EFL ASSESSMENT

Hasanuddin
Rasuna Talib

1

Editor : Novrianto Napu, MA, Ph.D.

ELT Assessment 1

Prof. Dr. Hasanuddin, M.Hum.
Dr. Hj. Rasuna Talib

deepublish
glorify and develop the intellectual of human's life

ELT ASSESSMENT 1

Hasanuddin
Rasuna Talib

Desain Cover : Dwi Novidiantoko
Tata Letak Isi : Emy Rizka Fadilah

Cetakan Pertama: Desember 2017

PENERBIT DEEPUBLISH
(Grup Penerbitan CV BUDI UTAMA)
Anggota IKAPI (076/DIY/2012)
Jl.Rajawali, G. Elang 6, No 3, Drono, Sardonoharjo, Ngaglik, Sleman
Jl.Kaliurang Km.9,3 – Yogyakarta 55581
Telp/Faks: (0274) 4533427
Website: www.deepublish.co.id
www.penerbitdeepublish.com
E-mail: cs@deepublish.co.id

Katalog Dalam Terbitan (KDT)

HASANUDDIN
ELT Assessment 1/oleh Hasanuddin & Rasuna Talib.--Ed.1, Cet. 1--
Yogyakarta: Deepublish, Desember 2017.

x, 86 hlm.; Uk:15.5x23 cm

ISBN **978-602-453-559-9**

1. English I. Judul
420

PREFACE

English Language Teaching Assessment is a subject matter that has to be taken by students of English Department. English Language Teaching Assessment is one of pedagogic competence that has to be mastered by English teachers especially in Indonesian context. This is book is very important to learn by students who will develop their competence in English Language Teaching and assessment. This book support to students who want to know about testing, assessing, and teaching, approach to language testing, testing and teaching, principles of language assessment, testing and curriculum, adopting, developing, and adapting language tests, developing and improving language tests, describing test results, and interpreting test scores. We hope this book has a significant impact to students for developing competency in teaching and assessing English. We would like to say thank very much for all friends who have supported to finish this book.

Gorontalo, October 2017

Hasanuddin
Rasuna Talib

CONTENTS

CHAPTER 3
PRINCIPLES OF LANGUAGE ASSESSMENT

CHAPTER 4
TESTING AND CURRICULUM

CHAPTER 7

DESCRIBING TEST RESULTS

CHAPTER 8

INTERPRETING TEST SCORES

CHAPTER 1
TESTING, ASSESSING, AND TEACHING

Short Description

This chapter simply describes (1) definition of a test (2) assessing and teaching, (3) informal and formal assessment, (4) formative and summative assessment. The four subsections are discussed in the following. The benefit of this material is to give knowledge of students about test, assessment, teaching, and relationship among them.

Basic Competence

Students are able to explain testing, assessing, and teaching that covers.

- Definition of a test
- Assessing and teaching
- Informal and formal assessment
- Formative and summative assessment

1.1 What is a Test?

A test can be defined as a way of measuring the ability, performance or knowledge of a person in a particular domain. Brown (2004) gives four components of a test definition. First, a test is a method that consists of techniques, procedures, or items that require performance on the part of the test-taker. To qualify as a test, a method must be expressed and organized. For example, multiple questions with prescribed correct answers, oral interview, etc. Second, a test must measure. There are tests that measure general ability while others focus on particular competencies. The way the results or measurements are communicated may vary, such as a classroom-based short answer essay test. Third, a test

measures a person's capacity, learning, or performance. The test analyzers need to comprehend who the test-takers are. What are their experience and foundation? Is the test properly coordinated to their capacities? In what manner should test-takers decipher their scores?

Fourth, a test measures performance. Most language tests measure one's capacity to perform a language skill, that is, to talk, compose, read, or listen to a subset of the language. Language performances consist of speaking, writing, reading and listening that are sometimes administered one language skill in one opportunity, but it may develop by using integrated tests that cover all language skills in one opportunity.

Finally, a test measures a given domain. For this situation of a language proficiency test, despite the fact that the real execution of the test includes just an inspecting of abilities is looking at the capability one's language competence – general competence in all skills of a language. Other tests may have more specific criteria. For instance, pronunciation tests, vocabulary tests.

