$r = \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

[19] In normally distributed data (and even in data with mound-shaped distributions which are slightly skewed or asymmetric), the following distribution of data points is implied. About 68% of the data points will fall within a range of 1 standard deviation above and 1 standard deviation below the mean; about 96% will fall between -2 and $+2$ standard deviations around the mean, and virtually all the data points will fall within 3 standard deviations above and below the mean (McClave and Benson 1985:83-87).

After all the z-scores are computed, then z_x and z_y are multiplied for each subject to produce the column, (z_x) (z_y), of (6). The sum of the (z_x) (z_y) column is then used to compute the correlation coefficient, r.

$$r = \frac{\Sigma z_x z_y}{N - 1} = \frac{7.61}{9} = .845$$

So there is a strong positive correlation of .845 of education with attitude scores in our sample. As noted above, $r = .845$ provides us with two pieces of information about the correlation. It is positive, and it has a certain magnitude. The coefficient of determination, r^2, however, although it obscures the directional information, gives us a more useful reading on the magnitude of prediction.

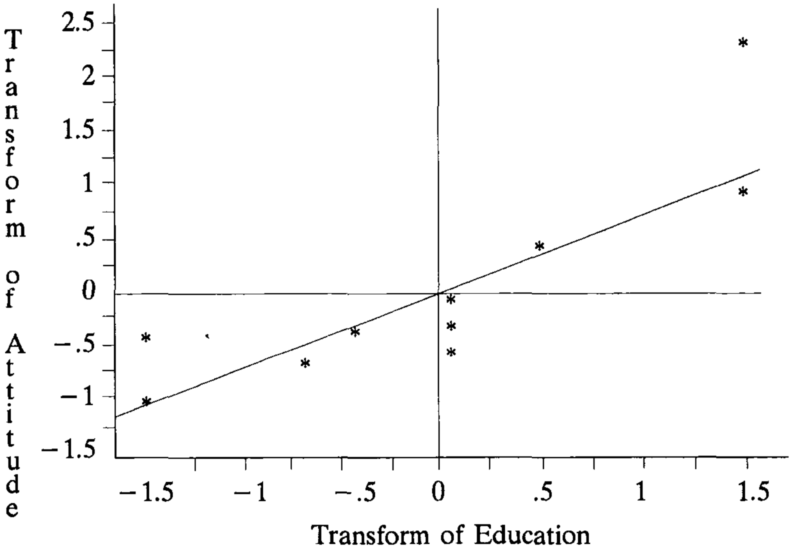
$$r^2 = (.845)^2 = .714$$

The usual reading of r^2 is that education correlated with attitude scores explains 71% of the variance. This statement still requires some interpretation. Recall that our best prediction of any individual attitude score using only what we know about attitudes is the mean. Variance is an index of the error generated by this best guess. When we added educational level to our predictive model and computed regression, our error was reduced. Compare r^2 to the computed proportional reduction of error (PRE) above.

As we can see, r^2, though computed differently, is actually a PRE measure. In the case of our data set, we were able to produce a relative reduction in our prediction error of 71% by measuring the association of educational level with attitude scores.

One further piece of information remains to be teased from r. Our formula for correlation actually compares the x and y scores in standard units (z-scores) which are directly related to the variances of the two measures. In essence, the Pearson correlation converts x and y to the same scale, and performs a regression analysis. The scattergram in (7) shows the attitude and education measurements plotted in standard units. Note that the regression line passes through the coordinate (0,0). This point represents the intersection of the two means. All of the data points are expressed in standard deviations from this point.

(7) Scattergram of attitude and education scores (transformed to *z*-scores) with regression line



At this point we recalculate the regression formula using *z*-scores instead of the normal scores. There is no need to calculate the Y intercept, *a*, since it will always be zero when the data points are expressed in *z*-scores. It only remains then to compute the slope, *b*.

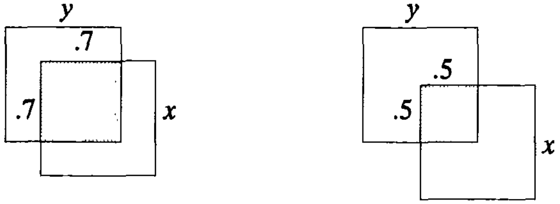
$$b = \frac{N(\Sigma(z_x)\,(z_y)) - (\Sigma z_x)\,(\Sigma z_y)}{N(\Sigma z_x^2) - (\Sigma z_x)^2} = \frac{10(7.61) - (0)(0)}{10(9) - (0)^2} = \frac{76.1}{90} = .845$$

Compare now our *z*-score slope, $b = .845$, with the correlation coefficient, $r = .845$. We find, then, that the correlation coefficient is equivalent to the slope of the regression line when our scores are expressed in standard units. Recall that the regression line represents our prediction of *y* scores from *x* scores. Since the Y intercept is always zero when *z*-scores are used, the slope becomes a direct measure of the prediction of *y*. For each *x* (expressed in *z*-scores), the predicted *y* will increase by .845 standard units.

Let us briefly review the relationship of the standard units (*z*-scores) of (7) to the variance of the variables. Z-scores are a transformation of the standard deviation which in turn is the square root of the variance. Standard units then, whether *z*-scores or standard deviations, when squared are equivalent to the variance. As we noted in the previous two paragraphs, *r* equals the slope when plotted in standard units. Stated another way, r equals a percentage of a standard unit of y accounted for by the regression of x on y. Since a squared standard unit is the equivalent of the variance, r^2 equals the percentage of the variance of y that is accounted for by the regression of x on y.

This relationship may be demonstrated visually as in (8). The sides of the squares are equivalent to one standard deviation of the variable for which they are named. The x and y squares in figures (8a) and (8b), then, graphically depict the variance in x and y. The shaded areas represent r^2, the amount of variance explained by the correlation. Recalling the discussion above, we can see this is because r is equal to the slope in the regression on a standard unit. Thus in (8a), representing an $r = .70$, x overlaps seven-tenths of one side of y which is analogous to a slope intersection of .70 plotted against standard units as in (7).

(8) Percent of variance accounted for


(a) 49% of variance accounted for $r = .70$

(b) 25% of variance accounted for $r = .50$

Superimposing the entire variance of x on the variance of y so that the two overlapping sides intersect at a point seven-tenths of the way down the sides of x and y in (8a) and five tenths down the sides of x and y in (8b) denotes the effect of squaring r illustrating the percentage of variance explained in each case. As we can see, 51% of the variance in y in (8a) as illustrated by the white space in the y variable's square is unaccounted for, and 75% of the variance in y is unaccounted for in (8b).

This leads us to the heart of what prediction means in correlational analysis. For an $r = .70$ as in (8a), accounting for 49% of the variance does not necessarily mean that we are making any correct predictions of y from x. It simply means that we have decreased the size of our error (the variance) from the entire y square which represents a prediction based only on the mean of y to just the unshaded white space in the y square, which represents the error that remains after we have made our prediction based on the correlation of x with y. By means of a correlation technique, we have drawn a tighter boundary on our error for an unknown y.

Before moving on to other matters, we should also note the relationship of correlation to variance in practical prediction. In statistical terms, all correlations of $r = .70$ are equivalent. If a particular y variable exhibits a very small variance, however, the narrowed error implied by a correlation coefficient of $r = .70$ may be so slight as to be of little practical use. If, on the other hand, a y variable exhibits a large variance, the practical implications of cutting the error approximately in half may be great.

Multiple correlations. Multiple correlations extend the notions inherent in simple pairwise correlations along with error adjustments (degrees of freedom) based on the number of factors under consideration. The multiple correlation coefficient, R, describes the magnitude and direction of prediction of a number of predictor variables against a criterion variable. As with simple pairwise correlation, the coefficient of multiple determination, R^2, represents the percentage of variance accounted for in the dependent, or criterion, variable.

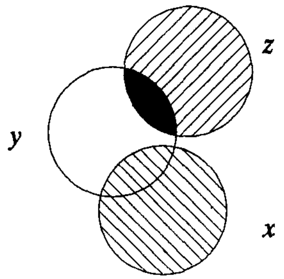
The primary tension in multiple regression is to find a set of independent variables that is optimally small and maximally predictive. The notion behind this is that given a potentially infinite number of predictors, we can fully account for the variance in whatever set of behaviors we choose to observe. On the other hand, the smaller the number of predictors, the more manageable the model becomes in practical settings.

Multiple regression analysis works by progressively introducing the predictor variables to increase the amount of variance accounted for in the criterion variable. As much as possible, the predictor variables should be independent of one another. In other words, the more highly any one predictor variable is correlated with another predictor variable, the less it will add to the predictive value of the set.

The Venn diagrams in (9), (10), and (11) illustrate the addition of z, a predictor variable of y, to x, another predictor variable.[20] The figure in (9) illustrates the situation where x and z are completely independent of one another. The correlation of x on z is nil ($r_{xz} = 0.00$). In this situation all of z's correlation on y contributes to the multiple correlation coefficient, R. Z's unique contribution to R is the area shaded black.

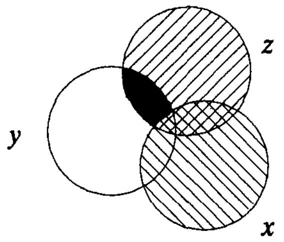
[20] In the figures of (9), (10), and (11), and the accompanying discussion, predictor variables x and z have identical zero order correlations with y, i.e., if we assume $r_{xy} = .4$, then $r_{zy} = .4$. This is only for convenience in illustration. The mechanics of accounting for variance would be the same if x and z exhibited unique r values. The Venn diagrams are not accurate in scale.

(9) Multiple correlation of x and z on y (nil correlation of x and z)



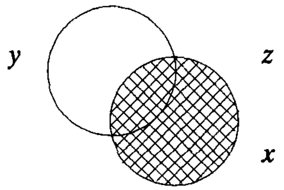
In (10), there is a moderate correlation of x on z. The figure illustrates the fact that z's contribution to the multiple correlation coefficient is partly unique (the area shaded black) and partly overlaps the contribution of x. In this instance z does account for part of the variance not covered by x, but its contribution to R is less than in (9). Even though the individual correlations of z and x to y are the same as in (9), since z and x are to some extent correlated with each other, the value of R is somewhat lower in (10).

(10) Multiple correlation of x and z on y (x and z moderately correlated)



In (11), x and z are perfectly correlated (r_{xz} = 1.00). Although z is correlated with y, it does not add any predictive value to x on y because it is not at all independent of x. In other words, z is utterly superfluous in (11). The multiple correlation coefficient, R_{xyz}, would have exactly the same value as R_{xy}.

(11) Multiple correlation of x and z on y (x and z perfectly correlated)



Statistical significance. We come now to the issue of statistical significance. What does it mean when we obtain an $r_{xy} = .40$ for a sample of thirty subjects, and upon examining a table of r values, we find that it exceeds the critical value of r at the 0.05 significance level in a nondirectional test?[21] Does it mean that we have proven the validity of our correlation? Does it mean that our obtained r is an adequate approximation of the true correlation, ρ (pronounced rho), of the larger population? For both questions the answer is no. We take them up in turn.

All significance tests in statistics are predicated on probabilities and sampling. If we measure some attribute of an entire population with a fully reliable instrument, then significance tests are irrelevant. Our measurements presumably reflect the true state of the attribute in the population. On the other hand, we very seldom can perform our measurements on an entire population. Rather, we are almost always in the position of drawing samples (hopefully, random)[22] in an attempt to infer the true state of the population. We are then able to evaluate our results (or draw inferences) using significance tests predicated on the laws of chance.

It is unfortunate, but true, that in using probabilities and samples, we can never prove anything. We can only formulate decision rules for evaluating probabilities and then estimate the probabilities associated with our obtained results. As it turns out, it is more difficult to accept our obtained results statistically than it is to reject an opposing hypothesis (called the null hypothesis, or H_0). Significance testing then is overtly designed to knock down a straw man, the null hypothesis. This allows us to accept our original hypothesis as a reasonable alternative by default.

[21] Most statistics textbooks will contain a table of critical values of r in an appendix. Almost all statistical software will, in addiiion 'io the r value, generaie the precise probability associated with the significance test.

[22] This is actually a rather strong hope. Statistical inference is founded upon the laws of chance, and without a random, or probability, sample, the assumptions guiding the inferences are not valid. To use a gambler's phrase, with a nonrandom sample "all bets are off."

Normally null hypotheses are hypotheses of no effect, and indeed, the most common null hypothesis for correlational studies is: H_0: $\rho_{xy} = 0.00$.[23] In prose, this hypothesis states that rho, the true correlation of x and y in the population, is zero. In other words, there is no correlation. When we feel certain x and y are correlated but have no idea whether the correlation is positive or negative, then it is usual to use a nondirectional alternate hypothesis: H_1: $\rho_{xy} \neq 0.00$.

If, prior to testing, we have reason to believe that the correlation is positive, then we are entitled to use the alternate hypothesis: H_1: $\rho_{xy} > 0.00$, or, alternately, if we feel the relationship is negative: H_1: $\rho_{xy} < 0.00$. The latter two hypotheses are directional and possess the advantage of requiring lower critical values of r to achieve significance than the nondirectional hypothesis.

Returning to our previous example of the correlation of educational level on attitude in the section "the meaning of correlation," if we make an *a priori* assumption of a positive correlation, we can test the following hypotheses:

$$H_0: \rho_{xy} = 0.00$$
$$H_1: \rho_{xy} > 0.00.$$

We also predetermine our decision rule: reject the null hypothesis if the obtained r achieves a significance level of 0.05 or better. This rule requires us to reject the null hypothesis only when the measured effect is large enough that the odds are only one in twenty or less that the obtained r is a random deviation from a zero correlation. In our case, with a sample of 10 and an obtained $r = .845$, we far exceed the critical value of r for a 0.05 significance level in a one-tailed test.[24] We are justified then in rejecting the null hypothesis and accepting our alternate hypothesis.

The logic behind the critical value tables and the probabilities reported in statistical software is that for a given sample size, a hypothetical normal distribution is constructed around $r = 0.00$ (the null hypothesis). For the normal distribution, we know the probabilities for scores which fall beyond any given point. Based on this distribution, a confidence interval is

[23] This null hypothesis is so common, in fact, that it is sometimes left unstated in research reports. This does not alter the fact, however, that significance testing is designed for the rejection of a null hypothesis and not for the validation of an alternate hypothesis. With few exceptions, critical value tables for r are overtly designed with the null hypothesis in mind. That is, the critical value represents the upper edge of the confidence interval around $\rho_{xy} = 0.00$.

[24] Butler (1985:180) is the source of the critical value table used here. It is oriented toward smaller samples and contains significance levels for both directional and nondirectional hypotheses.

constructed which encompasses 95% of the expected r scores for samples of 10 if the true correlation is zero. In a directional test, the 95% confidence interval for a sample of 10 encompasses all r values below $r = .549$. All r values of $r = .549$ and above lie outside the confidence interval and are considered to be significantly different from $r = 0.00$ at the 0.05 level of significance.

In answer to the first of the questions with which we opened this section, we really have not proven the validity of our correlation. What we have done is show that, given the magnitude of our correlation, it is likely that there is some true correlation between x and y in the population. The significance level has not given us an estimate of what ρ, the true correlation in the population, really is. Rather, it has given us an indication of what ρ is not. That is, we are reasonably certain that the true correlation is not $r = 0.00$.

