ulletin of Hokuriku University
Vol. 13 (1990)
pp. 67～84

An Analysis of Testing and Examinations

John D. Dennis

1.0 Introduction

Testing and examinations form a common, important part of the lives of students. Certain crucial tests, such as entrance examinations, can determine the course of a student's life. Because of this it is especially important to carefully study and prepare the tests we give to our students. This paper will attempt to characterize, analyze and classify both the various kinds of tests and the issues involved in their use in order that we can better understand the activity we are engaged in.

Testing is often looked upon as a necessary evil, and the student taking a test (the examinee) is often neglected and depersonalized by the testing process. In what follows, although the perceptions, feelings, and attitudes of students are not discussed directly, it should be remembered that neither students nor the teaching of language must be lost to the pursuit of testing. Though often stereotyped for their orientation towards data or numbers, researchers in the field of testing often do try to consider students on an individual, personal basis. Testing via an interview in a relaxed setting is one currently popular way to do this. Also, thorough, frequent and rigorous checks of the tests we use certainly will produce better tests than doing no investigation at all, and of course students can be treated more fairly, equally, and humanely by administering tests that are correctly suited to their skills. Thus, proper investigation of a test can increase its utility and improve how well it works. Correct placement of students can in many cases help a program or school operate more smoothly—by accurately evaluating student proficiency we can better control the type and quality of students in our classrooms. Critically examining tests is thus a kind of "quality control" that benefits students, schools and even teachers in the classroom.

As with testing in general, when investigating language tests there are three basic points to consider. Shohamy (1988 : 165) gives two of them by stating that "Language testing is concerned with the measurement of language: Language is the trait, and how we go about measuring it is the method." Peterson (1989 : 95) adds another step and states that:

* Faculty of Foreign Languages
外国語学部

1. We have to know what we want to measure,
2. we have to decide how to measure it, and
3. finally, we have to check to see that we have measured what we intended to measure.

While this deceptively simple scheme does summarize the requirements of good testing, there are a large number of questions that need to be answered along the way. Perhaps it would be best to begin by characterizing the general direction of testing at the present time.

Tests are becoming more efficient and, as mentioned, are being designed so that they are more adaptable to an individual's abilities. In designing language tests researchers are often attempting to test more than one skill or ability at once—i. e., to provide a test that gives an integrated evaluation of an examinee's ability. A second type of integration in which different skills support one another has also been developed. For example, listening comprehension tasks on a test can "feed" and "be fed by" other skills (Douglas, 1988 : 255-6). Tests that challenge examinees to use different strategies (linguistic, social, rhetorical, etc.) have become more common, tests have been "contextualized" by careful design and close attention to the social knowledge and backgrounds of the examinees, and finally, tests have been designed that require examinees to apply knowledge that they gained through one skill (reading, listening, etc.) to a variety of tasks that involve other aspects of language proficiency. These features reflect and summarize many of the recent trends in test development. Before moving to a discussion of test types and their characteristics, it is necessary to deal with certain global issues in the field of testing.

2.0 Global Aspects of Testing

There are certain questions, concepts, and problems that have an impact on most all tests, no matter what their type or style. The first question deals with what is being tested. Obviously, language teachers will want to test language—but just what aspects of language will they want to test? Will a test focus exclusively on pronunciation? Will it try to examine linguistic competence and exclude all types of "world knowledge"? At what level of proficiency can we expect students to display a grasp of the social and cultural knowledge possessed by the native speakers of a language? What about communicative ability? Should a test evaluate how well a student can function in specific situations and exclude consideration of grammatical accuracy? The decision about what to test can be a difficult one. Part of the reason for this is that the nature of the skills and abilities that allow one to produce language remains unclear. Also, the skills and abilities involved seem to develop in different ways and at different rates in different people. This diversity and uncertainty undermines the ability of the test designer to

correctly assess language ability.

The second global aspect of testing concerns whether or not a test is designed around a unifying theme or context, the theory being that a test designed in this way will provide a more motivating and more satisfying experience for examinees (Douglas, 1988 : 251). Some researchers argue that thematic unity can have a significant impact on the results of a test (Douglas, 1988 : 256). Typically, individual items on a test bear little relationship to preceding or following items. However, tests which have been designed around a central theme (a situation, a topic, or a place) produce different results than those which have not. Thus, the question for test designers is: Should an organizing theme be used for all or part of a test? If so, how can a theme be chosen that will not in itself skew the test results? Since certain topics are more interesting and appealing to various sections of any test population can a test be designed without giving an advantage to some special group of examinees?