1.2 Assessment and Teaching

Assessment is sometimes misunderstood in current educational practice. We might be tempted to consider testing and assessing as in the form of synonymous terms. However, this is not the case.

Tests are readied by authoritative systems that happen at identifiable circumstances in educational modules or curriculum when students ace every one of their difficulties to offer peak performance, realizing that their reactions are being measured and assessed.

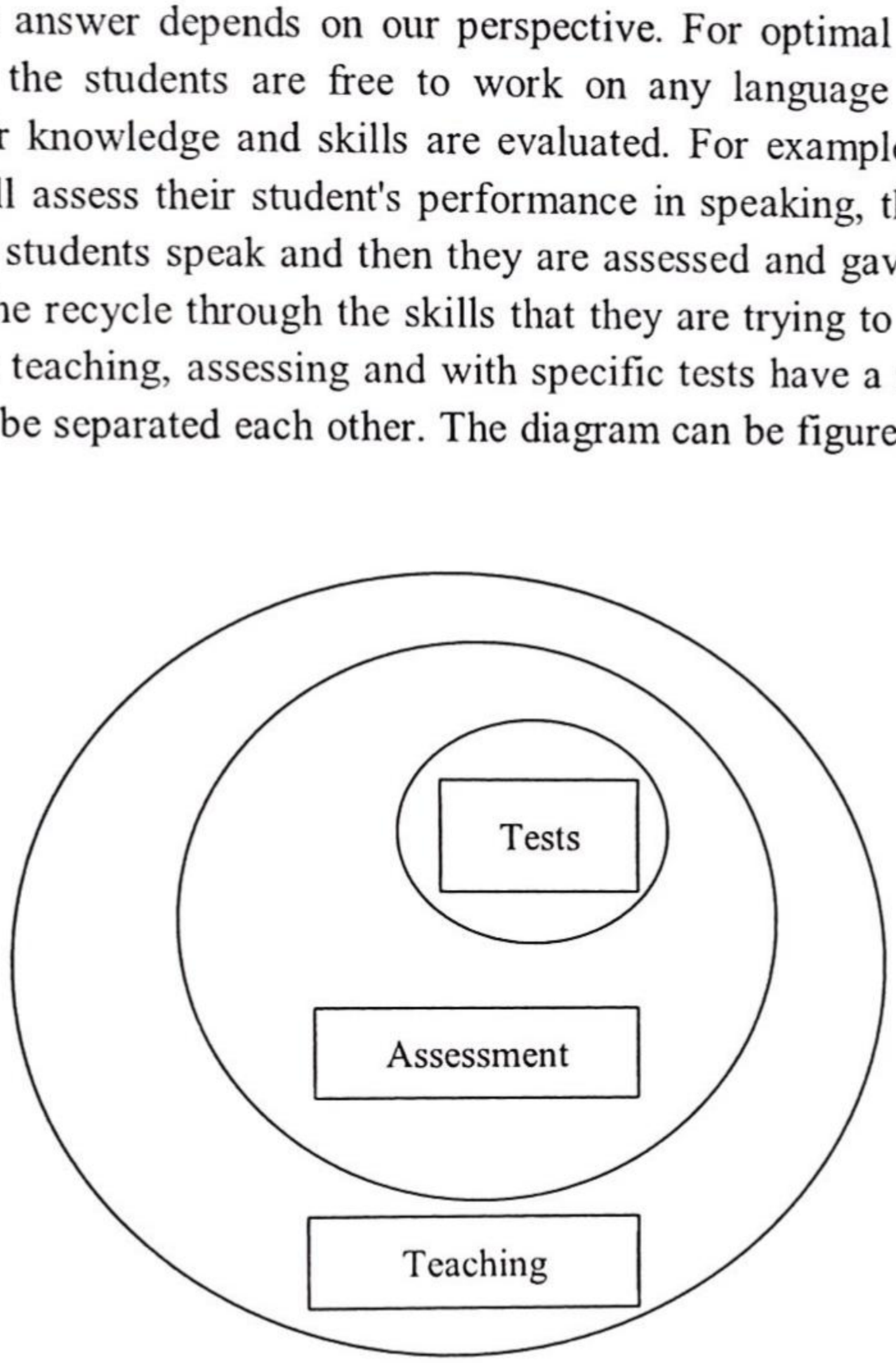
Assessment, on the one hand, is a progressing procedure that includes a considerably more extensive area. Each time a student responds to a question, offer a comment, or tries out another word or structure, the educator intuitively makes an appraisal of the performance of the student. Written work – from a jotted-down phrase to a formal essay – is assessed directly by the performers themselves, teacher and students.

