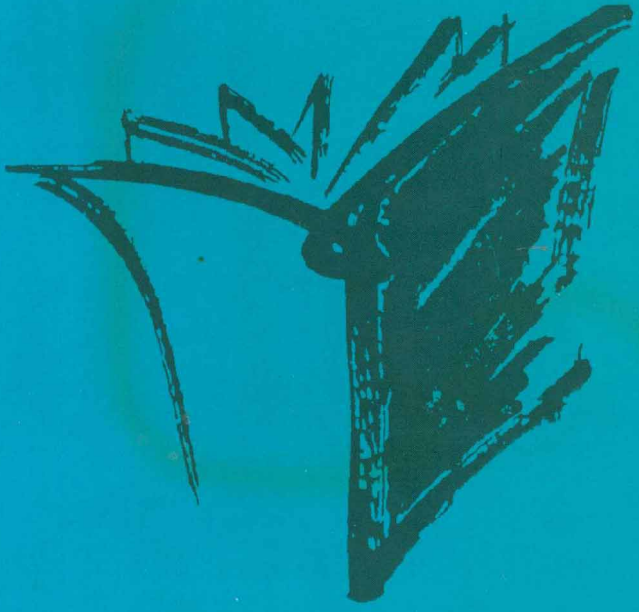
Encyclopaedia of

Language and Linguistics

Editor
Brian T. Riley



ENCYCLOPAEDIA OF LANGUAGE AND LINGUISTICS

VOLUME TWO

APPLIED LINGUISTICS

Editor
BRIAN T. RILEY

COSMO PUBLICATIONS
1999 INDIA

First Published 1999

ISBN 81-7020-904-8 (set)
81-7020-906-4 (volume 2)

Published by
MRS. RANI KAPOOR
for COSMO PUBLICATIONS Div. of
GENESIS PUBLISHING PVT. LTD.
24-B, Ansari Road,
Darya Ganj,
New Delhi-110002, INDIA

Typeset at
Cosmo Publications

Printed at
Mehra Offset Press

CONTENTS

Chapter 1
Introduction

Alan Davies

Chapter 1

It is a common criticism of applied linguistics—a criticism made by its practitioners as much as anyone—that there is no objectivity about it, that its views and hypotheses and conclusions are determined by fashion rather than by rigorous scientific procedure, that in fact there are no hard data because there is no way of establishing whether something is a result or a finding. This is a two-fold criticism. It is a theoretical criticism, denying that applied linguistics has any organized body of theory, and it is an experimental criticism, arguing that even if there is any body of theory there is no link between that and arguments as to how to proceed, i.e. how to teach and learn languages. As a result, in language teaching as in education generally, what determines change is the roundabout of fashion which seems recently to be moving back towards a modified grammar-translation method after a number of years in which such an approach to language teaching was anathema to many people. It may be that we shall always have to take account of changing fashion, simply because we have no way of finally establishing 'the best way' to learn or teach a language. Within such changes in fashion, however, there are smaller scale research operations which can be and need to be carried out and which will establish not the best way to teach language but a satisfactory set of procedures within an over-all theoretical approach. Since there is no easy way of evaluating the internal logic of a theoretical model of language, the question of what constitutes the best language-learning theory may not be a matter for experimental research at all, but a matter for philosophical argument about what kinds of aims we are interested in at any one time. Doubtless these will be influenced by the kind of within-theory experimentation we have been discussing. Certainly, our only hope of escaping from the tyranny of fashion is through submitting our guess-work to the rigour of Hypothesis and experimentation.

Experimental methods are typically the procedures by which scientific enquiry is carried out. One way of expressing this is to say that within what Kuhn has called a paradigm (Kuhn 1962) various theories are developed, expanded and tested by means of hypotheses. The purpose of a hypothesis is explanatory; but there are different kinds of explanation. Foss (1966) mentions seven of these:

causal, referring to some immediate cause
historical, referring to experiences or events in the past
purposive, referring to some purpose or goal
rule following, referring to some external rules of, for example, society
structural, referring to the components of the organism involved in the events
functional, referring to general rather than immediate causes
contingency, referring to some parallel, relatable event, hence correlational.

What all kinds of useful explanation have in common is that they demand generalization, i.e. they must be applicable to similar events and cases. Hence the need to say precisely how events and cases are judged to be similar, and what is meant by applicability.

The first of these needs is often discussed under the heading of validity, and the second under reliability. These concepts are usually discussed in connection with tests but they also have a relevance to experiments. In order to be accepted as a proper test of a hypothesis an experiment is set up in a highly formalized way by means of an experimental design which ensures that the events or subjects exposed to experiment are sufficiently like others of a similar kind for their results not to be treated as special cases. This area of experimental methods is discussed under sampling and under descriptive statistics; sampling attempts to make sure that the subjects chosen are randomly chosen, i.e., that others who are similar have an equal chance of being selected for the experiment. Descriptive statistics organizes the results, and

describes the sample in such a way as to make clear its relation to the population from whch it is drawn. Descriptive statistics then helps guarantee an experiment's validity.