The third global aspect of testing concerns the effect of background knowledge on test results. Some tests make common and frequent use of current events for test material. Others rely on certain genres of literature for content. How can test designers insure that by using these materials they are not prejudicing the results of their examination? Choosing a common topic or social situation on which to base a test is especially difficult where student populations come from a variety of cultural backgrounds. Related to the issue of background knowledge is the issue of how students have studied the target language. Some student groups have been thoroughly and repeatedly exposed to many types of test questions while other have not, and since a knowledge of testing techniques can be thought of as a kind of specialized background knowledge, won't those students who have a knowledge of testing score better than those who don't? How can a test be designed to circumvent this problem?

Other global problems concern dilemmas that test designers have encountered in practical situations. Some researchers have noted that while learners with similar competencies can get different scores on the same test, at other times learners with similar scores have gotten those scores as a result of differing competencies. Even the effectiveness of a good test can vary from one group of students to another, and it has been demonstrated that students may perform quite differently on different types of tests (Brown, 1989 : 66). Lantolf and Frawley (1988 : 183) claim that test tasks can never be accepted as natural by the examinee, that tests will always be artificial, and that tests can only mimic a reduced version of the world rather than showing it as it actually is. They state that the task of the test itself often overshadows, overpowers, and detracts from the task of demonstrating the abilities and skills that make up language proficiency. These authors state that they:

> ... cannot assume, therefore, that if students are successful in a decontextual-

> ized setting [i. e., the test] that they will also be successful in a different setting, ... The only environment we can be certain of is that established by the test situation itself and nothing more. (Lantolf and Frawley, 1988 : 190)

Lantolf and Frawley offer an extreme view of the perspective that an examinee may be more or less permanently disabled by the test situation from showing his or her true skills and abilities (1988 : 191). However, this extreme view must be put aside since it virtually prevents us from administering any type of test or using the results for any practical purpose. While a test may be a reduced (and thus inaccurate) model of the world, the general consensus is that it can still be used by students to demonstrate their abilities. Additionally, this type of reduced model can help students study and learn any number of language skills, and can help them gain confidence in the kinds of skills and abilities that they will eventually be using in the real world. Next, although the preceding issues and concepts have a broad, global impact on testing, there are two remaining areas with a similar global significance that must be discussed. The first of these is the concept of validity; the second is the concept of reliability.

2.1 Validity

Validity is a complex topic and one which test designers frequently fail to agree on. While a thorough treatment of validity is beyond the scope of this paper, this concept does provide a set of useful criteria for test analysis. Parenthetically, the confusion surrounding this concept may be the result of mistaking the particular type of validity that another person is talking about. Beginning with a general definition, Peterson (1989 : 95) suggests that "validity is related to the purpose for using a test, not to the test itself". Shohamy (1988 : 173) says that "it is by examining the reliability and validity of ... tests that we can assure that scores provide accurate and valid indications of ... language." There are five types of validity that will be described briefly in turn.

The first type of validity is *construct validity*. When determining construct validity for a test, we need to ask how well that test reflects current theories of testing and their requirements (Williams, 1990 : 50). Tests should incorporate the latest, generally accepted research in the fields of testing and language teaching.

The second type of validity is *face validity*. Face validity asks whether or not a test looks like the test that it is supposed to be (Williams, 1990 : 55). For example, does a test that purports to measure listening comprehension look like a test of listening to the students who are taking it? - -Or does it look like something else? Maintaining face validity is something like maintaining self-confidence: the test needs to support the illusion that it's really measuring what it claims to. Carter and Long (1990 : 220) suggest that a test should appear to be the reasonable outcome of any preceding classroom activity. As for university level entrance examinations, they should fall within the

expectations of the high school students that will be taking them. Face validity is a relatively weak form of validity and it is often claimed that it cannot act as proof of validity by itself (Shohamy, 1988 : 167).

The third type of validity is *content validity*. Here, the test should reflect the materials and content of the courses or curriculum that students will be studying (Williams, 1990 : 56). It is important to stress that testing and curriculum need to be coordinated. Students need to be evaluated based on their suitability for the program which they wish to enter. It makes little sense to evaluate students on the basis of one set of criteria and then to place them in a program that supports a different set of values.

The fourth type of validity is *concurrent validity*. This type of validity is a statistical concept—test designers must assess the extent to which test components or different tests which purport to measure the same skill actually correlate in statistical terms, and the extent to which the test correlates with other tests (Williams, 1990 : 56). Due to the number of calculations involved this type of validity check is possible only with machine-processed data and a computer. True, it can be done by hand, but the cost in time and effort would be prohibitive.

The fifth and final type of validity is *predictive validity*. When used with a proficiency test, this type of validity checks the extent to which scores on that test correlate well with the subsequent success of the examinees. To check the predictive validity of a university entrance examination a researcher would try to correlate scores on the exam with the later successes or failures of the students concerned. Students with high entrance exam scores should be the most successful students in the program that follows. Students with low scores on the test should be the weakest or need the most help in the program that follows. If not, then the entrance exam has poor predictive validity and the causes of this should be investigated. These five types of validity are suggestive of the kinds of questions and statistical analyses that can help improve tests and examinations. From here, we must move on to reliability, the next topic of discussion.