Tests, at that point, are a subset of appraisal; they are unquestionably by all account not the only type of assessment that an

educator can make. Tests can be valuable gadgets. However, they are just a single among numerous techniques and assignments that the instructors can at last use to evaluate learners.

However, now, we might be thinking if we make assessments every time we teach something in the classroom, does all teaching include assessment? Are instructors continually evaluating learners without interaction of assessment free?

The answer depends on our perspective. For optimal learning to take place, the students are free to work on any language hypotheses despite their knowledge and skills are evaluated. For example, when the teachers will assess their student's performance in speaking, the first step they let the students speak and then they are assessed and gave feedback. Therefore the recycle through the skills that they are trying to master that begins with teaching, assessing and with specific tests have a relationship that cannot be separated each other. The diagram can be figured out in the following.



ELT Assessment

1.3 Informal and Formal Assessment

Informal assessment consists of spontaneous responses, practices and the other feedback for the students. Examples include saying "Nice job!", "Good work!". However, this assessment is more emphasized on the outcome of classroom assignment without considering and scoring students' competence. The example at this end of the continuum is providing comments on written paper and essay, correct pronunciation instruction, reading strategy along with note-taking modification for the students.

Formal assessments, in contrast, is distinctively designed for skill and knowledge by using a systematic and purposive sampling technique to provide a learning achievement's assessment for the teacher and students. This is a course periodic diversion to expand the analogy.

It is widely known that all tests are categorized as formal assessment, but not all formal assessments are testing. For example, we may use student's journal or portfolio of materials as a formal assessment of certain course objectives.

1.4 Formative and Summative Assessment

Most of our classroom assessment is formative assessment: evaluating students in the process of "forming" their knowledge, ability, and skill to keep up that improvement process. For all practical purposes, virtually all kinds of informal assessment are (or should be) formative that focus on going process.

Summative assessment aims to measure or summarise, what a student has grasped and occurred at the end of a course unit of instruction. A summation of what a student has learned implies looking back and taking stock of how well that student has accomplished objectives. The proficiency examination is including in summative assessment.

Summary

This chapter simply describes (1) what is a test? (2) assessing and teaching, (3) informal and formal assessment, (4) formative and summative assessment.

Firstly, a test is a method that consists of techniques, procedures, or items that require performance on the part of the test-taker. Secondly, a test must measure. Thirdly, a test is to assess one's competencies and performances. Is the test appropriately matched to their abilities? How should test-takers interpret their scores? Fourthly, a test measures performance. Language tests are mostly conducted to evaluate one's language skills performance including listening, speaking, reading and writing. Finally, a test measures a given domain. For instance, pronunciation tests, vocabulary tests.

Assessment and Teaching

The test is the part of assessment, yet it is not the only way teacher can deal with. Does the teacher always test students without free assessment communication?

Informal and Formal Assessment

The informal assessment does not stop there. It is widely known that all tests are categorized as formal assessment, but not all formal assessments are test.

Formative and Summative Assessment

Most of our classroom assessment is formative assessment: evaluating students in the process of "forming" their knowledge, ability, and skill in order to keep up that improvement process.

Comprehension Question

1. What is a test?
2. Explain the relationship between test, assessment, and teaching.
3. Explain four components of tests.

References

Bachman, L.F. 1990. *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

Brown, J.D. 1996. *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Djiwandono, M.Soenardi. 1986. *Kemampuan Berbahasa dan Penilaiannya dalam Pengajaran Bahasa:* Pidato Pengukuhan pada Penerimaan Jabatan Guru Besar IKIP Malang. IKIP Malang.