The second area of experimental methods determines the applicability of the results to a more general situation, and to other cases; this is done under the heading of inferential statistics. Here the intention is to distinguish a particular finding from chance; to distinguish, in other words, the accidental from the experimental or the significant. This is the purpose of all tests of significance discussed under the heading of inferential statistics. Have we got a bizarre or accidental finding or have we got one that we can rely on ? One way of finding out is to repeat the experiment; another way, and one frequently used in the social sciences because of the sheer difficulty of repeating exactly all experimental conditions, is to make use of inferential statistics, which tell the experimenter how likely it is that he would get the same result again and again, ad infinitum. Not even a replicated experiment can tell us this, since replication can only be done a finite number of times. Inferential statistics, then, are concerned with the consistency of results, with reliability.

Experiment in the social sciences is very much more difficult than in the physical sciences. Sampling is more difficult; people do not repeat themselves, and consequently do not lend themselves to random sampling as do beans out of a bag. Laboratory conditions cannot be maintained for people. Often there is no laboratory except the real world, and experiments cannot even be set up with conditions held constant, let alone replicated. It is generally agreed, therefore, that experiment in the social sciences is weaker in scope than in the physical sciences. The question then arises of how weak is 'weaker'. Is teaching a class an experiment, or reading a book or interviewing an applicant for a job? (All these, of course, may be language learning situations.) The answer can only lie in the potential *use* made of the situations. From one point of view every task or experience is experimental. But this is not the viewpoint we wish to present at the

moment. An experiment into singing differs from the event of singing a song in that the experiment produces results that are meaningful in other situations and on other occasions. The event, on the other hand, is not generalizable in the scientific sense we are discussing here. Hence the buttressing provided by both descriptive and inferential statistics which are used to distinguish experiments from unique events, to make sure that the results that have been obtained have not occurred by chance, and that what has happened is an experiment and not an event. We are not, of course, making any exaggerated claims for experiment. We recognize that it is never possible to be completely objective, that all experiment is contaminated by the presence of the observer, not so much because he is there but because his results depend on his view of what he thinks he sees. Further, his discussion of his results, like his original conception of what he was trying to do, has a bias reflecting the experimenter's view of the world. So much is granted, but even so it still seems that some worth while experimentation is possible, and that there is no need to accept the infinite regression dilemma.

An experiment, then, is a series of controlled observations which will inevitably need the use of one or more tests. For example, an experiment into the effects of acceleration on the human body might consist of a series of observations of bodies, human and non-human, in a state of rest, or in a state of acceleration or deceleration. Each observation will be carefully arranged so as to ensure, as far as possible, that it is the right kind of body that is being observed, that it really is at rest, in a state of acceleration, etc., i.e., that the observations are valid. Furthermore, the observations will be designed in such a way that useful comparisons may be made, as between different bodies, different states (rest or acceleration), and so on. The actual making of such comparisons demands a test, or a series of tests.

A test, then, typically is the measurement of a comparison between effects or treatments. In the acceleration experiment the comparison might be between the effect of acceleration, decelera-

tion, etc., on the heart. So, taking the pulse or the blood pressure, however elaborately that is done, constitutes the test. In such cases the test instrument, the blood pressure machine, is already calibrated; hence the test is already in existence, and is widely accepted. It would be possible, though unlikely, for objections to be made to the machine, objections as to whether or not it really does measure blood pressure, objections also as to whether the markings on the scale are accurate. These would be objections to the validity of the machine, on the one hand, and to the reliability of the machine on the other. In applied linguistics such test instruments do not exist. The test therefore not only has to be used during the experiment, often it also has to be made. Hence the extra burden on the experimenter in applied linguistics, as in the social sciences generally, of constructing his own tests as well as conducting his experiments. Inevitably in these circumstances test construction is controversial in the sense that it exposes the experimenter to the criticism of nonvalidity. Normally speaking we do not question the validity of the blood pressure machine for measuring the reactions of the heart. Just as we do not question the validity of litmus paper for testing for acid; but we certainly do question the validity of a language test for testing for control of language and we do this because there is so little agreement as to what it means to 'know' a language. This is another way of saying what we have discussed above, that applied linguistics is not a theoretical discipline.

What most language tests do is to place one student in relation to a group of students—in terms of some particular language ability. The comparison for the test is therefore between students. The comparison for the experiment, on the other hand, is between effects, or treatments, or methods. Most tests compare students by establishing a rank order. A common assumption made in the construction of language tests is that language ability is normally distributed in the population, i.e. that the distribution of the ability among the population—all those who would properly take the test—is in accordance with the normal or bell-shaped curve. Whether or not this assumption about normality

is made, it is customary to use the descriptive statistics of *mean* and *standard deviation* as a simple way of summarizing the statistics for any sample who have taken the test.