2.2 Reliability

"Reliability is related to the interpretation of scores, ... not to the test itself" (Peterson, 1989 : 95). Tests of reliability are really tests of consistency—they answer the question: Do you consistently get the same results? In the area of oral proficiency testing, the people who give tests are called "raters" since they give ratings, or evaluations, to the examineses that they interview. Because a certain variation in opinion is to be expected when different raters interview the same examinee, test researchers have devised ways to check on "inter-rater" reliability to measure the extent to which different raters give the same evaluation to an examinee. Also, the concept of "intra-rater" reliability is used to measure the extent to which the *same* rater consistently evaluates the same students in the same way at different times. In both of these cases

the question here is really: Are the test results that we are getting consistent? As such, parallel questions can be asked of university entrance exams.

Try to conceive of a kind of "intra-test" reliability, where we could check on the extent to which the same test consistently evaluates the same students in the same way at different times. As for an entrance examination, it wouldn't be too difficult to administer the same test *a second time* to matriculating first year students. Although all the candidates who took the regular entrance examination would not be repeating the test, a reasonable sample (the matriculating class) would be. The retest could take place approximately two months after taking the regular entrance exam—not enough time for a significant improvement in language skills to take place. These students would retake the same test (probably without warning) and the results of their first and second attempts on the same test would be statistically compared to see how reliable the exam was. Thus, the same group of students would be taking the same test two times. Generally, if the scores on these two attempts *converge* (if the results are similar), then there is a good chance that the exam is really testing the language abilities and skills that it purports to. However, if there is little convergence or correlation the there is a good chance that the test is not functioning they way it is supposed to: it is not a reliable test for judging candidates' language abilities.

Next, there is anther type of test reliability that can be investigated. Again, "inter-rater" reliability checks on the extent to which different raters give the same evaluation to an examinee. This kind of reliability test can also be used with a university entrance exam. Instead of the *same* test being administered to the same group of students, a different but similar test is administered to the same group of students. How can this be done? Universities typically create and administer a new entrance examination each year. For any group of matriculating students there is a set of related entrance examination scores. Perhaps just before beginning their classes, this group of newly admitted students could retake the entrance exam from the previous year. Thus, the same group of students would have taken rather similar tests, designed by the same school, and used for the same purpose.

These two tests would yield two sets of scores for the same group of students —scores which could then be compared and correlated statistically to check whether or not the two tests were reliably measuring the same thing. Again, if the results of the two tests converge then there is a good chance that both examinations are measuring the language abilities they are designed to measure. However, if there is little convergence or correlation then there is a good chance that one (or both) of the tests is not functioning they way it is supposed to: it is not a reliable test for judging candidates' language abilities. This type of reliability is known as parallel-forms reliability, and as Shohamy states:

> For estimating parallel-forms reliability, there is a need to compare different versions of the same interaction, as it is varied by a number of contextual variables. The extent to which the scores of the two or more versions correlate is an indication of this type of reliability. (1988 : 173-4)

Thus, different types of reliability indicate the degree to which tests accurately and consistently measure what they are designed to measure. University entrance examinations exemplify the kind of test where it is especially important to achieve a high degree of reliability, since crucial decisions are based on the results of these tests. Finally, although it is not essential to investigate all types of validity and reliability, some basic checks should be a normal part of the design and administration of all tests (Shohamy, 1988 : 170).

2.3 Precautions

Before moving on to a discussion of different types of tests and their characteristics it is necessary to put forth some cautionary statements concerning language skills and abilities, testing, and the scores, data, and statistical results that an investigation into testing will inevitably produce. Though perhaps it seems obvious when stated directly, it must be remembered that just because tests produce numerical results this does not mean that the skill or ability being measured by the test is a linear, metric one (Lantolf and Frawley, 1988 : 185). This is especially the case with language, which consists of multiple skills and abilities, each of which develops in different ways and at different rates.

Second, it is important not to confuse the *method of testing* with the skill being tested. As Bachman (1988 : 15) states "...any measure of... ability must clearly distinguish the ability to be measured from the methods or procedures used to elicit evidence of this ability." That is, one must not confound the abilities to be measured with the elicitation procedures designed to produce them. Typically, one assumes that a particular language skill is best measured by using that same skill as a vehicle for the test. Thus, one would assume that the testing of listening skills would be best accomplished by using an elicitation procedure that involves listening, or that a test that purports to measure writing ability would naturally require the examinees to produce a writing sample. While using the same skill for elicitation as the one being measured does provide for face validity, it does not necessarily follow that such a test will "naturally" provide an accurate measurement. Such tests should continue to be the subject of ongoing validity and reliability checks.