Djiwandono, M, Soenardi. 1996. *Tes Bahasa dalam Pengajaran.* Bandung; Penerbit ITB.

Ebel, R.L. and Frisbie, D.A. 1986. *Essentials of Educational Measurements.* New Jersey: Prentice Hall, Inc.

Groundlund, Norman E. 1986. *Measurement and Evaluation in Teaching.* New York: Macmillan Publishing Company.

Hasanuddin.2003. *Language Testing.* Gorontalo: IKIP Press.

CHAPTER 2
INTRODUCTION TO LANGUAGE TESTING

Short Description

At the beginning of the explanation of language testing, this chapter will talk about (1) testing and teaching, (2) why tests, (3) what should be tested and what standard, (4) testing language skills and language areas, and (5) approaches to language testing. This chapter has significance for developing students' knowledge related to introduction to language testing than emphasizes on testing and approaches to language testing.

Basic Competence

Students are able to explain introduction to language testing that covers

- Why tests
- What should be tests and what standard
- Testing language skills and language areas
- Approaches to language testing

2.1 Testing and Teaching

Testing and teaching are associated with each other; ones cannot regard this as a separate matter. In other words, the close interrelation between these terms leads to a condition where people must not ignore the notion of testing if they discuss notions of teaching and vice versa. This is because tests are designed to know the level of students' language competency and even motivate them as a medium to boost their performance. Djiwandono (1996) states that language performance of the student refers to his ability to use language in every day or a real

communication. The language proficiency signifies one capability in using a language, i.e., to communicate, express feelings, and argue with others.

Language mainly serves a function to discover things that learners can do with language; this is primarily about practicing daily communication activities. Such a test would have a more beneficial impact on the learning of particular language than a mechanical test of the structure. "In the past, even good assessment of grammar, translation or language manipulation had a negative and even harmful impact on teaching" (Heaton, 1988). A good communicative language test, however, must embed the sense of giving the positive implications, such as an improved learning habit.

Luckily, some testing institutions are currently working on an attempt to examine the success rate of a candidate in performing purposeful, relevant tasks, as well as the actual skill in communicating with a particular language.

In this regard, such examinations undoubtedly contribute to a positive direction on syllabuses and teaching strategy compared to ones in the past. Still, some of the institutions heavily focus on comparing performances of a student with one another.

2.2 Why Test?

The function indicated in the following paragraph explains among answers to the question: "Why test?" However, it must be emphasized that the evaluation of the student performance for purposes for comparison or selection is only one of the function of a test. Giving a test, as far as the practicing teacher is concerned, is not the only way to evaluate students' performance, especially in schools.

Heaton (1988) states that a test will help a teacher to assess individual performance. A good classroom test will also contribute to locating the precise areas of difficulty encountered by the class or by the individual student. In addition, the idea of assisting the teacher in finding difficulties of certain parts of the language program should be embedded in

the test. By that, This enables the teacher to examine the effectiveness of the syllabus, the methods, and materials that he or she is applying.

A test also serves a purpose to effectively motivate students if it is designed appropriately without integrating trap questions. Such a test will provide the student with the opportunity to show their ability to perform certain tasks in the language. Giving the students their test outcome immediately will help them to ascertain their weaknesses. In other words, a well-designed test can be considered as a helpful tool in teaching.

2.3 What should be tested and to what standard?

The advancement of the recent theory in linguistics is beneficial for language teachers and testers in gaining awareness of the necessity to analyze the language that is being tested. The Latin-based prescriptive grammars are replaced by the Modern descriptive grammars since the entire complex system of language skills, and patterns of linguistic behaviors are being examined by the linguists. Indeed, The language skills are without question a complex matter that associates with the whole context of those nonlinguistic skills are used.

Before constructing a test, ones should consider the standards which are being set. These standards should meet the demands of foreign language learners. Another important question to answer is "should foreign language learners after a period of times be expected to communicate with the level and fluency of native speakers?

2.3.1 Testing The Language Skills

Four major skills in communicating through language are often broadly defined as listening, listening and speaking, reading and writing. Ones are required to integrate these skills conscientiously; this is to design good communicative tasks as many as possible. Thereby, the test writer is urged to focus on such test items. On top of that, such test items are undoubtedly related to the use of language for real-life communication, especially in oral interaction.