In regard to our second question from the beginning of this section, about whether our obtained r is an adequate approximation of the true correlation of the population, it is, and it isn't. It is a point estimate of ρ, and as such, we may expect it to vary considerably from sample to sample. It is possible, and often desirable, to improve the point estimate by calculating a confidence interval around an obtained r.[25]

The confidence interval around our obtained r gives us a range within which we know the probabilities of finding the true correlation of the population, ρ. For instance, in our example correlating educational level with attitude scores ($r_{xy} = .845$; n = 10), we can be about 95% certain that ρ lies in the interval from $r = .46$ to $r = .96$. With the same correlation coefficient, and a sample size of 50, we would be able to state with 95% certainty that ρ lies in the interval from $r = .74$ to $r = .91$. A principle emerges from this. As the sample is enlarged, the confidence interval around our obtained r shrinks, and the obtained r is more probably a close approximation to ρ, the true correlation in the population.

This interval estimate of ρ takes the magnitude of the correlation and the sample size into account to also give us an enriched interpretation of our correlation coefficient. Knowing the size of a confidence interval around an obtained r is analogous to knowing the standard deviation about the mean. It provides us a notion of how firmly established a particular correlation coefficient is given the size of the sample from which

[25] Woods, Fletcher, and Hughes (1986:163–65) describe a method for calculating the confidence interval around a correlation coefficient. While it is only valid for samples of 50 or greater, it does provide a fairly accurate approximation in smaller samples. In fact, Bernard (1988:406) appears to have used it to calculate a table of confidence intervals around r for samples as small as 30.

it is drawn. The smaller the range under the confidence interval, the better the estimate our obtained r is of ρ.

Type I and Type II errors and their practical consequences. In all statistical significance testing there are two potential errors that are always present. Type I, or α, errors occur whenever we reject a null hypothesis that is true. This is the case of "proving" our original idea (H_1) when it is really wrong. Type II, or β, errors occur when we accept a null hypothesis that is false. When we commit a β error, our original theory is right, but the results don't exceed the critical value for significance. In this case, our idea was right, but we couldn't "prove" it. Example (12) illustrates α and β errors in the context of a decision table.

(12) Type I (α) and Type II (β) errors

		Decision	
		Reject H_0	Accept H_0
Null hypothesis	True	α error	Correct
	False	Correct	β error

The probability of committing an α error is directly stated in the significance level. If we formulate a decision rule wherein we reject the null hypothesis ($\rho_{xy} = 0.00$) at the 0.05 significance level, then the probability of falsely rejecting a true null hypothesis when our r value exceeds the critical value is about five times out of 100.

To illustrate, suppose that we have drawn a large number of random samples, perhaps 500–1000, each containing 30 subjects, from a very large population in which the true correlation between two variables actually is $\rho_{xy} = 0.00$. If we use a nondirectional hypothesis and a 0.05 significance level in our decision rule, then we can expect about one in twenty of our sample groups to exhibit an $r_{xy} \geq 0.361$, our critical value, purely by chance. In other words, we would conclude in about one out of twenty cases that the relationship is significant when, in fact, it is not.

Our first reaction might be to assume that the solution would be to simply increase the confidence interval as much as possible, but that would increase our probability of making a β error that is, we would fail to reject a greater number of false null hypotheses. If we lower the significance level from 0.05 to 0.01, then we do decrease probability of incorrectly rejecting a true null hypothesis (an α error), from about one time out of twenty to about one time out of 100. At the same time, however, it becomes more difficult to reject a null hypothesis that actually is false, and conversely, becomes more difficult to accept our alternate hypothesis (which we have been ardently hoping to accept).

The bad news is that we cannot eliminate Type I and Type II errors altogether, and since they are linked, any tweaking of the α level also affects β. The good news is that there are other factors affecting β besides the α level.

There are at least four factors which affect β: the level of significance (as noted), the magnitude of the effect, variability of the population, and sample size.[26] The latter three factors are usually discussed in statistics texts under the heading of "power functions" or "the power of tests."

The power of a test (defined as 1 − β) refers to its ability to correctly reject a false null hypothesis (the lower left box of the decision table in (12)). This is the outcome we are normally aiming for in statistical inference. If our β level is .35, then the power of the test would be .65; if the β level is .2, then the power is .8, etc.

By manipulating those variables that increase a test's power, we can decrease the probability of making a β error. Either choosing variables which show large effects when present or using scales with better discrimination will help decrease the chance of β errors. Using SLOPE to evaluate bilingual abilities rather than recorded text tests is one example of controlling the magnitude of the effect through a more discriminant instrument to gain greater power.[27]

All other factors being held constant, the smaller the standard deviation of the population, the more powerful the test. One way to control for variability in the population is to draw our sample from a homogeneous population. This is essentially the strategy Walker (1987:71) employs in choosing single communities as his unit of analysis rather than entire language groups. The advantage gained by screening out bilinguals in dialect intelligibility testing results in increased power as we restrict variability in second language skills. Unreliable dependent variables also introduce unnecessary error and thereby increase the standard deviation.

Likewise, increasing the sample size increases the likelihood of avoiding a β error. It should be pointed out, though, that with large enough samples it is possible to obtain results that are statistically significant, but insignificant for all practical purposes. For example, with a sample of 200, an

[26] Shavelson (1981:378–81) contains a good discussion of these factors.

[27] James, Masland, and Rand's (1989) use of SLOPE as the baseline for evaluating recorded text tests and sentence tests as methods for assessing bilingual proficiencies is based on its greater discrimination. That is, they used the more powerful test to assess the efficiency of the less powerful ones.

r = .12 is significant at the 0.05 level, but it explains just 1.4% of the variance.

Since it is impossible to eliminate the chance of committing Type I and Type II errors, the recommended balance is to design a study to (1) use an instrument that has good discrimination (controlling the magnitude of the effect) and reliable variables, (2) set α at a conservative level[28] to minimize Type I error, and (3) get a sample which is large enough to be sensitive to theoretically and practically significant differences.

The practical consequence of α and β errors and power is felt when we use statistical inference in decision making. The decision rules we establish for evaluating our hypotheses are directly linked to the program implementation decisions we make. Our decision tables always imply certain costs and benefits attached to the possible errors. Because of this it is vital that we do at least an informal cost benefit analysis of our decision tables and rules as we decide how to decide.

The basic question is: given the costs, what are acceptable odds for being wrong in drawing inferences about Bible translation and literacy programs for particular peoples. An example of a hypothetical decision table based on intelligibility testing is shown in (13).[29]

(13) Hypothetical decision table for intelligibility testing

		You decide	
		They are different	They are the same
The speakers from test point Y are part of the same linguistic population as the speakers from reference point X (H_0)	True	unnecessary translation (α error)	Correct
	False	Correct	group with needs missed (β error)

[28] In quantitative scientific research the 0.05 and 0.01 significance levels have become conventionalized, but for other purposes an individual may decide it is worth the risk to accept a different significance level depending on the consequences of the potential errors.

[29] Significance testing with intelligibility tests normally involves the comparison of means, rather than correlation. While it would have been possible to devise a decision table based on a correlational hypothesis, the intelligibility based hypothesis was chosen because of its broad familiarity among surveyors and administrators.

Even if the speakers from test point Y can understand speakers from reference point X well enough to be considered practically the same population, we might by chance draw a sample that indicates they are different. Typically we use a 0.05 significance level in this type of decision rule accepting odds of one time in twenty saying two groups are different when they are, in fact, the same.

The costs of an α error in this situation would involve an unnecessary investment of personnel, money, and time to produce a New Testament translation. There might be some benefits to the people from Y, however, especially if we interpret Landin (1989) to imply that the presence of sympathetic missionaries promotes the use and valuation of the vernacular, but the New Testament produced would not have been strictly necessary.

If we decide, however, that a one in twenty chance of doing an unnecessary translation is too great, then we can tighten up the odds by setting the significance level to 0.01 (or one in 100 odds of making an α error). The problem with this approach, however, is that by lowering the odds of an α error, we automatically increase the odds of concluding that speakers of X and Y are members of the same population with respect to Bible translation and literacy needs when they are not (a β error).

The end result of a β error in this decision table is that a group who needs a separate translation would be missed. Financial and personnel costs of a β error would be negligible, but the spiritual impact and costs to the individuals of a group which needs its own New Testament but does not receive it, while intangible, should certainly be considered.

Conclusion

A number of the issues raised in this paper are directly relevant to the interpretation of the results of research using correlation as an analytical tool. The role of correlation in prediction has been a major focus. I have attempted to show that prediction, or accounting for variance, in correlation does not mean guessing more answers right; it means guessing more answers less wrong.

In addition, the virtues of the more humble, but more easily interpretable, coefficient of determination, r^2, have been showcased. Whereas the more commonly reported correlation coefficient, r, indicates both the magnitude and direction of a relationship, r^2 actually reports directly the proportion by which our error (the variance) has been reduced. It will be a great comfort in my old age if just half of those who read this far, ever after immediately think of r^2 and the percentage of variance explained whenever they see r.

In the same vein, I have spun a semi–cautionary tale about the use of statistical inference, the role of α and β errors, decision tables, and other quasi–mythical statistical beasts (possibly arising from an over–active calculator) which we really cannot escape. Since we must live with them, it is perhaps comforting to know they are at least somewhat tameable. Like a pet wolf or bear cub, however, they retain their teeth and a taste for the occasional nip at the unwary.

Anytime we get stuck messing about with a pile of numbers (especially somebody else's pile of numbers), a bit of caution is in order. We must learn to look beyond a statistical result and understand the relation of the numbers to the data. At some point, we must peel away the layered abstractions inherent in the numbers and the operations performed on them and contemplate the data. Quite often that is the hardest work of all, but we must come to understand what it is that the numbers really represent. Statistical significance is usually meaningless if the data are the result of poor sampling techniques, questionable methods, or poor representations of the research question.

From the other side of the coin, however, it is also not a good idea to throw up our hands in defeat and scurry away looking for the prose summary at the first sign of a statistical table. More than one researcher has gotten caught short by making statements about his findings which the numbers did not support. In part, the aim of this article has been to relieve one small part of our collective statistical anxiety by attempting to trace the path of logic from the rawest numerical data to the correlation coefficient and the coefficient of determination, hopefully stripping away some of the mystique of the formulas and replacing it with an understanding of the role and nature of correlation in prediction.

Perhaps a respect for both the power and the pitfalls of statistical reasoning (for statistics can be both illuminating and obfuscating) will, at times, inspire us to caution as we push forth our own theories and the theories of others which may lead to program decisions affecting our own personnel, our financial resources, and especially the lives of the minority peoples for whom the numbers toll.

[blank]

State of the Art: Dialect Survey Fifteen Years Later

Eugene H. Casad

The field formerly designated 'dialect survey' has changed considerably in the fifteen years since *Dialect Intelligibilty Testing* (Casad 1974) was published. Work in the Philippines, Sabah, Pakistan and other areas has both confirmed much of what I wrote in that monograph and has enriched our understanding of various aspects of the language assessment problem. Many of us now see several areas as being interrelated to one degree or another: linguistic similarity, dialect intelligibility (i.e., comprehension), bilingualism and language use, language attitudes, dialect extendability, and the acceptability of vernacular literature (Blair, 1990).[1]

We have a better understanding of the characteristics of the relevant indicators of language assessment and a much better conception of the relevant variables that underlie them. In particular, the sets of variables that underlie linguistic similarity are largely distinct from those that underlie intelligibility. In turn these two sets of variables only partly overlap with the variables that underly bilingualism. Likewise, those variables that determine

[1] I would like to thank a number of people who have helped make this paper possible. Special thanks are due to Joseph E. Grimes and Barbara F. Grimes who have helped update me on recent developments in the field and have made it possible for me to teach a section of their survey course. Besides that, they also read and made comments on this paper. My thanks are also warmly extended to Bruce Hollenbach, David Persons, Calvin Rensch, Ronald Stoltzfus, and John Wimbish for their input. Finally, my very deep thanks to the Indonesia branch for inviting me to hold a workshop at SIL-Umpatti, Ambon, Maluku in March of 1988. It was a stimulating and educational time for me and made this paper a better one.

attitudes toward language only partly overlap with those that relate to bilingualism. In short, the complexity of the problem is such that we cannot simplemindedly substitute a measure of one area of the problem as an adequate and sufficient indicator of the state of affairs in another domain of investigation. In other words, linguistic similarity is not an adequate single predictor of intelligibility (see Kroeger 1986; J. Grimes 1988a), nor is an intelligibility test an adequate single predictor of bilingual proficiency and so on.

We have seen an increased sophistication in our statistical treatment of survey data, moving all the way from making simple point estimates of sample mean scores to testing for correlations by way of linear and multiple regression analyses (cf. Simons 1977; Kroeger 1986; J. Grimes 1988a, 1988b; Walker 1987). To help us cope with this increased complexity, there has also been a pair of technological revolutions that have given us highly useful tools for collecting and analyzing survey data, i.e., the cassette recorder and the personal computer with a host of attendant software programs.

On the negative side, too many innovations in survey techniques have been made without any concern for either reliability, i.e., consistency in measurement, or validity, i.e., the test really tells you what you want to know. In this paper I discuss my concerns with a few of these innovations.

Increased complexity in the assessment task

The most noticeable change involves the increased complexity in the survey task as we now perceive it. Many of us now see the need for viewing several fields of investigation as being interrelated to one degree or another and, therefore, requiring a unifying treatment within an intricate, but integrated framework. These fields include linguistic similarity, dialect intelligibility, bilingualism and language use (both between genetically related dialects and between unrelated languages), language attitudes, dialect extendability and the acceptability of vernacular literature. Although I mentioned all of these areas in *Dialect Intelligibility Testing* (1974), I treated some of them in only cursory terms. By now we have in hand a few doctoral dissertations (Simons 1979; Quakenbush 1986; Walker 1987, to mention a few) and a number of interesting MA theses (Stahl 1988). All these works in one way or another have helped to fill out a more complex picture. In addition, there have been a number of very helpful articles that report on very well designed and carefully implemented research (B. Grimes 1985a, 1985b; 1986a; 1988; J. Grimes 1988a, 1988b; Kroeger 1986; Hurlbut and Pekkanen 1982; James, Masland and

Rand 1989; Radloff (1991) as well as several monographs (Blair 1990; Schooling 1990).

Out of all this, we are increasingly viewing survey as a set of stages of successive approximations to workable solutions. The figure in (1) devised by Barbara D. Grimes is presented as an effective means of displaying both the individual areas of assessment activities and the temporal and logical relations between them.