Third, test designers should abide by the comments of Lantolf and Frawley (1988 : 185) when they caution investigators not to lose sight of the object of their inquiry by focusing their attention too exclusively on the tools used to measure it. For language

teachers this means that the skills and abilities that comprise language should have a greater importance and priority than either the tests used to measure those skills or the kinds of analysis that those tests may be subject to. Having finished with these precautions, I will now move on to a discussion of types of tests.

3.0 Types of Tests

Researchers in the field of testing commonly agree that there are two basic types of tests. One type is called a *norm-referenced test* (NRT) and the other is called a *criterion-referenced test* (CRT). Norm-referenced tests will be characterized first.

Generally, NRTs are designed to measure overall language skills and abilities. The abilities measured could be broadly defined as overall proficiency or could be somewhat more restricted—e. g., a vocabulary or a listening comprechension test. For NRTs there is only one test and only one score for each individual who took the test. The scores produced by NRTs are interpreted relative to each other. Individual scores are frequently evaluated by their distribution around a statistical norm or mean. Indeed, that is the purpose of a NRT—to spread students along a continuum of scores so that those examinees with relatively little ability are placed at one end of the scale, while those with relatively high ability are placed at the other. Usually, the bulk of the scores is found in the middle of the scale, clustered around an average or mean score. Finally, prior to the test examinees are not given any information about the specific content of the test, although they may know something about its structure—the type of questions, for example (Brown, 1989 : 67-8).

On the other hand, CRTs are produced to measure specific instructional objectives. These objectives are often unique to a particular course or program and are dervied from the instructional goals that serve as a basis for the curriculum. Because of this relationship it is important for both teachers and students to know exactly what is expected of them. CRTs often employ a pretest/posttest scenario, but the pretest is often left undone and the students are assumed to have little or no knowledge of the material that will be taught and tested. Scores produced by CRTs are claimed to be "objective" and these results are considered to be absolute, i. e., they are not interpreted with reference to the scores of other students. Each individual score is thought to be meaningful in and of itself. Also, when placed on a scale the distribution of CRT scores is not the "normal" distribution of NRT scores. Instead, if all the students learn everything that is taught they will all receive the same high (or perfect) score on the test. Thus, CRTs measure the extent to which students have developed knowledge of a specific ability or skill, which has usually been specified in the goals or objectives for their program. In contrast to NRTs, above, for a CRT students usually have a very clear idea of the types of questions, tasks, and content to expect on the test (Brown, 1989 : 68).

The basic differences between NRTs and CRTs can be summarized with reference

to the following five points.

1. What is measured
2. How the scores are interpreted
3. The distribution of scores
4. The purpose of testing
5. Student awareness of test questions and content

Criterion-referenced tests appear to be the undeclared (and unattainable) ideal for the world of testing. The idealized version of a CRT is indeed objective, but it seems that there is always some way in which practical reality falls short of the ideal.

First, the establishment of a minimum criterion level is usually arbitrary. While it remains true that scores may be absolute, some sort of interpretative, arbitrary judgement is necessary to determine what level of performance is adequate. Should it be 60 points or 70 that is deemed to be a passing score? Why one and not the other? Second, Douglas claims that:

> There are so many ways, structural and strategic, to get something done with language that it is currently beyond our ability to establish a domain of tasks that would add up to a criterion. Only a very narrowly focused language-for-specific-purposes context might allow for a reasonable criterion-referencing. (1988 : 251)

In spite of this claim, which disallows CRTs because they cannot meet true standards of objectivity, it is still quite valuable to know as much as possible about what is being tested. Shohamy states that "In constructing language tests, it is essential to have a defined curriculum or set body of knowledge from which testers determine what to test" (1988 : 165).

4.0 Proficiency Tests and Achievement Tests

"Proficiency test" and "achievement test" are the popular terms for the two previously discussed types of tests. A proficiency test is a kind of norm-referenced test, while an achievement test is a kind of criterion-referenced test. As they are typically administered university entrance exams are norm-referenced proficiency tests, not achievement tests. University entrance exams fit the requirements for a norm-referenced test very well: 1) they are a broad measure of an examinee's skills, 2) scores are interpreted with reference to the scores of other examinees, 3) the graders of these tests hope for a normal distribution of scores, 4) the purpose of testing is to spread the candidates along a scale based on their scores, and 5) the examinees are familiar with the format or structure of the test without knowing the specific content that will be tested.

Because university entrance exams (proficiency tests) are so commonly accepted by examinees, institutions, and society they frequently escape critical examination of the concept of proficiency that they are based on. Since no two examinees are alike and because most of these tests produce only a single number or score as a measure, it can be claimed that proficiency tests have a homogenizing effect on any subcomponents of proficiency skill. The proficiency of a speaker can never be characterized in any absolute sense and it is not a concept "that can be formalized in terms of a taxonomy of items, no matter how long or genuine that taxonomy may be" (Lantolf and Frawley, 1988 : 189-90). This may be the reason that proficiency has been defined in more than one way by different researchers (Lantolf and Frawley, 1988 : 186-90). Thus the major weakness of proficiency tests is that while they purport to judge proficiency they fail to give any account of what proficiency is, where it comes from, and how it's determined. It would seem that a humanly devised concept (proficiency) is determining what the world should be like (the effect of the test on the matriculation of students) rather than the world (as empirical and objective as it is) determining what the concept of proficiency should be.