What has just been said about the establishing of a rank order on the basis of an assumed ability distribution is fundamental to what Ingram calls norm-referenced tests. At the moment it is unclear whether a similar claim could be made for criterion-referenced tests (see p. 13). It is useful to distinguish four kinds of language test according to the use that is being made of the test. These four kinds are: achievement (or attainment) tests, proficiency, aptitude and diagnostic tests. The achievement test is the typical end-of-course assessment which is intended to establish whether a student has learned what he is supposed to have learned. An achievement test, therefore, samples its items from the content of the syllabus on which the course is based. A proficiency test has no known syllabus to sample; the syllabus for the proficiency test has to come from within the mind of the tester. Since the use of the proficiency test is to establish some kind of standard of, for example, English for foreign students wishing to enter British universities, it must be available to all comers. Sampling of a known syllabus is therefore not possible. If the achievement test serves to answer the question: 'How much English have you learned in this course?', the proficiency test seeks to answer the question: 'Do you know enough English to study or work in the medium of English?'

An aptitude test has even less of a known syllabus than a proficiency test since it cannot even be based on some notional 'knowledge of English'. Instead the tester has to make a construct of language aptitude which serves as a guide to the kinds of abilities to test for. The problem still remains, of course, of actually relating these abilities to real language test items.

Finally the diagnostic test is best seen as a species of non-achievement test since it is usually based on a list of errors made by students. Thus its most obvious use is after an achievement or proficiency test when it is constructed from items that cause

many errors. The diagnostic test, then, multiplies items of a similar kind in order to check on patterns of error among students.

Different kinds of tests relate to different types of validity. For example, achievement and diagnostic tests make use first of content validity (Cronbach 1960), proficiency tests establish predictive or concurrent validity, aptitude tests claim construct validity. If possible a second validity type will also be established, e.g. content validity for proficiency tests and predictive validity for aptitude tests.

Language tests are often spoken of as though there were an abundant supply available. The number of published tests, however, is limited. The reason for this is that there is little point in publishing a test that is not standardized, i.e. provided with population statistics, or, more correctly, parameters. From one point of view, a test only really becomes a test when it is standardized, when it has been tried out on a random sample and thus shown to produce a normal distribution relevant to language ability. Most so-called tests which are not yet standardized are really trial tests, based on various unsubstantiated ideas and guesses. Useful lists of tests available are given in Buros (1972), CILT (1973) and Davies (1968).

Use of a standardized test in a particular situation does not necessarily provide the rank order we have spoken of, nor does it often do so. It is quite possible to get no spread or very little spread of scores. The reason for this is quite simple, namely that the sample being tested is not a random one. Ad hoc classroom tests devised by teachers during the course of their work often illustrate this lack of spread of scores. It is probably best to regard ad hoc tests as trial tests rather than as finished products. Only standardized tests can properly be called tests. Ad hoc tests rarely achieve this finish because by their nature they are made for a specific and temporary need. They can. be more objective than the usual subjective examinations but they must not be taken as final judgements because they do not describe a population.

As we have pointed out above, language experiments need

language tests as a means of checking the success or failure of the experiment. An experiment remains indeterminate unless it is put to the test. This is not to say the test is the experiment. Indeed, as we have seen, in normal science the test often exists already and can be used for different kinds of experiment. In this case the originality goes into the design and creation of the experiment itself. We have also seen that in applied linguistics it is often necessary to construct a test as well as to design the experiment. However from the point of view of the experiment the test remains an adjunct, a necessary one but subordinate to the experiment. Reports on applied linguistics experiments will refer to the tests used during the course of the experiment, and will state whether or not they have been specially created for the experiment. Looking at the experiment-test relation from the point of view of the test, it is possible to see a test under construction as in itself a mini-experiment. The tester's experimental hypothesis then is that the tasks he presents as test items will produce the rank order we have mentioned. Testing, retesting and analysing of the test results become the carrying out of the experiment and the final validation of the test becomes, as it were, the test of the test-experiment. Experiments then need and use tests, while test construction is a form of experiment. In some experiments tests need to be constructed (e.g. Smith and Berger 1968), while in some test construction, further tests need to be written in order to validate the main test under construction (e.g. Pimsleur 1963).

Test construction and test use are, therefore, quite distinct. The first relates to the experimental aspect of testing, the second to the instrumental aspect. One of the abuses of tests under construction—the ideas, guesses, trials we have referred to—is that they are often used as if they were already tests in use and employed in order to validate or assess experiments. The misuse often goes like this. A test is written and trials made. At this stage there are certain results available, though quite simple and yet fundamental things may he wrong, such as no random sampling; these results are then used to describe certain effects as if the test