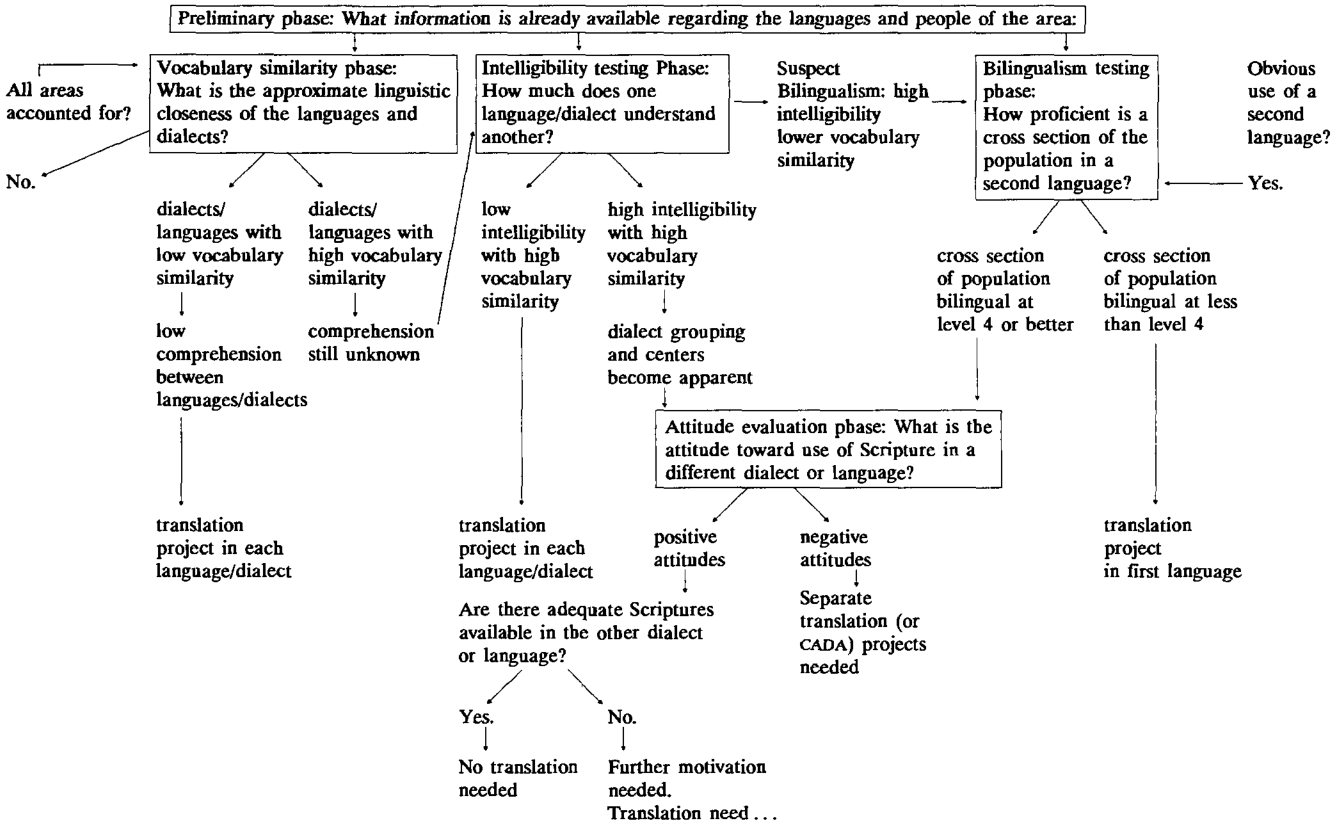
The preliminary phase of a survey consists of choosing an area for study, and doing background library research to determine what relevant information is already available regarding both the languages and the culture groups in the area. This phase also includes establishing public relations contacts among educators and local administrators, as well as the initial selection of the particular dialects that will be included in the study (Casad 1974:4–8; Sanders 1977:22–24). It is crucial to ensure comprehensive coverage of an area. Thus, during the course of a survey, supplementary information may turn up that will entail the collection of additional data and the specification of additional dialects.

The vocabulary similarity phase seeks to determine the approximate degree of linguistic proximity among the several dialects of a language. In those cases where the degree of vocabulary similarity is low (usually 60% or less), we can predict reasonably well that comprehension between those dialects will be low (J. Grimes 1988a, 1988b) and that, therefore, a separate language program will be needed for each of these distinct languages or dialects. On the other hand, for those dialects demonstrating medium to high vocabulary similarity levels (60% to 95%), the typical case is that we still can not predict what the degree of comprehension between those dialects will be (J. Grimes 1988a; Kroeger 1986). Thus we will need to test inter-dialect comprehension for these dialects.

The intelligibility testing phase tries to determine how well speakers of one language or dialect understand speakers of another one. If the results of the testing show that comprehension is low between a pair of dialects in spite of high vocabulary similarity, then it is likely that separate language projects will need to be initiated for each dialect or language. If the results show that comprehension is high in conjunction with high lexical similarity, then it is likely that dialect groupings with their respective centers will be discernible also. At this point the relevant dialects can be treated in terms of the attitude evaluation phase.

On the other hand, if the results show that a high degree of comprehension exists among speakers of distinct dialects that are marginal or low in vocabulary similarity (60% to 85%), then we can suspect that a marked degree of bilingualism exists between these groups and that it will be

(1) Survey: seeing the whole picture. (Barbara D. Grimes)


Preliminary phase: What information is already available regarding the languages and people of the area:
All areas accounted for?
No.
Vocabulary similarity phase: What is the approximate linguistic closeness of the languages and dialects?
Intelligibility testing Phase: How much does one language/dialect understand another?
Suspect Bilingualism: high intelligibility lower vocabulary similarity
Bilingualism testing phase: How proficient is a cross section of the population in a second language?
Obvious use of a second language?
Yes.
dialects/ languages with low vocabulary similarity
dialects/ languages with high vocabulary similarity
low intelligibility with high vocabulary similarity
high intelligibility with high vocabulary similarity
cross section of population bilingual at level 4 or better
cross section of population bilingual at less than level 4
low comprehension between languages/dialects
comprehension still unknown
dialect grouping and centers become apparent
Attitude evaluation phase: What is the attitude toward use of Scripture in a different dialect or language?
translation project in each language/dialect
translation project in each language/dialect
positive attitudes
negative attitudes
translation project in first language
Are there adequate Scriptures available in the other dialect or language?
Separate translation (or CADA) projects needed
Yes.
No.
No translation needed
Further motivation needed.
Translation need . . .

necessary to determine both the nature and degree of that bilingual proficiency.

The bilingualism testing phase, then, begins once we have pinpointed positive cases. At this point, it is necessary to determine how proficient a cross-section of that community is in the use of a particular second language. If the results of testing this stratified sample show that the subjects test out at less than an FSI level of 4 (as determined by SLOPE testing), then, other factors being equal, it will be necessary to establish a language program in the first language of that group. But should the results show that the sample of subjects are indeed bilingual at an FSI level of 4, then we need to consider that community in terms of the attitude evaluation phase.

Now the question becomes one of what kind of attitude the speakers have toward literature printed in the second language. If the peoples' attitudes are positive, then the question needs to be asked whether or not adequate materials are available in that second language. If the answer to this question is 'yes', we can justifiably conclude that no separate language program is needed for that group. Yet, if no adequate materials written in that second language are available, then further evaluation must be made, with the eventual possibility of initiating a language program in that group's first language and selling them on the idea of using it for literary and educational purposes.

Another possibility also exists. Should the population that appears to be adequately bilingual in a second language turn out to demonstrate negative attitudes toward the use of literacy materials written in that second language, then separate language projects or CADA projects will need to be implemented.

Top of the head figures from Barbara F. Grimes are that possibly fifty percent of language assessment needs can be handled at stage I on the basis of lexical similarity, whereas another thirty percent can be handled at stage II by testing for intelligibility. Stage III, bilingualism testing, may be needed for another fifteen percent of the cases, whereas attitude testing may prove crucial in as much as five percent of the cases (B. Grimes 1986a). This does not mean that stages II and III are less important than stage I, only that stage I, by the nature of the case, is logically prior to the others. These figures are impressionistic, but we can begin to firm them up by checking survey results from Mexico, the Philippines, and Sabah, at least. Finally, note that the sampling problem changes from stage to stage and, in general, the techniques employed become more difficult, time consuming, and subtle in their application and interpretation.

Variables underlying the assessment problem

A clear impression from a consideration of the literature that has come out over the past fifteen years is that we now have a much increased awareness of the characteristics of the relevant indicators of language assessment and the variables underlying them. To begin, linguistic similarity includes structured likenesses and differences at the syntactic, morphological, lexical, and phonological levels. SIL language surveys have typically focused only on lexical similarity (cf. Sanders 1977), but we really need a broader measure of things. For example, syntactic patterns can be borrowed, and they definitely do change.

At the phonological level, linguistic similarity is partly due to the operation of regular sound change, including splits, mergers, coalescences, and dissimilations. In addition, new segments frequently enter sound systems through borrowings, in many cases altering the set of phonemic contrasts in the indigenous sound system. Lexical similarity itself is partly due to historical relatedness; in part it also reflects borrowings from various sources. Older borrowings may be so well adapted to the native system as to be almost unrecognizable. Later borrowings are usually more easily recognized. In certain cases, individual lexical items can result from loan translations, or they may reflect marking reversals, as in the case of animal names in some Amerindian languages.

Morphological similarity can be affected by reanalyses of two sorts. In one case, a pair of distinct morphemes fuse to become a single one. In other cases, speakers impose a polymorphemic analysis on something that historically was either a single morpheme or had a different morphemic analysis. Morphemic processes such as the generalization of a particular morphemic shape throughout an entire paradigm affect linguistic similarity. Finally, morphophonemic processes of various kinds lead to a greater or lesser degree of lexical similarity among dialects.

Similar things can be said with respect to syntax. Syntactic patterns common to a pair of dialects may either reflect a common genetic base or they may be borrowed from one language or another. Commonly, sentence connectors and conjunctions are borrowed. The entire comparative construction has been borrowed from Spanish by a number of Mexican Indian languages. In addition, calques (loan translations) to some extent affect syntactic patterns. For example, Spanish *que Dios te lo pague* 'Thank you!' has appeared in Cora of Mexico as *Che'e Dios timuaatanajchite'en* and in Dos de Mayo Quechua as *Dyosolpaki* (Weber and Mann, 1979). In both cases, the indigenous phrase literally means 'May God repay you,' although the analyzability of *Dyosolpaki* into separate meaningful units has probably been lost.

Finally, numerous discourse functions and relations can be involved in linguistic change. Again, these changes may reflect either historical changes between genetically related languages or they may arise from borrowing from distinct languages. Sentence connectors and expletives are commonly borrowed this way, as are higher level discourse units.

I close this section with another pair of observations. First, linguistic similarity between a pair of dialects is typically symmetric: i.e., dialect A is as different from B as B is from A. Second, linguistic similarity is partially grounded in societal phenomena, but usually this involves considerable time lag.

The variables that determine intelligibility are largely distinct from those that determine linguistic similarity. In a nutshell, intelligibility is based partly on linguistic similarity between dialects, not just on lexical similarity. In addition, intelligibility is partly based on interdialectal learning, which I take to be a subclass of bilingualism (see Kamp, this volume) and is distinct from inherent intelligibility. All the societal attitudes and situations that underlie bilingualism also feed into intelligibility, but they probably feed into inherent intelligibility in one way, whereas they feed into intelligibility due to interdialectal learning in another way. We may not always be able to factor out these distinct effects, but we nonetheless need to continue trying to tease them out of the data. The time lag between societal contact and a discernible effect on intelligibility is quite distinct from the time lag associated with the impact of societal phenomena on linguistic similarity. Intelligibility is also typically asymmetric, i.e., speakers of dialect A understand speakers of dialect B better than speakers of dialect B understand those of A. (Grimes 1974, Casad 1974.) (Note: the myth of mutual intelligibility dies hard!) In addition, intelligibility is normally distributed within a population, i.e., it is not restricted to age groups, sex, or educational level (see Kroeger 1986). Nor is intelligibility restricted to particular domains of language use.

As for bilingualism, there are significant ways in which it differs from intelligibility. To begin, bilingual proficiency is not necessarily tied to the genetic status of the mother-tongue languages in question, whereas intelligibility is (see B. Grimes 1986a). Bilingualism derives more directly from societal phenomena, including the kinds of interactions between the social groups in contact and the attitudes the mother-tongue speakers hold with respect to various factors. The purposes of such interactions also are relevant. Bilingual proficiency, therefore, is often restricted by age, sex, educational level, and even the location of primary residence. It is often also restricted to particular cultural domains of language use. Finally, bilingual proficiency, like intelligibility, is often asymmetric. Thus the Kalagan of Northern Mindanao learn to speak B'laan, but speakers of

B'laan do not learn to speak Kalagan (Francis B. Dawson, personal communication).

Moving on to a consideration of attitudes, Roland Walker divides them into forces impinging upon the culture from the outside and forces arising from within the culture itself. External forces include language contact, government policies, immigration, economic forces, and intermarriage. Internal forces include ethnic identity, cultural resiliency, traditional values, vernacular literature, language use, and various other factors (Walker, 1987). Considerations such as the previous history of contact between the local culture and the outside are also highly relevant.

Finally, the question of the acceptability of vernacular literature is an additional aspect of the ball game. As I wrote in 1974, simply because a person can understand the speech of another dialect or language does not entail that he will accept materials written in that form of speech (Casad 1974:59, 70). Thus we need to determine more precisely what the variables are that relate to the acceptance of vernacular literature. Walker (1987) is the primary study that points us in this direction. Some of the factors that Walker mentions include the value of education to the community, community attitudes toward both the written and spoken languages, the differences between national language and vernacular language orthographies, the perceived ease of reading the vernacular, the amount of available vernacular literature, and the community leaders' involvement with all aspects of the literacy program. In passing, note that almost all of these variables presuppose the prior existence of literature. This is a strong suggestion that the question of the acceptability of vernacular literature must be settled at a substantially later stage than the specification of the minimal set of centers for the initiation of language programs.

There are actually numerous variables in all of these sociolinguistic categories that I have not mentioned here, but what I hope to have done is to have shown clearly two things: (1) There is some overlap among the sets of variables that underlie linguistic similarity, intelligibility, bilingual proficiency, language attitudes, and the acceptability of vernacular literature; and (2) The sets of variables that relate to these phenomena are nonetheless distinct. Thus, one must use distinct measuring devices to properly evaluate them. The interrelationships are enough to allow an investigator to discover some correlations between certain indicators of select sociolinguistic phenomena. The differences are great enough, however, to ensure that no single device used for measuring one phenomena will allow one to predict with any acceptable degree of success to a distinct one. In other words, linguistic similarity is not an adequate single predictor of intelligibility (sèe Kroeger 1986; J. Grimes 1988a), nor is an intelligibility test an adequate single predictor of bilingual proficiency and so on (B. Grimes 1986a).

Increased sophistication in treatment of survey data

In the past several years we have substantially upgraded our sophistication in the statistical treatment of survey data. In doing so, we have moved all the way from making simple point estimates of sample mean scores to testing for correlations by way of linear and multiple regression (cf. Simons 1979; Kroeger 1986; Grimes 1988a; Walker 1987; Radloff 1991). An optimization model has been profitably adapted for our use with survey data as both a heuristic device and a descriptive one (J. Grimes 1974, 1985; Casad 1974; Simons 1979). Concern for measures of variation, in contrast to the complete reliance on sample mean scores, is another positive development. Finally, Grimes has recently shown an insightful application of box plot analyses to survey data, allowing us to sort out several distinct situations associated with variation in sample scores (J. Grimes 1988b).

We are also beginning to see an appropriate concern with the need to rely on multiple indicators of given sociolinguistic phenomena (cf. Kamp, this volume; Quakenbush, this volume; Radloff 1991). This has been motivated by an increasing awareness of the complexity of the language assessment task and the need to validate our test instruments.