It is also important to note that while university entrance exams seem to easily satisfy the requirements of a norm-referenced proficiency test, there are ways in which the same test can be reconceived as a criterion-referenced achievement test. The first contrast between NRTs and CRTs concerns what is measured. Rather than considering an entrance exam to measure some general notion of proficiency, it could alternatively be thought of as measuring the common aspects of a set of relatively well-defined English curricula. This kind of elementary or basic proficiency might be more easily specified than proficiency for other, higher levels of skill. The second contrast concerns how scores are interpreted. Scores might be interpreted to characterize how well students have done at meeting those basic objectives rather than relative to the scores of other examinees.

The third contrast concerns score distribution. Not much reinterpretation can be done here—the scores present themselves as they are. If the distribution is unusual this could be reflecting actual differences in achievement of the students concerned. Perhaps an unusual distribution could be correlated with the high school the examinees attended, any extracurricular instruction they might have participated in, or some other factor (an unusual distribution could even be a by-product of the test itself). The fourth contrast, concerning the purpose of testing, is relatively less important than the others. Whether it's proficiency or achievement that's being measured there will certainly be a range of scores—some students will inevitably do better than others due to a variety of external factors. Either a proficiency or an achievement test can be designed to elicit a broad range of results, and too little or too much discrimination between student levels of proficiency can be a problem for either type of test.

The fifth contrast concerns the type of knowledge that examinees will have of a test before they take it. To the extent that university entrance exams are testing a basic proficiency, and to the extent that these exams remain unchanged in design and content from year to year, students (or their tutors) will be able to predict more and more accurately both the kind and type of questions and the content that they will contain. Finally, regardless of which type of test an entrance exam is considered to be, the scores are reported as numbers. Again, it is best not to lose sight of the reason and purpose for testing by focusing too exclusively on scores.

5.0 Test Construction

In section 3.0, above, two basic types of tests were characterized. While the temptation is strong to label this section "Types of Tests (Part 2)" it is perhaps better to call it "Test Construction," since the issues dealt with here are concerned with the actual production and design of a test. Tests involve types of tasks, the testing of all or part of an examinee's knowledge of a language, and various types of skills and knowledge. We will now examine these different things to get a feel for various types of language tests.

When designing a test one of the first decisions concerns the type of *tasks* that will be used in the test. Douglas (1988 : 246) lists five types of tasks that can be used on language tests. The first task is called "listen [or read]-and-give-the-right-answer." This type of question is common on proficiency examinations such as TOEFL. The answer format can be multiple choice, cloze, fill-in-the-blanks, or perhaps a longer answer. The second type of task is called a "reduced-redundancy task." This type of task tests listening comprehension when some features are missing or when they have been masked by some disturbance or environmental noise. The third type of task is called a "repetition-imitation task." Here, competence is displayed in the form of short term memory. The theory is that if students know something they will remember it more easily than if they don't. The fourth type of task is called an "interaction task." A task of this type might require students to produce or select the appropriate responses in conversations. The fifth and last type of task is called a "media-transfer task." Students are asked to respond to video or pictorial input on this type of task. The Test of English for International Communication (TOEIC) uses this type of task to test basic listening skills.

Next, a decision must be made about what aspect of the examinee's knowledge is going to be tested. Is the test going to check their language ability in a global, holistic fashion, or is it going to assess different skills separately? Bachman says that:

> The evidence from language testing research is generally consistent with the hypothesis that language proficiency consists of several distinct abilities that

> are either related to each other or that are related to a higher order general ability. (1988 : 155)

Another researcher, Elana Shohamy, states that:

> Holistic scales define global knowledge, whereas analytic scales focus on specific aspects such as grammar, fluency, strategies, sociolinguistic factors and pronunciation. (1988 : 173)

Neither of these statements offers any evidence for or against testing holistically or testing various skills independently. This decision is one that would best be made with reference to the kind of student that a program or school would most like to attract. Or, it may be best to consider the type of curriculum a school is offering. Does it give a higher priority to one or another of the language skills (reading, writing, speaking, or listening)? Or does it give an equal value to these skills when considering students for admission? These questions and others like them need to be considered for a test to adequately reflect the goals and objectives embodied in a curriculum.