The following are examples of assessment of four main language skills:

- Listening (auditory) comprehension, i.e., dialogues, talks, and lectures are given to the testees;
- Speaking ability, i.e., interview, a picture, description, role play, and a problem-solving task involving pair work or group work;
- Reading comprehension, whereas the set questions aims to test the students' ability to understand the gist of a text and obtain the main information on particular points in the text; and
- Writing ability, i.e., letters, instructions, messages, reports, memos, and accounts of past events, etc.

2.3.2 Testing Language Areas

In an effort to insulate the areas of language learning, a substantial number of tests involve units on (1) grammar and usage; (2) vocabulary (concerned with word meanings, word formation, and collocations); (3) phonology (concerned with phonemes, stress, and intonation).

The three testing language areas have different areas of measurement. Tests of grammar and usage measure students' ability to recognize appropriate grammatical forms and to manipulate structures. A test of vocabulary evaluates students' understanding of related aspects, i.e. meaning of a word, the use of a particular word, as well as its collocations. This may test their active vocabulary (the use of words in the spoken and written language) or their passive vocabulary (the words someone or when they are reading). Tests of phonology are designed to test phonology in order to assess the following sub-skills. They are: (1) the ability to recognize and pronounce the significant sound contrasts of a language, (2) ability to recognize and use the stress patterns of a language, and (3) ability to hear and produce the melody or patterns of the tunes of a language (i.e. the rise and the fall of the voice).

2.4 Approaches to Language Testing

The classification of language tests consists of four major testing approaches: (1) the essay translation approach, (2) the structuralist

approach, (3) the integrative approach, and (4) the communicative approach. This list of approaches is, however, not limited to certain periods in the development of language testing.

2.4.1 The Essay Translation Approach

This approach is commonly referred to the pre-scientific stage of language testing. The essay translation approach requires no special skill or expertise in testing. Instead, such an approach requires the subjective judgment of the teacher the most. Tests basically comprise of essay writing, translation, and grammatical analysis (often in the form of comments *about* the language being learned). Public examinations resulting from the essay translation approach integrate an oral component at the upper intermediate and advanced levels – though this has sometimes been considered in the past as something additional and no way an integral part of the syllabus or examination.

2.4.2 The Structuralist Approach

This approach believes that language learning focus on the systematic acquisition of a set of habits. The grounding of this approach involves several aspects of structural linguistics, such as the importance of contrastive analysis and the urgency to classify and examine learners' mastery of separate elements of the target language, i.e., phonology, vocabulary, and grammar. Assessing learners' mastery is by giving a test using words and sentences separated from any context based on the considerations that the examining the whole language competence of a person can be done in the test in a comparatively short time; in other words, it is essential to test one aspect at a time. As a result, testing the four main language skills, i.e., listening, speaking, reading and writing are conducted separately.

Such feature of the structuralist approach is, of course, still valid for certain types of test and certain purposes. For instances, the needs to focus on the test takers' ability to write by trying to separate a composition test from reading (i.e., by making it entirely independent of the ability to

read long and complicated instructions of or verbal stimuli) is commendable in some ways.

2.4.3 The Integrative Approach

The major focus of this approach is about meaning and the total communicative effect of discourse; it involves the context-based language test. By that, integrative tests are constructed to assess learners' capability in using more than one skill at a time instead of assessing the language proficiency separately. Thus, integrative tests have an interrelation with a global view of competence – underlying language competence or 'grammar of expectancy,' which it is argued every learner possess regardless of the purpose of which the language is being learned. Integrative testing involves 'functional language' but not the use of functional language. Integrative tests are best characterized by the use of cloze testing and dictation.

The tenet of cloze testing is based on the Gestalt theory of closure (closing gaps in patterns subconsciously). Therefore, cloze tests aim at examining the ability of readers to decode 'interrupted' or 'mutilated' messages through producing the most acceptable substitutions from all the contextual clues available.