Too many innovations in survey techniques have been made without any concern for either reliability, i.e., consistency in measurement, or validity, i.e., the test really tells you what you want to know. The concern that the Asia area survey teams show for these considerations by the first-class methodology lying behind their development of the Sentence Repetition Test (SRT) is both a personal encouragement and an example for the entire field. Furthermore, too much of the effort put into innovating test instruments has been directed to altering the form of the instrument in order to save time. This approach to things can be very counterproductive, as Wilson pointed out in his handbook to scientific research:

> Mere experimental convenience should not be allowed to outweigh the more basic considerations. The easy experiment may not answer the right questions. (Wilson 1952:40)

A classic innovation of the sort I am bothered by is seen in Simons' use of group testing in lieu of individual testing through earphones. He comments that he did not feel that it was right to exclude bystanders from the testing (1979:28). This is the main rationale he gives for making a fundamental change in a test design that has been successfully applied in Mexico, the Philippines, Ghana, Sabah, India, Indonesia, and Pakistan. My own feeling is that had Simons done his public relations more thoroughly he would have found no major obstacles to the testing. My comment here does not mean that doing public relations is easy. It is not. It can be very stressful and time consuming, as well as requiring finesse and even the proper social matching of the investigating team (cf. Radloff 1991; Stahl 1988). The other commonly cited rationale for Simons' modifications is the ease and quickness of test construction and administration (Simons 1979:10–11, 29). Wilson's comment above is very pertinent in this regard.

There are several problems in the group testing method. For one, it often turns out that a single person in the group is a particularly strong personality and succeeds in imposing his own will on the whole. How his actual understanding relates to the true population value is anybody's guess, but he is often a person with a much broader background than that of the other members of the group, which is what affords him the additional leverage. On the other hand, one or two members of the group may understand the dialect being tested much better than the other group members. In this case the score would be skewed higher with respect to the population value. In a third case, the members of the group could pool their knowledge to give a score that would also be skewed high. A general weakness of all of this is that the investigator is often going to be left with a score based on a sample size of one, rather than the sample size of ten that he would have had if he had tested individually. This is a crucial point, especially in view of the information that can be gleaned from the variability that occurs in the responses of a sample of ten (cf. J. Grimes 1988b). It should be obvious that a sample of one affords no variablity to measure. I am not satisfied that Simons' controls for these are sufficient (cf. Simons 1979:25).[2]

Simons also changed the method of scoring the intelligibility test from a percentage scale to an interval scale of "1" to "4". He did this claiming that "the results of methods which yield percentage scores are already too precise for the level of statistical significance that can be attached to them" (1979:28). As it stands, the statement is misleading. On that basis, however, he switches to an interval scale such that 3 = full intelligibility, 2 = partial intelligibility, 1 = sporadic recognition and 0 = no understanding (1979:10). Although interval scales are widely used, Simons' switch at this juncture automatically means that we cannot compare statistically his results with those deriving from more standard approaches (cf. Kroeger 1986:319).

A number of these nonvalidated versions of intelligibility tests have come about from good intentions—people were concerned about the presence of error in our survey results. So they tried to design an error-free instrument. All such efforts are bound to fail, however. The best we can

[2] In fairness to Simons, I need to point out that he himself now sees "too many problems for group testing to be useable" (Gary Simons, personal communication).

do is control for whatever disturbing factors are in the situation and then try to estimate how much our error is. As Joe Grimes put it to me, the role of methodology is as follows:

1. Good methodology makes you aware of experimental error; the most solid fact underlying the theory of experimental design is that all measurement contains some degree of error (Wilson 1952:34). Furthermore, this error creeps in from various sources, not just from the test instrument itself. Experimenter bias and subject bias are also present and may significantly affect the results. Neither word lists, questionnaire data, comprehension tests nor observations are immune to such disturbing influences (B. Grimes 1988). The pervasiveness and subtlety of it all is why so many controls had to be specified for the design of the recorded text test method used for testing dialect intelligibility. A recent claim that such studies are "uncontrolled" is misleading and untenable.
2. It helps you to estimate the size of the error. This is based on the theory of sampling. Particular kinds of statistics can be appropriately employed to estimate the range of possible values of a given characteristic if one's sample has been appropriately selected. Even if you cannot meet the assumptions needed for the appropriate use of standard parametric statistics, nonparametric methods can help us immensely, as J. Grimes' boxplot analyses suggest.
3. It tells you how much a given error hurts you. In certain cases the error may not have much of any bearing on your use of the results. More likely, however, it will have sufficient enough effect that you will need to be concerned about it, and in certain cases, the error may have devastating consequences. You need to have the entire design well conceptualized in order to distinguish and control for these possible eventualities.

Improved technological tools

With all the mention that I have made of complexity in this paper, I need to remind the reader that there is no need for throwing one's hands up in despair. Increasingly, a powerful array of evaluative instruments is being put at our disposal and powerful computational tools are being developed to help us catalogue, describe, and analyze our survey data. Limited space only allows me to barely mention some of these tools here.

Commercial software programs such as Minitab and Number Cruncher are available that allow the user to perform a full range of statistical calculations from the keyboard. These programs not only do the standard calculations for you, but they will also do scattergrams, box plots and locate regression lines within the scattergram. Joe Grimes has, for quite some time now, been developing programs to handle more specifically survey-oriented data. One such program is INTELL, which will figure out arithmetic mean scores, give the range of a sample of scores, calculate standard deviations and confidence intervals, as well as perform t-tests for significant differences between a pair of sample mean scores.

Other programs have been developed to help us do the dog-work of various kinds of linguistic comparisons. Frantz' (1970) COMPASS (for Comparativist Assistance) is a good example of this. It sorts word lists that have already been marked for whether pairs of lexical items are possible cognates or not, calculates and prints out regular correspondences, notes their frequency, and goes back to examine each pair of words in the list segment by segment. On the basis of the frequency of correspondence, this program then calculates a value for strength of correspondence between each pair of words. Above a certain value, pairs are considered probably cognate (i.e., historically related). Wimbish has also been working on a family of programs that he calls WORDSURV. This set of programs includes a modified version of Frantz' COMPASS as well as programs for calculating phonostatistical values. These are not all the tools at our disposal. For example, I have passed over the entire range of literature related to CADA, the Computer-Assisted Dialect Adaptation view of things. At the very least, I hope this brief listing will be a small encouragement to survey technicians.

Conclusion

In general, then, the assessment picture is very much more complex than what we had thought at first. None of it is easy and all of it is important as well as time consuming. In some cases we have developed our own measuring instruments, more generally we have adapted instruments originally devised by others. In any event, we now have a range of measuring instruments that help us collect, analyze, and evaluate survey data. All these distinct instruments have their proper roles and their own built-in limitations. Such complexity almost seems overwhelming. But the good news is that we understand the picture better and we have, or are, developing the tools to do the job. We can do it, and do it well if we are stubborn enough, meticulous enough, and intellectually honest enough.

Part V: Postscript

On Making Decisions about Language Projects

Calvin R. Rensch

More than a decade ago SIL began to consider a set of 'Faith Goals for the '80s.' These dealt with aspects of the work in which progress was crucial if the work of the organization was to move forward. There were no ready solutions or programs. We agreed to call upon God's help in a united way as well as to take new initiative ourselves in these areas.

One of these areas of concern dealt with survey and the identification of those languages that yet need Bible translation.[1] During the decade, substantial progress has been made in the number of languages surveyed, the number of workers involved in survey activities—many of them full-time survey workers—and the refinement of techniques for gathering information relevant to decisions about new language projects.

Toward the end of the decade, about a year ago, scores of survey specialists and administrators from every part of the SIL world met at Horsleys Green to share experience in language assessment, expound new survey techniques, and set directions for new research. During the conference a statement on language assessment criteria was discussed and adopted by the participants. That statement was later edited and approved by the area directors and vice presidents and has been placed in circulation

[1] References in this paper to "language projects" are intended to refer to projects in which Bible translation activities are included as a prominent part of the project. This more inclusive term is used to indicate the wide range of activities typically included in projects undertaken by the Summer Institute of Linguistics.

provisionally by the Board Committee on Academic Affairs with the expectation that there will be ongoing opportunity for discussion of these issues.

A large part of the presentations and discussion at the Language Assessment Conference dealt with issues of bilingualism, patterns of language use in multilingual communities, and language attitudes. This emphasis reflects the fact that the range of data being collected and interpreted by survey specialists and administrators extends beyond issues of linguistic similarity and dialect intelligibility.

When my wife, Carolyn, and I began to work with the Chinantec people in Mexico, language use there was quite straightforward. Spanish was used in the local school (principally by the teacher) for making purchases from a few resident and itinerant merchants (although most of the itinerant merchants had learned Chinantec), and in occasional contacts with officials above the village level. Other than that, the Chinantec language was the unchallenged vehicle of communication in the dozens of Chinantec hamlets.

On one occasion our supervisor came to visit us in the village and was introduced to the elected leader of the village, who had just returned from a trip to the central town of the district. Our supervisor asked the village leader in Spanish, "How was your trip to Choapan?" To this the village leader replied, "Yes," not understanding even that simple question. The level of proficiency in Spanish for most adults at that time was obviously quite low. Not surprisingly, the Chinantec people believed that their own language was the finest one for nearly all purposes. Therefore, when the branch conducted a survey among the various groups of Chinantec people, the nearly exclusive focus of that survey was to map out the network of mutually unintelligible Chinantec languages.

The situation in which many survey specialists are working today is quite different from that. Frequently encountered complicating factors include the following:

1. More than one vernacular (local, ethnic) language is spoken in a single community;
2. Multiple varieties of vernacular languages, associated with different castes or other social groups, are encountered;
3. More than one major language (language of wider communication) is used in the area—one for education and others for government, business or religion(s);
4. Both a major language and the vernacular are used in the home, depending on the topic of conversation or the combination of family members involved in the conversation;
5. The local dialect of the major language is so different from the standard dialect of that language as to be essentially unintelligible

with it, but the standard and nonstandard dialects are called by the same name and little recognition is given to this difference;

6. Reading and writing are firmly established in connection with a language other than the vernacular; sometimes the only language considered suitable for writing is not one's mother tongue but is acquired only through education;
7. Attitudes of the people are rather negative toward the language they understand best or language loyalty may be divided among the various languages of the community.

In some areas even the concept of mother tongue becomes muddy. In Singapore, for example, the language reported as mother tongue often is the language associated with the clan of one's father, whether or not the child has ever learned a word of that language—and frequently he has not. In some communities in northwestern South America the mother's language regularly is not the language of the father nor of most households of the community. Perhaps even more perplexing are those situations of rapid language shift, frequently encountered in urban India and Singapore, in which young people adopt a new language for use in their own families and in the community, restricting use of their own mother tongue to communication with members of their parents' generation, with resultant reduction of their mother tongue both in level of proficiency and in domains of usage.

With this array of complex factors facing both survey specialist and administrator, it is not surprising that both appreciated the opportunity afforded by the conference to discuss these factors.[2]

Relevant factors for selecting language projects

Increasingly, in various parts of the world a range of factors are being considered when decisions are made concerning language projects. These factors often include the following:

1. Dialect intelligibility: intelligibility among linguistically related varieties of a language.
2. Bilingualism: proficiency of speakers of vernacular languages in a second language or in a distinct standard dialect.

[2] In some contexts each language identified as being clearly distinct and into which the Scriptures have not been translated is labeled as a "translation need." Since this ambiguous term is used in a variety of senses by different writers, I have avoided using it altogether in this paper.

3. Language use: distribution of languages (and dialects) in daily life in the ethnic community.
4. Language attitudes: attitudes toward the vernacular language and other languages spoken in the community.

Discussions of such criteria sometimes distinguish between objective and subjective data and between linguistic and sociolinguistic (or sociological) factors. In at least some situations these distinctions prove difficult to maintain.

Techniques such as word-list collecting and studying intelligibility through recorded-text testing are often thought to be objective, and less subject to the judgment of the investigator than techniques used to study the social use of language. However, there is growing evidence that the selection of items for a standard word list, the technique the investigator uses when eliciting the word lists, and his method of determining the pairs of items to be counted as similar can all affect the results, which are usually stated in mathematical terms, which heightens the impression of objectivity.

Similarly, there is concern over the comparability of results of recorded-text testing. Techniques for administering this type of test have been described in detail in Casad's (1974) *Dialect Intelligibility Testing*. Nevertheless, questions have arisen about whether results are comparable when the texts or sets of questions used vary in difficulty or cultural relevance. Fortunately, members of the South Asia survey team and perhaps others are undertaking to study the effects of varying techniques in both word-list counting and recorded-text testing.

It may also be difficult to distinguish linguistic from sociolinguistic factors. For example, in many cases the extent to which a speaker of one dialect understands a related dialect results from a blending of the linguistic factor of linguistic similarity with social factors such as being positively or negatively disposed toward the related dialect or being prepared through previous experience to handle such testing.

Let us now consider some of the factors which are relevant to the decision-making process and some of the procedures which have been developed to assess these factors.

Intelligibility patterns among related language varieties. Recorded-text testing has become established as the standard technique for investigating the extent to which speakers of one dialect understand a linguistically related dialect. Often the linguistic variation within a dialect network is sampled by collecting word lists and calculating the percentage of similar vocabulary. Word-list comparison rarely, if ever, provides sufficient information about comprehension among related dialects, but it can provide useful data regarding which dialects should be included in the recorded-text testing.

If the scores of ten subjects who have answered questions about the content of the same text are quite similar, we usually assume that their scores accurately represent the understanding of their fellows. However, if there is a broader range of scores, we may infer that some of the subjects have had greater opportunity than the rest to learn the other dialect and that, therefore, the scores reflect a combination of inherent similarity of the dialects and varying degrees of dialect learning. In such a case a larger group of subjects, representing various subgroups in the community, must be sought for the testing since we are in this case faced with a form of bilingualism.

Proficiency in a second language. If we accept the possibility that some groups may have effective access to the Scriptures in a second language in which they are proficient, it is important for us to evaluate carefully the proficiency of such groups in the second language, whether the proficiency be in a major and genetically unrelated language or be in a distinct, but standard dialect of some particular vernacular. Members of SIL are usually oriented to recognize the value of work in vernacular languages, so they are probably more easily persuaded than others that a group does not have adequate second-language proficiency. However, we are increasingly seeking to involve others, including speakers of major languages, in either preparing or distributing the Scriptures in vernacular languages. In these cases it becomes all the more important to demonstrate clearly the proficiency (or lack of it) of the vernacular-speaking community in the major language. So, our techniques for evaluating second-language proficiency need to be persuasive.

Typically, some members of a speech community have greater opportunity to learn a second language than other members. Therefore, it is important to test the proficiency of a wide range of members of the community. Variables that increase contact with the second language and often found to be significant are: (a) sex, (b) age, (c) level of education, and (d) amount of travel or residence in areas where the second language is spoken.

Several types of tests have been used to probe second-language proficiency. In some surveys a recorded-text test in the second language has been administered to a range of subjects. This technique has the appeal of using a test which is rather easy to develop and which can easily be added to a battery of recorded-text tests administered for dialect intelligibility testing. However, it is doubtful that a single recorded-text test can probe the range of language difficulty required for bilingualism testing and especially the higher levels of difficulty. Since the factor of second-language proficiency becomes especially critical when there is the possibility that a high percentage of speakers in various subgroups have high levels of second-language proficiency, it is precisely at the upper end of the difficulty scale where our evaluation must be accurate.