If a decision is made to test the sub-skills of language proficiency it still has to be determined at what level those sub-skills will be tested. Is it listening comprehension that is going to be tested? Or should it be some particular aspect of listening? Douglas cites research that identifies nine different sub-skills of listening comprehension (1988 : 246). Thus it may not be enough to simply pick a skill and begin developing a test for it. Given that there are nine sub-skills to listening comprehension it still must be decided whether to test for all or part of them. Finally, Matthews cautions that "It is... illogical to allocate equal marks for the various sub-skills as if the relationship between them was one of simple addition" (1990 : 118).

Third, is a test going to be designed to emphasize functional abilities and skills (showing an awareness of social status, formality, and context), grammatical and linguistic knowledge, or content (knowledge of a certain set of topics)? Although to a certain extent this decision will be based on the competency level of the students to be tested, these are the types of issues that have to be addressed when designing an examination.

Finally, when considering types of tests sometimes the discussion turns to the various types of questions and answers that can be included on them. Although it is common to hear a test referred to as "a multiple choice test" or as "an objective test" this use of popular terminology is misleading, and it is a way of describing tests which is rare in the literature of this field. A multiple choice test could be either a norm-referenced proficiency test or a criterion-referenced achievement test. As for the term "objective" it is usually used in connection with criterion-referenced tests since their scores are usually interpreted objectively—with reference to an absolute scale rather

than being interpreted in relation to the scores of other students.

6.0 University Entrance Examinations

Many of the ideas and issues presented up to this point have a direct impact on university entrance examinations. I would like to review some of the ideas discussed so for with special attention to these exams. Too often questions of test design and the nature of the relationship of a test to a school's curriculum go unnoticed in the process of test development. It cannot help but be beneficial for a school to clearly ask these questions and to define the type of test it would like to use *before* the process of test development begins. Based on their knowledge of present or future curriculum design, should test designers give priority to linguistic ability, to functional or communicative skills, or to knowledge of content, certain topics, or cultural information?

Second, since a test is generally believed to be better when it is based on or related to the curriculum of the school concerned, is it the case that university entrance exams are connected in this way to their curricula? Malu refers directly to this situation in her discussion of the effectiveness of different types of tests by stating that:

> ...curent thinking concerning tests and testing appears to be that tests and testing procedures that clearly relate to our classroom assignments, curricula, and pedagogy may be very effective and efficient tools for measurement. (1989 : 209)

If a school is indeed concerned about its testing program one of the best ways to begin the process of improvement would be to establish a connection between its classroom assignments, curriculum, and instructional methods on the one hand and its testing program on the other. A faculty survey could provide a list of typical assignments, course objctives, and any minimum requirements instructors may be using. By designing a test with this information in mind a school stands an excellent chance of selecting better-suited students for matriculation.

Additionally, it is good to have a test that also functions as a teaching tool. If instructors can use entrance exam material in their classes for pedagogical purposes this is a good sign—it shows a certain degree of test-curriculum integration. Actually, a good way to evaluate a test that has been used would be to ask the faculty whether or not it possible to include material from that test in their courses. If the answer is "yes" then there is at least minimal convergence between the testing program and curriculum. If the answer is "no" then this is a sign of a possible mismatch and the suitability of the test for the curriculum that follows should be questioned.

6.1 Technical Analysis of Entrance Examinations

Technical analysis of an examination can be under taken using a variety of statistical tests. There are a number of tests that can be used, and all of them require that the final results of an exam (whether it is a proficiency or achievement test) be represented by a number. Some statistical tests even require that individual items on an examination (i. e., individual questions) be graded as either "correct" or "incorrect" rather than being given points on a scale depending on how well that problem was solved. While there are certainly disadvantages to representing language skills with numbers, it is also important to consider the valuable information that can be produced by the statistical analysis of a test or series of tests. Also, it quickly becomes difficult and impractical to evaluate a growing number of exam candidates in any other way.

For these reasons it is often necessary to introduce a certain amount of mechanization into the examination process. Frequently, this time-and labor-saving mechanization takes the for of multiple choice, computer-graded answer sheets. While some instructors bemoan the introductiion of such "dehumanizing" and "impersonal" test methods, this solution to the problem does allow those persons who are responsible for testing to spend their time on test design and production, and later on an analysis of the test, rather than on grading. Finally, while the machine-grading of tests is usually handled by clerical staff, the work of hand-grading tests is usually handled by the faculty.

Thus, there are two reasons which support the reporting of results in the form of numbers. First, this allows a school to investigate the validity and reliability of a test by using various statistical tools to analyze it, thereby gaining valuable information about its strong and weak points. It should be noted that this type of analysis is one of the few sources of information that can help improve a test. (Although it is possible to have an evaluative brainstorming session after hand-grading a large number of exams, these sessions are certainly more subjective and impressionistic than statistical analysis. Fatigue of the grading staff is also a factor.) Second, limited time and/or resources often prevent a test from being designed and administered in any other manner. Let us now examine some of the statistical tools that can be used to analyze a test.