2.4.4 The Communicative Approach

The communicative approach to language testing sometimes relates to the integrative approach. Despite this similarity, there are fundamental differences between these two approaches. Communicative tests, to certain extents, focus on the use of languages in communication. Success can be measured by the effectiveness of the communication rather than the accuracy of formal linguistics.

Divisibility hypothesis is regarded as a grounding in measuring different language skills in this approach. Communicative testing aims at obtaining different profiles regarding a learner's language performance. For example, a learner may score low in an informal-style-oriented speaking test but may score high on the tests of reading comprehension.

Consequently, the outcome of one's language performance is displayed in separated measures of each proficiency as depicted in the following Table 1 (each with six boxes to indicate the different levels of students' performances).

Table 1 Four basic skills in a Communicative test

	6	5	4	3	2	1
Listening						
Reading						
Listening & Speaking						
Writing						

The foremost concern of communicative tests is to involve accurate and detailed information regarding the needs of the learners. An example of this is the testing of English for specific purposes. Nevertheless, considering such a test is only exclusive to ESP, or even mature learners with short term goals is a misconception. Young learners who study English have this type of test possessing the same notion as those for adults; yet, the tasks might be different.

Lastly, the qualitative modes of assessment have been integrated with communicative testing in preference to those of quantitative's. This is done through a system well-known as a language band system; it is used to show information regarding learner's capability in different skills of language. Furthermore, the details of each skill of performance act as the reliability of the scoring in which assist the examiner in decision-making process based on the thorough and well-constructed criteria.

Summary

At the beginning of the explanation of language testing, this chapter will talk about (1) testing and teaching, (2) why tests, (3) what should be tested and what standard, (4) testing language skills and language areas, and (5) approaches to language testing.

Testing and Teaching

Testing and teaching cannot be separated each other. One of the aims of designing a test is to bolster learning and motivate students in addition to assessing students' language competencies. Djiwandono (1996) states that language performance of the student refers to his ability to use language in every day or a real communication. The language ability signifies the level of language mastery in aspects, such as how people communicate, express their feeling, and argue to others.

On the other hand, a good communicative test of language should have a much more positive impact on learning and teaching and should result in improved learning habits.

Heaton (1988) states that a test will help a teacher to evaluate individual performance. A well-designed classroom test will provide the student with the opportunity to show their ability to perform certain tasks in the language. Questions about "what should be tested and to what standard" are to be considered the most.

There is a rise in the awareness of language teachers and test designers regarding the urge to assess the test language since the development of modern linguistic theory has helped them to make such progress.

Testing the language skills

Ways of evaluating performance in the four major skills may take the form of tests of:

Testing language areas

The three testing language areas have different areas of measurement. Tests of grammar and usage measure students' ability to recognize appropriate grammatical forms and to manipulate structures. Tests of phonology are designed to test phonology to assess the following sub-skills. Approaches to Language Testing

There are four main approaches to language tests: (1) the essay translation approach, (2) the structuralist approach, (3) the integrative approach, and (4) the communicative approach.

Tests usually consist of essay writing, translation, and grammatical analysis (often in the form of comments *about* the language being learned). As a result, integrative tests mostly consider the way a learner uses two or more language skills simultaneously instead of assessing each skill separately. Integrative testing involves 'functional language' but not the use of functional language. Integrative tests are best characterized by the use of cloze testing and dictation. The communicative approach to language testing somehow relates to the integrative approach. Communicative tests are concerned primarily (if not totally) with the use of a language in communication.

A view of language well-known as the divisibility hypothesis is the grounding of assessing one's language skills in the communicative test. This test attempts to find out each of the profile of a learner's language performance; it is usually displayed in the form of language band systems.

Comprehension Question

1. Why tests are necessary for teaching
2. What should be tests and what standard?
3. What are testing language skills and language areas?
4. Develop testing language skills and language areas of English.
5. Explain four Approaches to language testing and give examples of each.

References

Bachman, L.F. 1990. *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

Brown, J.D. 1996. *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Djiwandono, M.Soenardi. 1986. *Kemampuan Berbahasa dan Penilaiannya dalam Pengajaran Bahasa:* Pidato Pengukuhan pada Penerimaan Jabatan Guru Besar IKIP Malang. IKIP Malang.