Therefore, two other types of test have been developed recently to test second-language proficiency: the Second-Language Oral Proficiency Evaluation (SLOPE) and the Sentence Repetition Test (SRT). The first method is an adaptation for unsophisticated subjects of the oral interview developed by the Foreign Service Institute. It seeks to probe both the active and passive proficiency of the subject. Results of the interview are stated in terms of the FSI levels of 0 to 5. The second method is an adaptation of a technique employed in speech pathology to screen numbers of subjects into groups without attempting to diagnose specific dysfunctions of any subject. In this test the subject is asked to repeat a set of tape-recorded sentences which have been graded for difficulty and ability to discriminate differences in proficiency. The accuracy of the repetitions is scored by the administrator. In the development of a sentence-repetition test, the scores for that test are calibrated in relation to levels of second-language proficiency called Reported Proficiency Evaluation (RPE). Results of the SRT are stated in terms of RPE levels.

The advantages of the SLOPE are that it is a direct test of second-language proficiency and the results are stated in the familiar terms of the FSI scale. However, administering the interview requires the services of a team of three, at least one of whom must have extensive training and certification from a central certifying group. In field tests, conducting and scoring each interview has required considerable time. The SRT is an indirect test of second-language proficiency which infers such ratings from the subject's performance in repeating sentences. It takes considerable time to develop and calibrate such a test for each test language. However, this type of test has proved to be acceptable and nonthreatening to unsophisticated subjects, and it can be administered in just a few minutes, making it feasible to test a large number of subjects in a short time frame. Furthermore, administrators can be trained quickly to score the subjects' performances.

In situations where it is essential to use a direct test or to have results stated in terms of FSI levels and if a longer time is available, the SLOPE test is appropriate for evaluating communal bilingualism. In situations where it is important to test a wide range of subjects in a limited period of time, and if a centrally certified administrator cannot be present to administer all tests, the SRT is appropriate.

Distribution of languages in the ethnic community. It is possible for two speech communities to use the same two languages in the course of daily living yet use those languages in very different ways. One community may use the major language only for formal education and central government functions while using the vernacular language in the home, community, local trade, religion, and other domains of community life. The other may use the vernacular when discussing some topics in the home but use only the major language in all other contexts of daily life. The inventory of languages used in the two communities is the same, but the patterns of language use are sharply contrasting. In the former case, the vernacular is in a dominant position with the second language occupying a few rather marginal public roles, while in the latter case the vernacular appears to be severely threatened and even linguistically reduced by the dominance of the second language.

Of course, for our interests, the language that is used in the home domain is very important since that is the environment in which beliefs and values are usually communicated. Typically, this is the stronghold of the vernacular and is one of the reasons for our belief in the power of vernacular languages. On the other hand, we need also to be interested in the overall distribution of languages as in the cases mentioned above. Especially we should not fail to understand the influence exerted by languages used in powerful and prestigious domains, such as education and religion, or in such media as writing, radio, and television. Typically, these are the domains of second languages. If such languages are used extensively in powerful domains and in the media, the potential for development of the vernacular is probably restricted.

In this connection it is helpful to try to understand both the language policies of the nation, where they exist, and the informal language policies, which might be called traditions, of the local community. Both kinds of policy regulate language-use practices. Longstanding, formal language policies can be expected to constrain the direction of change in language use to a considerable extent. Naturally, if formal language policies have not been developed or are changed frequently, language-use patterns will be more unpredictable.

Observation through participation in community life is our most common method for understanding the distribution of languages. Eliciting information from members of the community may also be useful. However, there may be discrepancies between self-reported language use and observations on this topic, which probably reveal less about language use than about language attitudes held, i.e., what the speakers of the language believe—or would like others to believe—about which languages they use in different contexts.

Attitudes toward the vernacular and other languages. It is important to know something of the attitudes speakers hold regarding both vernacular languages and second languages that are spoken in their communities. Attitudes of those who are primarily speakers of the vernacular are important, of course, but attitudes of those who are primarily speakers of the major language may also be important.

This is the type of information which is probably the most difficult to collect. In some cases, members of the speech community may be hesitant to share their feelings with outsiders. However, at least as significant an obstacle is the fact that they often have not thought about language attitudes and do not know how to express them. Consequently, information on this topic is likely to be anecdotal and collected informally in the course of living in the community. Researchers sometimes ask speakers of one language how they feel about various kinds of social interaction with speakers of other languages on the assumption that attitudes toward a speaker and attitudes towards his language are intimately related. However, some kinds of social interaction among speakers of different languages are restricted for cultural reasons.

Nevertheless, responses concerning permitted interaction can reveal language attitudes. Indirect tests are used by some researchers for discovering language attitudes, but these have not yet found much application in the unsophisticated speech communities with which we typically deal.

It is also helpful to explore the beliefs which speakers hold concerning the extent of their language community, i.e., to ask them which varieties are part of their language. It would be difficult to carry out a single language project for two communities who speak intelligible varieties of speech but who for social reasons regard their speech varieties as separate languages. By contrast, we should seriously consider the consequences of beginning totally separate language projects for groups which intelligibility testing shows to have distinct languages but which the speakers regard as the same language—not just the same ethnic group. Such a situation may call for a complex language project which would produce for some purposes separate bodies of literature and for other purposes a single body of literature, thus giving recognition to perceptions of linguistic unity.

Decisions about beginning language projects are often based partially on predictions, either stated or unstated, regarding the vitality of the vernacular language and the expected direction and pace of language shift. Such predictions of language vitality are frequently made on the basis of (a) perceptions about language attitudes, (b) observed or reported increases in second-language proficiency on the part of vernacular-language speakers, and (c) observable changes in the contexts in which the various languages of the community are used. It hardly needs to be mentioned that such predictions are quite hazardous and not infrequently prove to be incorrect. However, they may be more valid if they are founded on documented changes rather than on presumed trends.

Inaccuracy is not the only hazard related to predictions of language vitality. In addition, we should guard against the assumption that new language projects are justified only for languages that show promise of being spoken into the future. Of course, if a language is spoken by only a handful of elderly people it is doubtful that a project could progress far before the language reaches extinction. However, if there is evidence that the language will be spoken for another generation, that in itself should be sufficient since we have a primary obligation to minister effectively to those of our own generation without being unduly concerned about what may be the speech habits of future generations.

In this section and in the two previous ones, I have discussed issues of bilingualism, language use, and language attitudes as though they were totally independent factors. In fact, these factors illuminate various aspects of an implicit competition in many communities between two or more speech varieties regarded by those communities as different languages. In some communities with stable bilingualism, especially those with diglossia, the roles of the languages are apparently fixed, but in many other communities there is evident "jockeying for position" between the languages, with the roles appearing to be in the process of shifting. In either situation the factors of these three types are often interlocking in a kind of "conspiracy" in which the effect of a factor of one type reinforces the effect of a factor of another type. For example, a strong belief in the practical value of using the second language usually reinforces the strength of that second language in various domains of public life for both speaking and writing and promotes rising levels of second-language proficiency. By contrast, low levels of second-language proficiency and acknowledgement of that by speakers of the vernacular tend to limit the domains in which the second language is used in daily life and may reduce the speakers' view of the value of knowing or acquiring that second language.

Toward a decision-making process

The types of information discussed in the preceding sections are useful in making various kinds of decisions regarding language projects. However, not all types of information are employed for the same purpose or at the same stage in the decision-making process.

The various types of information can be incorporated in a four-step decision-making process as follows:

1. Identify the clearly distinct languages.
2. Study the extent of second-language proficiency. Make a decision about whether to begin a language project.
3. Study language-use patterns and language attitudes. Make a decision about what kind of project should be started.
4. Reassess periodically the factors studied in steps two and three and evaluate the suitability of previous decisions.

Step one. A clearly fundamental question is that of identifying distinct languages. This is done primarily through dialect intelligibility testing. Through such testing it can be determined that each language variety falls into one of three categories:

1. It is a clearly distinct`language; that is, it is not inherently intelligible with any other language.
2. It is clearly not a distinct language; that is, it is inherently intelligible with another language variety.
3. It is a marginally distinct language; that is, the data suggest that this variety may be sufficiently intelligible with another language variety to enable speakers to use a common body of literature. With some language varieties of this type it may be wise to wait until materials in the other language variety are available for testing of comprehension and acceptability.

Step two. Since the possibility should be considered that literature can be effectively provided for a group of people in a second language in which they are proficient, it is important to undertake a study of their second-language proficiency. If that possibility is not evaluated at the time of the initial decision, it is likely to be raised later on in a way that will be disturbing to those engaged in the project.

However, if there is no immediate prospect of a language project being started for such a group, it may be wise in some situations to defer the study of bilingualism until there is such a project. Levels of second-language proficiency sometimes change rapidly. If the decision to start a language project is not made until some years after the bilingualism data are collected, parts of the bilingualism study will almost certainly need to be redone.

It is often wise to initiate a study of second-language proficiency with a pilot study, which looks for indirect and more easily collected indications of bilingualism levels.

Observations made about the frequency with which the second language is used may provide indirect evidence about levels of proficiency in the second language since frequent use often, but not always, leads to proficient use.

The contexts in which the second language is used in daily life may provide further evidence based on the level of proficiency demanded by those contexts. For example, using a second language in an occupation which is dependent on language usually requires a higher level of proficiency than does using a second language in an occupation, such as fishing or agriculture, not heavily dependent on language.

A study of the sources of second-language proficiency may be helpful. For example, if higher levels of proficiency are gained only through secondary school and most people in the community complete only primary school, it is unlikely that many attain high levels of proficiency in the second language. By contrast, if most young people in such a community complete secondary school, a more extensive study of second-language proficiency is probably called for.

Reports of vernacular speakers about their second-language proficiency and evaluations by mother-tongue speakers of the second language may also prove helpful in drawing conclusions about general levels of second-language proficiency.

If the pilot study suggests that high levels of second-language proficiency are widespread in various sub-groups of the community, a more extensive bilingualism study is called for. In that case a profile should be developed of a village thought to be representative. For developing such a profile, information is gathered about the members of each household. This information includes factors expected to be significant in the distribution of bilingual proficiency throughout the community—sex, age, education, occupation, travel, and other widespread language-contact factors. After the size of each of these subgroups in the community is calculated, a sampling of subjects is sought which will accurately represent the various subgroups. Even if it proves impossible to select a sampling that truly reflects the numerical strength of the sub-groups, the profile can guide in determining the weight to be attached to the scores of the subjects in each subgroup.

Steps three and four. After a bilingualism test of the sort described earlier has been administered to a representative group of subjects, levels of proficiency in the various subgroups of the community are examined so as to learn whether high levels of second-language proficiency are restricted to one or two advantaged subgroups or whether they are well distributed throughout the community.

In many cases a decision can be made at this point about whether a language project should be initiated. In such cases a project is warranted if (a) the vernacular is a clearly distinct language, and (b) high levels of second-language proficiency are either lacking or are not well distributed throughout the community.

In some other cases, especially if the levels of second-language proficiency are only moderately high, the proper decision may be unclear. In such cases it may be wise to defer a decision until further types of information such as language-use patterns and language attitudes can be considered.

A decision not to initiate a language project because of factors which can change should be made with special caution since it is at this point that a group which really needs vernacular literature could be denied it because of an inappropriate decision. Of course, the factors in a situation such as this should be reassessed periodically and the decision reviewed.

Specifying the nature of the project. Even after a basic decision has been reached to begin a language project, it is not always clear just what type of project is warranted. Some projects must include a major effort in introducing the skills of literacy, whereas others can simply reinforce the efforts of other people to promote literacy or produce literature. In some projects the public use of vernacular Scriptures can be expected, whereas in others there is a need to focus on promoting the private use of vernacular Scriptures, at least at first, because public use of Scriptures in the major language is already well established. Some projects call for early production of vernacular literature of various sorts to meet an obvious interest, whereas others must focus first on changing local attitudes toward using the vernacular in written form.

Decisions about what kind of language project should be started should be based on a wide spectrum of information. Data about patterns of language use in the community and about attitudes held toward the inventory of languages used by the community should be available before such decisions are made.

However, use of such information is not entirely limited to decisions about what kind of project should be started. As stated above, information of this type is sometimes required for reaching a decision about whether to start a project in a marginal situation, especially where the implications of bilingualism are unclear.

Periodical reassessment. When a decision has been made to start a language project of a certain type and after the project is underway, it is important to reassess the situation periodically to make sure that the factors which suggested that type of program at the original decision point have not changed.

It is not so likely that factors which led to a decision not to start a language project will change so that starting such a project will be necessary. However, the possibility of such a change cannot be ruled out, and the more changeable factors, such as bilingual proficiency and language attitudes, should be periodically reassessed.

Beyond decisions of whether a language project should be started and what should be the nature of such a project lie decisions of which language projects should receive attention first.

Certainly, there are factors apart from the ones discussed above which affect what we do. In some cases speakers of a language express an interest in having a language project and demonstrate that interest by providing workers who are to be trained and helped to carry out a project in their own language. In other cases, such interest is expressed by Christians ministering in the area even though they are not themselves mother-tongue speakers of the language. In still other cases, an expatriate team may sense a special burden from God for a given language group. Almost invariably we seek to develop a project of some sort in any of the above situations if feasible, even though other factors may not suggest special urgency.

Areas for further discussion and study

Many of the issues discussed here are matters on which there seems to be consensus. General agreement about their importance was reflected in the discussions at the Language Assessment Conference.

We seem to agree that decisions about where to begin language projects must be based on information concerning intelligibility among related dialects. Information about linguistic (especially lexical) similarity is helpful in guiding us to places where intelligibility testing should be carried out and gives us a preliminary indication of the results that may be expected from testing. However, an understanding of the limits of the area where the language is spoken (or the range in intelligible dialects) is generally only derived from the results of intelligibility testing.

We regularly interpret the results of intelligibility testing by assigning each variety a status which is "clear" (clearly a distinct language or clearly not a distinct language) or "marginal" (marginally distinct). With the many varieties which are marginal we feel keenly the need for further kinds of information.

We also regularly gather and interpret data concerning bilingualism, language-use patterns, and language attitudes, and make estimates of language vitality. There is probably less agreement on the precise role that these factors should play in the decision-making process. Some of my beliefs regarding their use have been stated in the previous section.

Interestingly enough, there appears to be agreement on some factors which we do NOT regard as critical in making decisions about language projects. I mention these here because they have been mentioned in recent discussions as factors which may be prominent in the thinking of some members. I would like to mention three such factors:

1. Size of the language group
2. Expectation of favorable response to Scripture in any language or medium
3. Approval by local leaders

As a group, I believe that we do not consider the size of the language group as critical once it has been established that there are enough speakers to form a viable speech community (however defined) and that they are distributed in age groups other than just the most elderly. I believe that we are prepared to undertake a language project for any small language group which includes speakers who need and could use vernacular materials by the time they would be ready.

As a group, I believe that we do not consider as a critical factor the likelihood of a favorable response to the Scriptures. That is to say, I believe that we are prepared to support the translation of the Scriptures for a group even when the group has not accepted the Scriptures in a major language of the area and when religious and cultural factors do not encourage us to expect a favorable response, at least not initially.