6.2 Statistical Procedures

Most analyses of data deal with different kinds of relationships, and the comparison and contrast of two groups of data is one of the simplest relationships to investigate. The simplest comparison of two (or more) groups is a comparison of their centers (defined as either an average or mean). This kind of test is called a *t-test* or *t-statistic*. In successive years students taking an entrance exam are similar in most respects, except that their means might be different. Because of this similarity it can be assumed

that the population variances of any two groups of exam candidates are equal and statistical procedures that use pooled variance estimates can be employed. Given that this assumption is correct the corresponding tests for differences between the means will be more powerful and will produce more significant information about the two groups. If cannot be assumed that the underlying variances for each group of students are equal then a *two sample t-test* must be used. Since two estimates of variance are used the results of the test will show less information about the two groups, i. e., it will be less powerful. The trade-off is between the assumptions one makes and the kind of information that is gained. A stronger initial assumption yields more information, but as with all assumptions one runs the risk of being incorrect. A more conservative initial assumption about the variance of the two groups provides some insurance against the risk of being incorrect but yields less significant information.

There is anther type of t-test that is called a *paired-t statistic*. This statistical tool can be used when each case in the first group corresponds naturally to a case in the second group. A common situation that produces naturally paired cases is when the same individuals make two judgements. Also, when data are naturally paired, a paired-t test will yield more powerful results. This tool could be used with entrance exams in the following ways. First, it could be used to evaluate separate sections of a single exam. If there are two (or more) sections on an entrance exam these sections could be graded separately and then compared as though they were paired judgements made by the same individual.

If a school were interested in re-administering the same entrance exam (to check its reliability) a paired-t test could be used for this type of analysis. For this type of study, any group of matriculating students (students who have just taken the regular entrance exam) could retake the same exam, perhaps early in April just as their classes begin. As described above in the section on reliability, the higher the correlation of these paired scores, the more confidence a school could have in the reliability (the quality and consistency) of its entrance exam. A slight variation on this strategy could also be used to compare separate entrance examinations. Universities usually create a new entrance examination every year. If a school wanted to compare two entrance examinations they could re-administer an entrance exam from a previous year and then compare the results of that exam with the results of the regular entrance exam that had just been used, still using a paired-t procedure.

A final variation on this strategy could be used to compare the results of an entrance examination with a proficiency test such as the "Eiken," the TOEIC, or perhaps a less difficult exam such as the SLEP test ("SLEP" stands for "Secondary Level English Proficiency"). This last test is used to evaluate the English skills of high school students, rather than businessmen or applicants to an American university. Although the content of an entrance examination and one of these publicly offered proficiency

tests would diverge more than a pair of successive entrance examinations from the same university, it would still be an interesting and enlightening comparison to make. Such a comparison would perhaps shed some light on the differing notions of English proficiency that were embodied in various tests.

Another more powerful statistical procedure that could be used to study entrance examinations is called *Pearson Product-Moment Correlation.* This is the statistic that is commonly known as correlation. This test is a measure of the extent to which each person with a high (or low) score on one test will tend to get a correspondingly high (or low) score on the other; thus, it is very useful for comparing different proficiency tests. Correlation measures linear association so it is important to note that only variables which have such a linear association can succenssfully analyzed using this tool. If an analysis using this statistic does not show a correlation it is still possible that the variables are closely related but that their relationship is a non-linear one. Perhaps it is also important to note that correlation has no units—an advantage if the original units of data contain sensitive information which should not be revealed (such as the actual examination scores of students).

A second type of correlation test is called *Spearman Rank Correlation.* When two variables are not linearly related but there is a consistent trend between them, the Spearman Rank Correlation will correlate the respective *ranks* of the two variables. Though less powerful than the Pearson statistic, Spearman's rho can analyze variables that are not linearly related.

Either the Pearson Product-Moment Correlation or the Spearman Rank Correlation can be used to analyze the results of the multiple testing strategies described above. These correlation procedures can compare: 1) sections of the same test, 2) different administrations of the same test, or 3) different tests administered to the same population (the same group of students). Comparisons based on any of these three procedures can be used in an analysis of university entrance examinations.

Apart from the types of statistical procedures that have just been described, there are some techniques for measuring the performance of single items (questions) on a single examination. The first such technique is called *item facility.* Item facility (IF) is a single number that represents the proportion of examinees who answered a given item correctly. It is written as a decimal fraction, so that while an item facility of .02 means that 2% of the examinees answered a given question correctly, an item facility of .95 means that 95% have answered correctly. This kind of analysis can show how well single questions are functioning on an entrance examination. A high item facility shows that a question is too easy, that virtually all examiness are answering the question correctly. A low item facility shows that few examinees are answering a question correctly.

Since questions with high or low item facility numbers are not distinguishing well

between examinees, they are correspondingly less useful for spreading examinee scores along a scale. And since the purpose of a proficiency exam (i. e., a university entrance exam) is to discriminate between various levels of language ability, questions with either a very high or very low item facility are not useful for the purposes of the examination. These questions should be marked for revision or removal from any subsequent test form, while questions with an item facility close to .50 or .60 should be retained.