Djiwandono, M, Soenardi. 1996. *Tes Bahasa dalam Pengajaran.* Bandung; Penerbit ITB.

Ebel, R.L. and Frisbie, D.A. 1986. *Essentials of Educational Measurements.* New Jersey: Prentice Hall, Inc.

Groundlund, Norman E. 1986. *Measurement and Evaluation in Teaching.* New York: Macmillan Publishing Company.

Hasanuddin.2003. *Language Testing.* Gorontalo: IKIP Press.

CHAPTER 3
PRINCIPLES OF LANGUAGE ASSESSMENT

Short Description

This chapter explains how principles of language assessment can and should be applied to formal tests, but the ultimate recognition that these principles also apply to assessments for all kinds. These principles will be used to evaluate an existing, published or created test. It will center on how to use those principles to design a good test. Therefore, this chapter has significance for the understanding of principles of language assessment.

Basic Competence

Students are able to explain and apply principles of language assessment that covers

- Practicality
- Reliability
- Validity
- Authenticity

3.1 Practicality

According to Brown (2004) "an effective test is practical that means that (1) it is not expensive, (2) it stays within appropriate time constraints, (3) it is relatively easy to administer, and (4) it has a scoring/evaluation procedure that is specific and time-efficient. A test that is expensive is impractical. A language proficiency test that takes a student five hours to complete is impractical in which it consumes more time". Besides that, a test that is easy to administer makes easy to a test taker and

tester. For example, multiple choice tests are easy to administer and score. Moreover, a test that is easy to score is also effective to know and evaluate the test results. Therefore, the decision of test results can be found earlier.

3.2 Reliability

A reliable test is consistent and dependable. If we give the same test to the same student or matched students on two different occasions, the test should yield similar results. Brown (2004) states the issues of reliability of a test may be considered a number of factors. First is student-related reliability in which the most common learner-related issue in reliability is caused by temporary illness, a bad day, anxiety, and other physical or psychological factors, which may make an 'observed score deviate from one true score. Second is rater reliability takes place if the result scores are inconstant in the same test and human error, subjectivity, and bias may enter into the scoring process. Rater-reliability issues are not limited to contexts where two scores are involved. Inter-rater reliability is a common occurrence for classroom teacher because of unclear scoring criteria, fatigue, and inclination towards general recklessness of students. The third is a test administration reliability. Unreliability could be the result of the conditions in which the test is administered. Unreliability is due to the test administration state, and it is from various documents copying, the amount of light in different parts of the room, variation in temperature, and the condition of desks and chairs. The last is test reliability. If there are too many test items, the test takers may find it difficult to stay focus on it, so that it will affect their answers.

3.3 Validity

The most complex criterion of an effective test and the most important principle is validity, in which the inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment. There are five types of evidence of validity. First is content related evidence if the subject test sample concerns with the drawn conclusion and demands the test taker to perform the assessed

task. The other way to understand the content validity is by taking into account the difference between direct testing that involves the test taker in performing targeted work and indirect testing. In an indirect test, learners are not performing the task itself.

Second is criterion-related evidence in which the criterion of the test has been reached. Criterion related validity usually falls one into two categories: concurrent and predictive validity. The test has concurrent validity if its results are supported by other concurrent performance beyond the assessment itself. The predictive validity of an assessment becomes necessary in the case of placement tests, admission of assessment batteries, language aptitude tests, and the like.

The third is construct-related evidence that is commonly referred to theory or hypothesis constructed to describe the phenomena in our universe of perception. Finally, face validity is the suitable test to measure one's abilities with the subjective scoring of the test takers. A qualified face validity will be highly achieved if the students deal with (1) a well-constructed, expected format with similar tasks, (2) a test that is clearly doable within the allotted time limit, (3) items that are clear and uncomplicated, (4) directions are clear, (5) tasks that relate to their course work (content), and (6) a difficulty that presents a reasonable challenge.

3.4 Authenticity

Bachman and Palmer in Brown (2004) define authenticity as "the degree of the correspondence of the characteristics of a given language test task to the features of a target language task, and then suggest an agenda for identifying those target language tasks and for transforming them into valid test items." In a test, authenticity can be presented by composing: (1) natural language, (2) contextual test items, (3) important topic, (4) thematic creation test items, and (5) reality-based tasks.