As a group, I believe that we do not regard as essential for beginning a language project either a request for, or approval of, a language project from local leaders, political or religious. Of course, community support for a project is an advantage of enormous proportions. It is probably especially important in communities where higher levels of bilingualism or ambivalent attitudes toward the vernacular are found. Lack of such support has significant implications for the nature and emphases of any language project undertaken for that group. However, I believe that we are prepared to begin a language project, if needed for reasons of language and if feasible, for a language group in which we cannot see evidence of support from local leaders.

The conditions in different nations and regions vary considerably. So, factors which seem crucial in one area may appear to have little relevance or importance in another. Consequently, the details of making decisions about language projects will probably continue to vary and will involve the judgment of local administrators, who, while considering agreed-upon factors and procedures, will in many cases need to exercise their own judgment.

In spite of these differences it is important that we agree upon the basic factors which we all consider when making decisions about beginning (or continuing) language projects. There appear to be some matters in this area about which we do not have agreement. I will mention a few of them here in the hope that further discussion will show us that in fact we do have essential agreement or else will help us reach consensus on these issues.

1. Should we assume that the mother tongue is always the most effective language in which to use the Scriptures? Are there situations in which vernacular speakers who have effective access to a second language could, for various social reasons, be better served by using the Scriptures in that second language?
2. Are we committed to begin or encourage others to begin a language project for every clearly distinct language if at all feasible? Are there situations in which we would conclude that it is not worthwhile to begin a project for a clearly distinct language without the Scriptures?
3. Do we need to collect and consider information other than dialect-intelligibility information? Are bilingualism level, language use, language attitudes, etc., relevant for deciding where to begin a language project? Are there other types of information relevant for deciding what type of language project should be started?
4. Some types of information can be ranked along a scale of more favorable or less favorable to the use of vernacular literature or can indicate that a group has greater dependence or less dependence on Scriptures in their vernacular language. If we are to use such information in making decisions about beginning language projects, should we rank potential language projects according to a system of priorities? If so, what should that system of priorities be?
5. What levels of second-language proficiency are required for a person to make adequate use of the Scriptures in a second language? How widely must such levels of second-language proficiency be distributed through the speech community? Can we (or, need we) reach consensus on this matter?

[blank]

Appendix A
Interview Score Sheet

Name:
Interviewer:
Comments:
Score:
Weighting score:

Proficiency level	A	B	C	D	E	F	
Accent							
Grammar							
Vocabulary							
Fluency							
Comprehension							
						TOTAL	

Proficiency description for weighting procedures
Accent:

A. Pronunciation frequently unintelligible.
B. Frequent gross errors and a very heavy accent make understanding difficult, requires frequent repetition.
C. Foreign accent requires concentrated listening; mispronunciations lead to occasional misunderstanding; apparent errors in grammar or vocabulary.
D. Marked foreign accent and occasional mispronunciations that do not interfere with understanding.
E. No conspicuous mispronunciations, but would not be taken for a native speaker.
F. Native pronunciation, with no trace of foreign accent.

Grammar

A. Grammar almost entirely inaccurate except in stock phrases.
B. Constant errors showing control of very few major patterns and frequently preventing communication.
C. Frequent errors showing some uncontrolled major patterns and causing occasional irritation and misunderstandings.
D. Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding.
E. Few errors, with no pattern of failure.
F. No more than two errors during the interview.

Vocabulary

A. Vocabulary inadequate for even the simplest conversation.
B. Vocabulary limited to basic personal and survival areas (time, food, transportation, family, etc.).
C. Choice of words sometimes inaccurate, limitations of vocabulary prevent discussion of some common professional and social topics.
D. Professional vocabulary adequate to discuss special interests; general vocabulary permits discussion of any nontechnical subject with some circumlocutions.
E. Professional vocabulary broad and precise; general vocabulary adequate to cope with complex practical problems and varied social situations.
F. Vocabulary apparently as accurate and extensive as that of an educated native speaker.

Fluency

A. Speech is so halting and fragmentary that conversation is virtually impossible.
B. Speech is very slow and uneven except for short or routine sentences.
C. Speech is frequently hesitant and jerky; sentences may be left uncompleted.
D. Speech is occasionally hesitant, with some unevenness caused by rephrasing and groping for words.
E. Speech is effortless and smooth, but perceptibly nonnative in speed and evenness.
F. Speech on all professional and general topics as effortless and smooth as a native speaker's.

Comprehension

A. Understands too little for the simplest type of conversation.

B. Understands only slow, very simple speech on common social and touristic topics; requires constant repetition.

C. Understands careful, somewhat simplified speech, with considerable repetition and rephrasing.

D. Understands normal educated speech quite well, but requires occasional repetition or rephrasing.

E. Understands everything in normal educated conversation except for very colloquial or low-frequency items, or exceptionally rapid or slurred speech.

F. Understands everything in both formal and colloquial speech to be expected of an educated native speaker.

Appendix B
Karao Survey Data

The personal data for each of the subjects tested is given along with their individual proficiency scores as determined by each of the four distinct instruments employed in this study, as well as an average proficiency and average comprehension score. These data are also separated into three groups, corresponding to the particular sets of subjects evaluated by each of three distinct testers.

Group 1: Tester A

			Educa-	Self-	Self-	Interview			Average	Tape			Average
No.	Sex	Age	tion	score	test	1	2	3	profic.	1	2	3	compn.
1	F	50	6	3+	3+	3	3	2+	3	89	72	77	E
2	F	43	1HS	3+	3+	3+	3+	3+	3+	100	100	93	E
3	M	49	COLL	4	3+	4	4	4+	4	100	100	97	F
4	F	26	COLL	3+	3	4	4+	4+	4+	100	100	97	E
5	F	43	6	4	3+	3+	3+	3+	3+	100	100	100	E
7	F	20	3COLL	3+	3+	3	3	3+	3	100	94	100	E
8	F	45	COLL	3+	3+	4+	3+	4	4	100	94	97	E
9	F	55	2	1	2+	0+	1	1	1	94	89	53	D
11	F	55	0	2+	3	1+	1+	1+	1+	83	100	77	C
12	F	45	4	2	1+	1	1	1	1	83	67	73	E
13	M	21	3COLL	3+	3+	4	3+	4	4	100	94	100	E
14	M	32	COLL	3+	3+	4	3+	4	4	100	100	100	E
15	M	59	COLL	4	4+	4+	4+	4+	4+	100	100	100	F
16	F	59	6	3+	4+	3	3	2+	3	100	100	100	E
17	M	13	1HS	2+	3	3+	3	3	3	100	100	90	E
18	M	26	1COLL	3+	4	3	2+	2	2+	89	100	93	C
19	F	18	2COLL	4	3+	3+	2+	3	3	100	100	93	D

21	M	32	2COLL	4	3+	4	3+	3+	3+	100	100	83	E
22	F	26	COLL	4	3+	3+	3+	3+	3+	100	100	93	E
23	M	50	6	2+	4+	3	3+	3+	3+	89	89	90	D
24	M	16	HS	3+	3+	3+	3	3+	3+	100	89	83	E
25	F	43	1HS	3+	3+	3+	3	3	3	100	89	93	E
26	M	30	COLL	4	3+	4	4	4	4	100	94	97	E
28	M	23	3COLL	3	1+	4	3	3+	3+	100	100	100	E

Group 2: Tester B

			Educa-	Self-	Self-	Interview			Average	Tape			Average
No.	Sex	Age	tion	score	test	1	2	3	profic.	1	2	3	compn.
29	F	19	1COLL	3	3+	3	3+	4	3+	100	100	93	D
31	F	41	COLL	3+	3+	4	4	4	4	100	100	93	E
32	M	22	2HS	3+	3+	2+	2+	3	2+	100	94	93	D
35	M	28	3HS	3	3+	3	3	3	3	89	94	87	D
36	M	31	3COLL	3	3+	3+	3	3+	3+	89	94	93	E
38	M	25	1COLL	3	3+	1+	2	1+	1+	100	94	93	C
39	M	35	5	3	3+	3	2+	2+	2+	100	100	83	D
40	M	48	2	3	3+	2	1	1+	1+	95	83	57	C
41	F	19	2COLL	3	2+	2+	2	2	2	100	100	97	D
42	F	28	COLL	3+	3+	3+	3+	4	3+	100	94	97	E
43	M	60	2	2	3+	2	2	1+	2	100	78	77	D
44	M	17	6	3	2	1	1	1+	1	100	61	70	B
46	M	35	6	2+	2+	2	2+	2+	2+	100	94	63	D
47	M	17	6	2+	3+	1	1+	1+	1+	100	94	90	D
48	F	19	1COLL	3	3	1	1	1	1	100	89	83	C
49	F	27	3HS	3	2+	3+	3	3+	3+	100	94	93	D
50	M	15	1HS	3	2+	1	0+	1	1	95	67	73	C
51	F	16	6	2+	3+	0+	1	1	1	100	83	97	C
52	F	13	4	2	3	0+	0+	0+	0+	100	67	83	B
53	F	14	6	1+	0+	0+	1	1	1	100	89	80	C
54	M	65	6	3	3+	2	1	1+	1+	100	100	93	D
55	F	23	4	1	0+	0+	1	1+	1	100	100	90	C
56	M	61	1	2	2+	3	3	2+	3	90	61	73	D
57	M	24	4	2+	3+	2	1+	2+	2	95	72	73	D
58	F	30	1HS	2	2+	2+	2	1+	2	100	100	80	D
59	F	32	2	1	0+	0+	0+	0+	0+	100	100	100	C

Group 3: Tester C

			Educa-	Self-	Self-	Interview			Average	Tape			Average
No.	Sex	Age	tion	score	test	1	2	3	profic.	1	2	3	compn.
60	F	21	HS	2+	2+	1	1	1	1	100	94	97	D
61	M	31	HS	2	2+	3	3	3	3	100	94	100	D
62	F	25	6	1+	1+	1	1	1	1	100	83	90	C
63	M	14	1HS	2+	3+	2+	2	2+	2+	100	100	100	D
64	F	38	2COLL	3	4+	2	1+	2	2	100	100	100	D
65	M	14	1HS	2+	3+	2+	3	2+	2+	100	100	100	D
66	F	30	HS	2+	3+	1+	1	2	1	100	94	97	C
68	F	42	2HS	2+	3+	2	1+	1+	1+	100	100	100	D
69	F	22	COLL	3+	3	3	2+	3	3	100	94	100	D
70	F	19	1COLL	3	3+	1	0+	1+	1	100	100	97	C
71	M	43	COLL	3+	3+	3+	3+	4	3+	100	100	97	E
72	M	27	1HS	3	3+	1	1+	1	1	100	100	90	C
73	F	22	COLL	3+	3+	3+	3+	4	3+	100	100	100	E
74	F	36	HS	2+	3+	2+	3	3	3	100	94	100	D
75	M	63	6	2+	3+	1+	1+	1+	1+	100	100	87	D
76	M	54	1HS	3+	3+	2+	2+	2+	2+	100	94	90	E
77	M	63	2HS	2+	3	2	1+	1+	1+	100	100	97	E
78	M	14	6	2	1	1	1	1+	1	100	100	97	D
79	F	13	5	2+	2+	1	0+	0+	0+	100	94	93	C
80	M	13	4	1+	0+	0+	0+	0+	0+	100	89	90	C

Appendix C
Karao Self-test

S-0+ Can you speak Ibaloi just a little bit?

S-1 (A) Could you explain the way from here to . . . ?

(B) Can you understand and respond properly to questions about where you are from and if you are married?

S-2 (A) Can you describe in detail your present work?

(B) Can you describe your family, your house, and the weather today?

(C) Can you hire someone to work for you and arrange such details as salary, hours, and specific duties?

(D) Do Ibaloi speakers understand you nearly all of the time?

S-3 (A) Can you listen to a conversation among Ibaloi speakers and be able to summarize what you have heard?

(B) Can you debate well in Ibaloi?

(C) Can you arrange with Ibaloi speakers to build a new house, explaining just how you want it built?

(D) Are you sometimes unable to finish a sentence because you don't know how to say something in Ibaloi?

S-4 (A) In discussions with Ibaloi speakers, can you always say exactly what you want to say?

(B) Can you accomplish whatever task in Ibaloi, just as if it were in Karao?

(C) Can you speak Ibaloi well, even when you're angry?

(D) Do you sometimes make mistakes when you speak Ibaloi?

S-5 (A) Can you use as many words in Ibaloi as in Karao?
(B) Sometimes is it easier to think in Ibaloi than it is in Karao?
(C) Do you speak Ibaloi as well as an Ibaloi speaker?
(D) Do people know that you are not an Ibaloi by the way that you speak Ibaloi?