There is a second technique that is related to item facility which is known as *item discrimination*. Item discrimination (ID) measures the extent to which a question separates students with more ability from those with less, and it is calculated in the following manner. First, the scores of all examinees are rank-ordered and the scores of students who have scored in the upper third and the lower third of all examinees are separated into groups. Second, item facility numbers are calculated for all test questions for examinees in both of these groups. There are now two item facility numbers for each test question—one for examinees who did well on the test, and another for examinees who did poorly. Third, for each question, the item facility number for the low group is subtracted from the item facility number for the high group. The resulting number represents item discrimination.

For example, if the lower group of examinees had a composite item facility of .12 (12% of the examinees answered the question correctly) and if the higher group had a composite item facility of .92 (92% answered correctly) the resulting item discrimination number is produced by subtracting .12 from .92, which in this case yields .70. Generally, the higher the item discrimination number, the better a question is working—a higher number means that a question is efficiently separating examinees with different levels of ability—the common purpose of entrance examinations. While this description characterizes item analysis for norm-referenced proficiency tests, similar and slightly more complex item analysis can also be performed for criterion-referenced achievement tests.

When used together or individually, these tools for statistical analysis can provide a broad range of valuable information about the tests that a school is using to judge the proficiency of its potential students. With them, one can analyze individual test questions, sections of tests, whole tests, or more than one test in a variety of ways.

7.0 Conclusion

Careful preparation of the tests we give our students is but one aspect of the process of testing. These tests must also be critically examined in a number of ways to insure that students are evaluated fairly and in accordance with the demands of the curriculum that they will be studying. Also, thorough and frequent checks of the tests we use will produce better and more efficient tests that are suited to the particular skills of the students we are testing. Accurate and proper placement of students can improve student satisfaction with a school, and it can also make a language instructor's job less

difficult by grouping students in ways that match the material to be taught. Such analysis will help prevent the repetition of deficiencies that would otherwise go unnoticed.

BIBLIOGRAPHY

Bachman, Lyle F. "Problems in Examining the Validity of the ACTFL Oral Proficiency Interview." Studies in Second Language Acquisition. Vol. 10/2 June 1988, pp. 149-164.

Brown, James Dean. "Improving ESL Placement Tests Using Two Perspectives." TESOL Quarterly. Vol. 23, No. 1, March 1989, pp. 65-83.

Buck, Gary. "Written Tests of Pronunciation : Do They Work?" ELT journal. Vol. 43/1, January 1989, pp. 50-56.

Carter, Ronald, and Michael N. Long. "Testing Literature in EFL Classes: Tradition and Innovation." ELT Journal. Vol. 44/3, July 1990, pp. 215-221.

Clark, John L. D. and Ray T. Clifford. "The FSI/ILR/ACTFL Proficiency Scales and Testing Techniques." Studies in Second Language Acquisition. Vol. 10/2 June 1988, pp. 129-147.

Douglas, Dan. "Testing Listening Comprehension in the Context of the ACTFL Proficiency Guidelines." Studies in Second Language Acquisition. Vol. 10/2 June 1988, pp. 245-261.

Lantolf, James P. and William Frawley. "Proficiency : Understanding the Construct." Studies in Second Language Acquisition. Vol. 10/2, June 1988, pp. 181-195.

Lennon, Paul. "Conversational Cloze Tests for Advanced Learners." ELT Journal. Vol. 43/1, January 1989, pp. 38-44.

Malu, Kathleen F. "Entrance Testing and Course Placement at the UN International School, New York City." ELT Journal. Vol. 43/3, July 1989, pp. 206-212.

Matthews, Margaret, "The Measurement of Productive Skills : Doubts Concerning the Assessment Criteria of Certain Public Examinations." ELT Journal. Vol. 44/2, April 1990, pp. 117-121.

Peterson, Scott. "Cautionary Notes on Oral Language Testing." Cross Currents. Vol. XVI, No. 2, Fall 1989, pp. 95-100.

Shohamy, Elana. "A Proposed Framework for Testing the Oral Language of Second/Foreign Language Learners." Studies in Second Language Acquisition. Vol. 10/2 June 1988, pp. 165-179.

Smith, Jan. "Topic and Variation in ITA Oral Proficiency: SPEAK and Field-Specific Tests." English for Specific Purposes. Vol. 8, 1989, pp. 155-167.

Valdman, Albert, "Introduction." Studies in Second Language Acquisition. Vol. 10/2 June 1988, pp. 121-128.

Williams, K. L. "Three New Tests for Overseas Students Entering Postgraduate and Vocational Training Courses." ELT Journal. Vol. 44/1, January 1990, pp. 55-65.