Appendix D

Computer Program in BASIC to Calculate Correlations

Runs on the Sharp PC-5000, Kaypro 2000,
and other IBM PC compatibles.

```
1     ‘SIMINT - Correlate vocabulary similarity with intelligibility
2     ‘Joseph E. Grimes, 1986 October 13
3     ‘
10    LI=100 : LN=10 : L=60
20    DIM SD(LI), ID(LI)
30    DIM SL(LN), IL(LN), RROW(LN), NROW(LN)
40    NS=9 : NI=7
50    ‘
100   FOR I=1 TO NS : READ SL(I) : NEXT I       ‘Similarity threshholds
110       DATA 0,60,65,70,75,80,85,90,95
120   FOR I=1 TO NI : READ IL(I) : NEXT I       ‘Intelligibility thresholds
130       DATA 95,90,85,80,75,70,0
140   ‘
200   INPUT "File or device name for output";O$
210       OPEN O$ FOR OUTPUT AS #2
215       IF O$="LPT1" OR O$="lpt1" THEN PRINT#2,CHR$(27);"*1"
220   ‘
230   FOR Z=0 TO 1 STEP 0
235       IF O$<>"LPT1" AND O$<>"lpt1" THEN BEEP
240       INPUT "File name and extension [NNNNNNNN.XXX] for the data";F$
250       IF F$="" THEN CLOSE #2 : STOP
260           OPEN F$ FOR INPUT AS #1
270   ‘
300       FOR K=1 TO LI                         ‘Read the data
310           INPUT #1, SD(K), ID(K)
```

```
320           IF SD(K)=0 AND ID(K)=0 THEN 350
330        NEXT K
340        K=K+1
350        K=K-1                                    'K pairs read in
360        CLOSE #1
370
400        PRINT#2,TAB(4);F$;TAB(5+NS*5+3+4);F$ : L=L-1
410        PRINT#2,"Intelligibility> =I%";          TAB(5+NS*5+3);
           "Intelligibility> =I%" : L=L-1
420        FOR I=1 TO NI
430           IN=IL(I)                              'By intelligibility thresholds
440           PRINT#2,USING"##% ";IN;
450           GOSUB 500                             'Correlate with similarity
460        NEXT I
470        GOSUB 900                                'Bottom of graph
480     NEXT Z                                      'End of one data set
490     '
500     FOR J=1 TO NS                               'Correlate with similarity
                                                    (450)
510        SIM=SL(J)                                'By similarity thresholds
520        GOSUB 600                                'Print the correlation
530        NROW(J)=N : RROW(J)=R                    'and save it for the graph
540     NEXT J
550     GOSUB 800                                   'Print the graph
560     RETURN
570     '
600     N=0 : SX=0 : SY=0 : XY=0 : X2=0 : Y2=0      'Correlation (520)
610     FOR M=1 TO K                                'Go through the data
620        X=SD(M) : Y=ID(M)
630        IF X<SIM OR Y<IN  THEN 650 ELSE N=N+1
640        SX=SX+X : SY=SY+Y : XY=XY+X*Y : X2=X2+X*X :
           Y2=Y2+Y*Y
650     NEXT M
660     '
700     IF N<5 THEN R=0 : PRINT#2,"     *"; : GOTO 780
710     A=SX/N : B=SY/N : C=XY/N : D=X2/N : E=Y2/N
720     YX=XY-(SX*SY) : XX=A*A-D
730     BB=YX/XX                                    'Slope of regression line
740     AA=B-BB*A                                   'Intercept of regression line
750     R=(XY-(SX*SY)/N) / SQR(ABS(X2-(SX*SX/N))*ABS(Y2-(SY*SY/N)))
                                                    'Correlation
760     IF R1 OR R1 THEN PRINT#2,"   ??"; : GOTO 780
770     PRINT#2,USING" #.##";R;
780     RETURN
790     '
800     PRINT#2,USING"    ##%";IN;                  'Print the graph (550)
```

```
810  FOR J=1 TO NS
820      IF NROW(J)<5 THEN PRINT#2," *"; : GOTO 850
830      IF RROW(J)<0 THEN PRINT#2," -"; : GOTO 850
840      IF RROW(J)>=0 THEN PRINT#2,USING" #";INT(RROW(J)*10);
850  NEXT J
860  PRINT#2,"" : L=L-1
870  RETURN
880  '
900  PRINT#2,TAB(10);                          'Bottom of graph (470)
910  FOR J=2 TO NS : PRINT#2,"-----"; : NEXT J
920  PRINT#2,TAB(5+NS*5+6);
930  FOR J=1 TO NS : PRINT#2,"--"; : NEXT J
940  PRINT#2,"" : L=L-1
950  '
1000 PRINT#2,TAB(5);
1010 FOR J=1 TO NS : PRINT#2,USING"   ##%";SL(J); : NEXT J
1020 PRINT#2,TAB(5+NS*5+6);
1030 FOR J=1 TO NS : PRINT#2,USING" #";INT(SL(J)/10); : NEXT J
1040 PRINT#2,"" : L=L-1
1050 '
1100 PRINT#2,TAB(5+NS*5+6);
1110 FOR J=1 TO NS : PRINT#2,USING" #";SL(J)-INT(SL(J)/10)*10; : NEXT J
1120 PRINT#2," %" : L=L-1
1130 PRINT#2,TAB(12);"Similarity>=S% for"; K; "pairs"; TAB(5+NS*5+6+3);
     "Similarity>=S%"
1140 PRINT#2,TAB(5);"* N<5 too few to correlate   - R<0 reversed 0 to 9 R=.0
     to .9"
1150 PRINT#2,"" : PRINT#2,"" : L=L-4
1160 IF L-NI-8<=0 THEN L=60 : PRINT#2,CHR$(12)
1170 RETURN
1180 END
```

Appendix E
Proficiency Questions Used in Agutaynen Survey

S-0+ Can you speak x just a little bit?

S-1 (A) Can you understand and respond correctly to questions about where you are from, if you are married, your work, date and place of birth?
(B) Could you explain the way from here to the high school to someone who did not know?

S-2 (A) Can you describe in detail your present or former work?
(B) Could you give a brief account of your lifestyle and plans for the future?
(C) Could you hire someone to work for you, arranging his wages, qualifications, hours, and responsibilities?

S-3 (A) Sometimes do you not know how to say something in x?
(B) Do you debate well in x?
(C) Can you listen to and give a brief summary of conversations in x on topics that you are interested in?

S-4 (A) If x-speakers are debating, are you always able to say to them whatever you want?
(B) Do you speak x well even when you're angry?
(C) Can you accomplish whatever task in x, just as if it were in Agutaynen?
(D) Do you make mistakes in x?

S-5 (A) Can you use as many words in x as in Agutaynen?
(B) Sometimes is it easier to think in x than in Agutaynen?
(C) Do you speak x as well as an x-speaker?
(D) Do people know that you are not an x-speaker by the way you speak x?

Appendix F
Proficiency Scores For Brooke's Point Test

Respondent	Self report	Direct test	Tester one	Tester two
1	4+	5	5	5
2	5	5	5	5
3	4+	5	5	5
4	2+	4	4	4
5	4+	5	5	5
6	3	3+	4	3+
7	5	5	5	5
8	1+	4	4	4
9	1+	4	4	4+
10	3	4+	4+	5
11	3	3+	4	3
12	4	3+	4	3
13	3+	3+	4	3
14	3+	5	5	5
15	4+	5	5	5
16	2	3+	4	3
17	3+	3	3	3
18	5	4+	5	4+
19	5	3+	4	3
20	2	4	4	4
21	3+	3+	4	3
22	4	3+	4	3
23	5	4	4	4
24	3+	4	4	4

Respondent	Self report	Direct test	Tester one	Tester two
25	3+	3	3+	3
26	3+	3	3	3
27	3+	3	3	3
28	2+	2	2	2
29	3+	4	4	4
30	4+	4	4	4
31	4+	4	4	4
32	1	0+	0+	0+
33	3+	4	4	4
34	4+	4	4	4
35	4+	4	4	4
36	5	4	4	4
37	5	4	4	4
38	2	3	3	3
39	2+	2	2	2
40	3+	2	2	2+

Appendix G
Walker's Attitude Questionnaire

The following questions are from Walker's original questionnaire which was included as a part of his dissertation (Walker 1987:238–45). They have been renumbered here for ease of reference; the numbers in parentheses are the original questionnaire numbers. Responses are recorded as applying to the vernacular language (VL) and the national language (NL).

Criterion variables

1. (43) How many people have purchased (or wanted to receive as a gift) Scriptures (either the New Testament or Scripture portions)? VL = __ NL = __

(44) How many people have purchased (or received) other types of literature (i.e., not Scriptures)? VL = __ NL = __

2. (45) What is the percentage of the population who can read narratives with understanding? __%

VL = __% age 10–25; __% age 26–40 NL = __% age 10–25; __% age 26–40

3. (46) What percentage of the population spend time reading (any kind of literature) weekly in informal settings (i.e., outside church and school)?

VL = __% age 10–25; __% age 26–40 NL = __% age 10–25; __% age 26–40

4. (47) For each church in the community, are Scriptures read aloud in church meetings? 3 = every meeting; 2 = most meetings; 1 = some meetings; 0 = not at all

(list each church)	VL	NL	Average Attendance
__________	3 2 1 0	3 2 1 0	__________
__________	3 2 1 0	3 2 1 0	__________, etc.

Predictor variables

1. (8) How many hours' travel (by the commonest [sic] mode) is it to a town where the NL is widely used? ____ hours

2. (27) Intermarriage. Estimate the percentage of homes in the community in which one spouse is not a mother-tongue speaker of the VL. ____%

3. (21) Estimate the percentage of homes in the community where people live who are not native to the community and do not speak the VL. ____%

4. (19) Use the Rating Scale below to estimate PROFICIENCY IN THE NL for each of the categories below. Put the percentage of that category of the people in the boxes below the appropriate proficiency level. (See the Example—a situation in which 40% of the males age 10–25 are at level 0 and 60% are at level 1.)

Rating scale

Level 0. No ability.

Level 1. Can carry out minimal activities in daily living in the language.

Level 2. Can respond to opportunities and interact in routine social situations and limited work requirements.

Level 3. Can satisfy normal social and work requirements with sufficient structural accuracy and vocabulary to meet these limited needs.

Level 4. Can communicate effectively with vocabulary that is always extensive and precise enough to convey exact meaning.

Level 5. Native speaker fluency.

Example:

sex/age	0	1	2	3	4	5	
males 10–25	40	60					= 100%

Proficiency in the NL (% levels)

sex/age	0	1	2	3	4	5	
males 10–25							= 100%
males 26–40							= 100%
females 10–25							= 100%
females 26–40							= 100%

5. (30) What is the average number of years of formal education completed by adult males? ____

6. (29) What percentage of VL readers could read the NL first? ____%

7. (25) Economically, for the people in the community... (Check one)

( ) 0–most can earn a living as they traditionally have

( ) 1–a few are beginning to leave the community to find jobs

( ) 2–more and more are leaving to find jobs on the outside

( ) 3–many people leave the community to work for wages

8. (26) How important do the people feel proficiency in the NL is to economic advancement? (Circle a number) Not important – 0 1 2 3 4 – Very important.

9. (24) What is the % of the adult population needing the NL to carry out their occupation? (See no. 4 for proficiency levels) List common occupations:

___ % men need spoken proficiency at Level ___ ____________

___ % men need written proficiency at Level ___ ____________

___ % women need spoken proficiency at Level ___ ____________

___ % women need written proficiency at Level ___ ____________

10. (33) What is the prevailing attitude of local government officials, who are not VL speakers (e.g., schoolteachers or whoever is most influential in the community) to the development and use of the VL for literacy?

Negative – 0 1 2 3 4 + Positive

11–14. (9) Circle which language is most dominant in each domain for spoken use.

	Language		Domain[1]
	VL	NL	home
	VL	NL	community
11.	VL	NL	church/religion
12.	VL	NL	occupation
13.	VL	NL	school classroom
	VL	NL	government
14.	VL	NL	singing

15. (12) Estimate the percentage of the community (of any religious affiliation) who aim at living their lives according to the Bible. ___%

[1] The domains of home, community, and government did not receive sufficient response to be included in Walker's final model.

16. (22) List the number of symbols in the VL orthography that are not found in the NL orthography or which have different phonemic values.

Number	Items	Symbols
	consonants	
	glottal stop	
	vowels	
	nasalized vowels	
	vowel length (phonemic or ballistic [sic])	
	accent	
	tone	
	other	

17. (23) How difficult do people in the community view reading the VL? (Check one)

( ) 0–It is much more difficult to read than the NL

( ) 1–It is fairly difficult to read compared to the NL

( ) 2–It is about the same as reading the NL

( ) 3–It is fairly easy compared to the NL

( ) 4–It is very easy compared to the NL

18. (39) To what extent were community leaders involved in orthography decisions?

( ) 0–Actively opposed to the SIL produced orthography

( ) 1–Not involved, neutral

( ) 2–Involved and supportive of the orthography

( ) 3–Enthusiastic promoter/s of the orthography

19. (40) To what extent were community leaders involved in other aspects of the VL literacy program?

( ) 0–Opposed to it

( ) 1–Not involved at all

( ) 2–Involved to some degree

( ) 3–Actively involved

( ) 4–Enthusiastic promoter/s

References

Adams, Marianne Lehr and James R. Frith, eds. 1979. Testing kit: French and Spanish. Foreign Service Institute. Washington, DC: U.S. Government Printing Office.

Agheyisi, Rebecca and Joshua A. Fishman. 1970. Language attitude studies: A brief survey of methodological approaches. Anthropological Linguistics 12(5):137–57.

Aguilana, Esther D. 1978. Teacher reactions towards the use of Pilipino in high school mathematics. MA thesis in Education. Baguio City: Saint Louis University.

Anderson, James A. 1985. A theory for attitude and behavior applied to an election survey. Behavioral Science 30(4):219–29.

Babbie, Earl R. 1975. The practice of social research. Belmont, California: Wadsworth Publishing Company.

Ball, Peter, Howard Giles, and Miles Hewstone. 1984. Second language acquisition: The intergroup model with catastrophic dimensions. In Henri Tajfel (ed.), The social dimension, 668–94. Cambridge: Cambridge University Press.

Barcelona, Herminia. 1977. Language usage and preference patterns of Filipino bilinguals: An NMPC survey. In Emy M. Pascasio (ed.), The Filipino bilingual: Studies on Philippine bilingualism and bilingual education, 64–71. Quezon City: Ateneo de Manila University Press.

Bateson, Gregory. 1980. Mind and nature: A necessary unity. New York: Bantam Books.

Bautista, Ma. Lourdes S., et al. 1977. The Filipino bilingual's orientation. In Emy M. Pascasio (ed.), The Filipino bilingual: Studies on Philippine bilingualism and bilingual education, 72–82. Quezon City: Ateneo de Manila University Press.

Bendor-Samuel, David. 1991. Two questions surveys must answer. In Gloria E. Kindell (ed.), Proceedings of the Summer Institute of Linguistics International Language Assessment Conference, Horsleys Green, 23–31 May 1989, 9–19. Dallas: Summer Institute of Linguistics.

Bendor-Samuel, John. 1982. What answers obviate the need for further research? Notes on Linguistics Special Publication 2:9–14.

Berk, Richard A. and Thomas F. Cooley. 1987. Errors in forecasting social phenomena. In Kenneth C. Land and Stephen H. Schneider (eds.), Forecasting in the social and natural sciences, 247–65. Dordrecht, Holland: D. Reidel.

Bernard, H. Russell. 1988. Research methods in cultural anthropology. Newbury Park, California: Sage Publications.

Blair, Frank. 1990. Survey on a shoestring: A manual for small-scale language surveys. Summer Institute of Linguistics and The University of Texas at Arlington publications in linguistics 96. Dallas.

Bourhis, Richard Yvon, Howard Giles, and Doreen Rosenthal. 1981. Notes on the construction of a 'subjective vitality questionnaire' for ethnolinguistic groups. Journal of Multilingual and Multicultural Development 2(2):145–55.

Brewster, E. Thomas and Elizabeth S. Brewster. 1976. Language acquisition made practical: Field methods for language learners. Colorado Springs: Lingua House.

Brudner, Lilyan A. and Douglas R. White. 1979. Language attitudes: Behavior and intervening variables. In William F. Mackey and Jacob Ornstein (eds.), Sociolinguistic studies in language contact: Methods and cases, 51–68. The Hague: Mouton.

Bruhn, Thea C. 1989. 'Passages:' Life, the universe, and language proficiency assessment. In Georgetown University Round Table on Languages and Linguistics, 1989, 245–54. Washington, DC: Georgetown University Press.

Butler, Christopher. 1985. Statistics in linguistics. Oxford: Basil Blackwell.

Casad, Eugene H. 1974. Dialect intelligibility testing. Summer Institute of Linguistics publications in linguistics and related fields 38. Norman, Oklahoma: The Summer Institute of Linguistics and the University of Oklahoma.

———. 1990 (and this volume). State of the art: Dialect survey 15 years later. In Gloria E. Kindell (ed.), Proceedings of the Summer Institute of Linguistics International Language Assessment Conference, Horsleys Green, 23–31 May 1989, 143–55. Dallas: Summer Institute of Linguistics.

Cooper, Robert L. 1975. Sociolinguistic surveys: The state of the art. In Jonathan Pool (comp.), Proceedings of the international conference on the methodology of sociolinguistic surveys, 28–45. Washington, DC:

Center for Applied Linguistics. (Reprinted in Applied Linguistics 1(2):113–28, 1980.)

——— and Michael King. 1976. Language and university students. In Bender et al. (eds.), Language in Ethiopia, 273–80. London: Oxford University Press.

De Gaay Fortman, Clasina. 1978. Oral competence in Nyanja among Lusaka schoolchildren. In Sirarpi Ohanessian and Mubanga E. Kashoki (eds.), Language in Zambia, 243–67. London: International African Institute.

de Vaus, David A. 1986. Surveys in social research. London: George Allen and Unwin.

Downie, N. M. and R. W. Heath. 1974. Basic statistical methods, 4th edition. New York: Harper and Row.

Early, Robert. 1991. Access in Vanuatu: Optimal or maximal? In Gloria E. Kindell (ed.), Proceedings of the Summer Institute of Linguistics International Language Assessment Conference, Horsleys Green, 23–31 May 1989, 69–79. Dallas: Summer Institute of Linguistics.

Ehrlich, Howard J. 1969. Attitudes, behavior, and the intervening variables. The American Sociologist 4(1):29–34.

Fasold, Ralph. 1984. The sociolinguistics of society. Oxford: Basil Blackwell.

Fishbein, Martin and Icek Ajzen. 1975. Belief, attitude, intention, and behavior: An introduction to theory and research. Reading, Massachusetts: Addison-Wesley.

Fishman, Joshua A. 1990. What is reversing language shift (RLS) and how can it succeed? Journal of Multilingual and Multicultural Development 11:1, 2:5–36.

Fitz-Gibbon, Carol Taylor, and Lynn Lyons Morris. 1978. How to calculate statistics. Beverly Hills: Sage Publications.

Flay, Brian R. 1978. Catastrophe theory in social psychology: Some applications to attitudes and social behaviour. Behavioral Science 23:335–50.

Fowler, Floyd J., Jr. 1984. Survey research methods. Beverly Hills: Sage Publications.

Frantz, Donald G. 1970. A PL/1 program to assist the comparative linguist. Computational Linguistics 13:353–56.

Frith, James R., ed. 1980. Measuring spoken language proficiency. Washington: Georgetown University Press.

Gal, Susan. 1989. Language and political economy. Annual Review of Anthropology 18:345–67.

Giles, Howard, Richard Y. Bourhis, and Donald M. Taylor. 1977. Towards a theory of language in ethnic group relations. In Howard Giles (ed.), Language, ethnicity, and intergroup relations, 307–48. New York: Academic Press.

———, Miles Hewstone, Ellen B. Ryan, and Patricia Johnson. 1987. Research on language attitudes. In Herausgegeben von Ulrich Ammon, Norbert Dittmar and Klaus J. Mattheier (eds.), Sociolinguistics: An international handbook of the science of language and society, 585–97. Berlin: Walter de Gruyter.

Gray, Francis A. 1987. The use of language data in broadcast research. Paper presented at the Asia Area Conference on Survey Data Collecting and Interpreting, 12–13 October, 1987. Baguio City, Luzon, Philippines.

Grimes, Barbara F. 1984a. The ethnologue: Languages of the world, 10th edition. Dallas: Summer Institute of Linguistics.

———. 1984b. Second language proficiency report. Notes on Linguistics 31:26–30.

———. 1985a. Comprehension and language attitudes in relation to language choice for literature and education in preliterate societies. Journal of Multilingual and Multicultural Development 6(2):165–81.

———. 1985b. Language attitudes: Identity, distinctiveness, survival in the Vaupes. Journal of Multilingual and Multicultural Development 6(5):389–401.

———. 1986a. Evaluating bilingual proficiency in language groups for cross-cultural communication. Notes on Linguistics 33:5–27.

———. 1986b. Regional and other nonstandard dialects of major languages. Notes on Linguistics 35:19–39.

———. 1987a. Good surveys: Diagnosing vernacular need. Notes on Linguistics 38:26–30.

———. 1987b. How bilingual is bilingual? Notes on Linguistics 40:3–23.

———. 1988. Why test intelligibility? Notes on Linguistics 42:39–64.

———. 1992. Notes on oral proficiency testing. This volume.

Grimes, Joseph E. 1964. Measures of linguistic divergence. In Horace G. Lunt (ed.), Proceedings of the Ninth International Congress of Linguistics, Cambridge, Massachusetts 1962, 44–50. The Hague: Mouton.

———. 1974. Dialects as optimal communication networks. Language 50:260–69.

———. 1975. The thread of discourse. The Hague: Mouton.

———. 1985. The interpretation of relationships among Quechua dialects. In Veneeta Z. Acson and Richard L. Leed (eds.), For Gordon H. Fairbanks, 271–84. Honolulu: University of Hawaii Press.

———. 1988a (and this volume). Çorrelations between vocabulary similarity and intelligibility. Notes on Linguistics 41:19–33.

———. 1988b. Interpreting sample variation in intelligibility tests. In Thomas J. Walsh (ed.), Georgetown University Round Table on Languages and Linguistics 1988. Synchronic and diachronic approaches to linguistic variation and change, 138–46. Washington, DC: Georgetown University Press.

——— and Frederick B. Agard. 1959. Linguistic divergence in romance. Language 35:598–604.

Guilford, J. P. 1956. Fundamental statistics in psychology and education. New York: McGraw-Hill.

Hatch, Evelyn and Hossein Farhady. 1982. Research design and statistics for applied linguistics. Cambridge: Newbury House Publishers.

Hollenbach, Bruce. 1989. Mandate and "success" in SIL. Notes on Scripture in Use and Language Programs 22:25–31.

Huff, Darrell. 1954. How to lie with statistics. New York: W. W. Norton.

Hurlbut, Hope and Inga Pekkanen. 1982. Dialect comparison and intelligibility testing in the Upper Kinbatangan River area (Sabah). Philippine Journal of Linguistics 13:17–33.

Huttar, George. 1977. Identification of bilingual groups. Pre-Conference Seminars on National Involvement, 47–49. Mexico City: Summer Institute of Linguistics.

———, ed. 1982. Sociolinguistic survey conference. Notes on Linguistics Special Publication 2.

James, Heidi, Elizabeth Masland, and Sharon Rand. 1987. An investigation into bilingualism measurement methods. Dakar, Senegal: Société Internationale de Linguistique.

———. 1989. An investigation into bilingualism measurement methods. Meeting Handbook of the International Language Assessment Conference, Horsleys Green, England, 24–31 May, 1989, 381–94. Dallas: Summer Institute of Linguistics.

Kamp, Randy. 1987 (and this volume). Bilingualism testing in the Philippines. Paper presented at the 1987 Asia Area Survey Conference, Baguio City, Luzon, 13–14 October 1987.

Kashoki, Mubanga E. 1978. Between language communication in Zambia. In Sirarpi Ohannessian and Mubanga E. Kashoki (eds.), Language in Zambia, 123–43. London: International African Institute.

Kroeger, Paul R. 1986. Intelligibility patterns in Sabah and the problem of prediction. In Paul Geraghty, Lois Carrington, and S. A. Wurm (eds.), FOCAL I: Papers from the Fourth International Conference on Austronesian Linguistics. Pacific Linguistics C-93:309–39.

Labov, William. 1972. The study of language in its social context. In Pier Paolo Giglioli (ed.), Language and social context, 283–307. Harmondsworth: Penguin Books.

Ladefoged, Peter, Ruth Glick, and Clive Criper. 1968. Language in Uganda. Nairobi: Oxford University Press.

Lambert, Wallace, R. C. Hodgson, R. C. Gardner, and S. Fillenbaum. 1960. Evaluational reactions to spoken languages. Journal of Abnormal and Social Psychology 60:44–51.

Landin, David. 1989. Some factors which influence success or failure in SIL vernacular language programs. Meeting Handbook of the International Language Assessment Conference, Horsleys Green, England, 24–31 May, 1989, 148–63. Dallas: Summer Institute of Linguistics.

Langley, Russell. 1968. Practical statistics. London: Pan Books.

Loether, Herman J. and Donald G. McTavish. 1974. Descriptive statistics for sociologists. Boston: Allyn and Bacon.

Longacre, Robert E. 1989. Two hypotheses regarding text generation and analysis. Discourse Processes 12(4):413–60.

McClave, James T. and P. George Benson. 1985. Statistics for business and economics. San Francisco: Dellen Publishing Company.

McFarland, Curtis D. 1980. A lingusitic atlas of the Philippines. Study of Languages and Cultures of Asia and Africa Monograph Series 15.

McIver, John P. and Edward G. Carmines. 1981. Unidimensional scaling. Quantitative Applications in the Social Sciences 24. Sage Publications.

Mueller, John H., Karl F. Schuessler, and Herbert L. Costner. 1977. Statistical reasoning in sociology. Boston: Houghton Mifflin.

Olonan, Zenaida A. 1978. Language use in a multilingual community. PhD dissertation. Manila: University of Santo Tomas.

Phillips, John L., Jr. 1988. How to think about statistics. New York: W. H. Freeman.

Polome, Edgar C. and C. P. Hill, eds. 1980. Language in Tanzania. Survey of Language Use and Language Teaching in Eastern Africa Series. Oxford University Press.

Quakenbush, John Stephen. 1986. Language use and proficiency in a multilingual setting: A sociolinguistic survey of Agutaynen speakers in Palawan, Philippines. PhD dissertation, Georgetown University. Published in 1989 by the Linguistic Society of the Philippines.

———. 1988. Surveying language proficiency. A paper presented at the Fifth International Congress on Austronesian Linguistics, Auckland, New Zealand, January 1988. In Ray Harlow (ed.), Western Austronesian and contact languages, VICAL 2. Auckland: Linguistic Society of New Zealand, in press. (Reprinted by permission in this volume.)

Radloff, Carla. 1991. Sentence repetition testing for studies of community bilingualism. Summer Institute of Lingusitics and the University of Texas at Arlington publications in linguistics 104. Dallas.

Rensch, Calvin R. 1976. Comparative Otomanguean phonology. Indiana University Publications, Language Science Monographs 14. Bloomington, Indiana: Indiana University.

———. 1992. Sociolinguistic community profiles. This volume.

Romesburg, H. Charles. 1984. Cluster analysis for researchers. Belmont, California: Lifetime Learning Publications.

Ryan, Ellen Bouchard. 1979. Why do low-prestige language varieties persist? In Howard Giles and Robert N. St. Clair (eds.), Language and social psychology, 145–57. Baltimore: University Park Press.

Sanders, Arden. 1977. Guidelines for conducting a lexicostatistic survey in Papua New Guinea. In Richard Loving and Gary F. Simons (eds.), Language variation and survey techniques. Workpapers in Papua New Guinea Languages 21, 21–44. Ukarumpa, Papua New Guinea: Summer Institute of Linguistics.

Sankoff, David and Joseph B. Kruskal, eds. 1983. Time warps, string edits, and macromolecules: The theory and practice of sequence comparison. Reading, Massachusetts: Addison-Wesley.

Savage, Dale. 1992. A review of Walker's research on assessing attitudes and vernacular literacy acceptance. This volume.

Schooling, Stephen J. 1990. Language maintenance in Melanesia: Sociolinguistics and social networks in New Caledonia. Summer Institute of Linguistics and The University of Texas at Arlington publications in linguistics 91. Dallas.

Serpell, Robert. 1978. Comprehension of Nyanja by Lusaka school children. In Sirarpi Ohanessian and Mubanga E. Kashoki (eds.), Language in Zambia, 144–81. London: International African Institute.

Shavelson, Richard J. 1981. Statistical reasoning for the behavioral sciences. Boston: Allyn and Bacon.

Shell, Olive A., ed. 1988. Papers presented at the miniconference on vernacular literacy at Stanford, July 24–25, 1987. Notes on Literacy 54.

Simons, Gary F. 1979. Language variation and limits to communication. Ithaca, New York: Cornell University, Department of Modern Languages and Linguistics. Reissued by the Summer Institute of Linguistics, Dallas, 1983.

Stahl, James Louis. 1988. Multilingualism in Kalam Kohistan. MA thesis, University of Texas at Arlington.

Stevens, S. Smith. 1946. On the theory of scales and measurement. Science 103:667–80.

Summer Institute of Linguistics. 1987. Second language oral proficiency evaluation (SLOPE). Notes on Linguistics 40:24–54.

———. 1989. Meeting handbook of the International Language Assessment Conference, Horsleys Green, England, 24–31 May 1989. Dallas: Summer Institute of Linguistics.

Tesser, Abraham. 1980. When individual dispositions and social pressure conflict: A catastrophe. Human Relations 33:393–407.

Trudgill, Peter. 1984. Sex and covert prestige: Linguistic change in the urban dialect of Norwich. In John Baugh and Joel Sherzer (eds.), Language in use: Readings in sociolinguistics, 55–66. Englewood Cliffs, New Jersey: Prentice Hall.

Walker, Roland. 1982. Measuring language attitudes and language use. Notes on Linguistics Special Publication 2:15–25.

———. 1987. Towards a model for predicting the acceptance of vernacular literacy by minority-language groups. PhD dissertation, University of California, Los Angeles.

———. 1988. Toward a model for predicting the acceptance of vernacular literacy by minority-language groups. Notes on Literacy 54:18–45.

———. 1991. Sociolinguistic surveys for identifying priority language projects in Irian Jaya, Indonesia. In Gloria E. Kindell (ed.), Proceedings of the Summer Institute of Linguistics International Language Assessment Conference, Horsleys Green, England 23–31 May, 1989, 79–97. Dallas: Summer Institute of Linguistics.

Walton, Charles. 1977. A Philippine language tree. Paper presented at the Austronesian Symposium, University of Hawaii.

Weber, David J. and William C. Mann. 1979. Prospects for computer-assisted dialect adaptation. Notes on Linguistics Special Publication 1.

Whitely, W. H., ed. 1974. Language in Kenya. Nairobi: Oxford University Press.

Wicker, Allan W. 1969. Attitudes versus actions: The relationship of verbal and overt behavioural responses to attitude objects. Journal of Social Issues 25:41–78.

Wilson, E. Bright, Jr. 1952. An introduction to scientific research. New York: McGraw-Hill.

Wimbish, John. 1989. WORDSURV: A program for analyzing language survey word lists. Occasional Publications in Academic Computing 13. Dallas: Summer Institute of Linguistics

Woods, Anthony, Paul Fletcher, and Arthur Highes. 1986. Statistics in language studies. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Woolard, Kathryn A. 1985. Language variation and cultural hegemony: Toward an integration of sociolinguistic and social theory. American Ethnologist 12:738–48.

———. 1989. Double Talk: Bilingualism and the politics of ethnicity in Catalonia. Stanford: Stanford University Press.

Index

[blank]

Windows on Bilingualism

Publications in Linguistics 110

This volume consists mainly of a collection of papers which were presented at the Asia Area Conference of the Summer Institute of Linguistics on Survey Data Collecting and Interpreting, held in the Philippines in 1987. Several papers not presented at the Conference are also included because they reinforce and expand on the subject of survey data collecting, and treat issues which must be handled when making decisions about program planning.

The book deals with such matters as calculating lexical similarity, correlations between vocabulary similarity and intelligibility, sociolinguistic community profiles, and determining language proficiency. Also discussed are various methods of bilingualism testing, calibrating results of such testing, and the use of statistical techniques for handling survey data.

The volume should be a useful one to all those interested in survey techniques and language assessment.

ISBN: 0-88312-809-8

SUMMER INSTITUTE OF LINGUISTICS

UNIVERSITY OF TEXAS AT ARLINGTON