Summary

These principles will be used to evaluate an existing, published or created test. Reliability

Validity. A test that is expensive is impractical. Besides that, a test that is easy to administer makes easy to a test taker and tester. For example, multiple choice tests are easy to administer and score. Moreover, a test that is easy to score is also effective to know and evaluate the test results. Therefore, the decision of test results can be found earlier.

Reliability

A reliable test is consistent and dependable. If the same test is given to the students on two different occasions, the test should yield similar results. Inter-rater reliability takes place if the results are inconstant in the same test. The third is test administration reliabiiity. Unreliability is due to the test administration state. The last is test reliability. If a test is too long, test-takers may become fatigued by the time they reach the following items and respond incorrectly.

Validity

Direct testing is done by having the students work directly right on the test. In an indirect test, learners are not performing the task itself.

The test has concurrent validity if its results are supported by other concurrent performance beyond the assessment itself. The predictive validity of an assessment becomes necessary in the case of placement tests, admission of assessment batteries, language aptitude tests, and the like.

Authenticity

In a test, authenticity can be presented by composing: (1) natural language, (2) contextual test items, (3) significant topic, (4) thematic creation test items, and (5) reality-based tasks.

Comprehension questions and tasks

1. What are practicality, reliability, validity, and authenticity?
2. Why are practicality, reliability, validity, and authenticity so important in testing English?
3. Take a sample of English tests that you have administered and then analyze their reliability and validity.

References

Bachman, L.F. 1990. *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

Brown, J.D. 1996. *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Djiwandono, M.Soenardi. 1986. *Kemampuan Berbahasa dan Penilaiannya dalam Pengajaran Bahasa:* Pidato Pengukuhan pada Penerimaan Jabatan Guru Besar IKIP Malang. IKIP Malang.

Djiwandono, M, Soenardi. 1996. *Tes Bahasa dalam Pengajaran.* Bandung; Penerbit ITB.

Ebel, R.L. and Frisbie, D.A. 1986. *Essentials of Educational Measurements.*New Jersey: Prentice Hall, Inc.

Groundlund, Norman E. 1986. *Measurement and Evaluation in Teaching.* New York: Macmillan Publishing Company.

Hasanuddin.2003. *Language Testing.* Gorontalo: IKIP Press.

CHAPTER 4
TESTING AND CURRICULUM

Short Description

This chapter will discuss (1) the place of tests in curriculum planning that consists of need analysis, goals and objectives, language testing, material development, language teaching and program evaluation; (2) the place of tests in curriculum implementation that deals with ELI as a language program, initial screening and proficiency procedures, placement procedures, second-week diagnostic procedures, achievement procedures, and testing as an integrated system.

Basic Competence

Students are able to explain the relationship between testing and curriculum that covers

- The place of tests in curriculum planning
- The place of tests in curriculum implementation

4.1 The Place of Tests in Curriculum Planning

Curriculum planning or development is viewed here as a series of activities that provide a support framework that helps teachers to design effective activities and learning situations to promote language learning. Brown (1996) describes six broad types of activities that are often identified in the curriculum design in order to promote good teaching and learning: needs analysis, goals and objectives setting, testing (both NRT and CRT), materials development, teaching, and program evaluation.

4.1.1 Needs Analysis

In language teaching, needs analysis is often seen as the identification and selection of the language forms that the target students are likely to require in actually using a particular language. Most often, the focus is on the learners and what they need to learn, and these needs are usually expressed in linguistic terms. Such a focus seems reasonable because learners are the primary "clients" in a language program, and the curriculum should be designed to serve the clients' needs.

Two dangers, however, may arise in taking needs analysis. First danger results from the fact that teachers, administrators, employers, institutions, societies, and even nations have needs that may influence the delivery of language teaching and the effectiveness of language learning that follows. A second danger arises from the fact that a needs analysis usually focuses solely on the linguistic factors involved. Students are people and therefore have needs and concerns that are not solely linguistics. The solution to this problem may be found in making the linguistic forms the focus of the needs analysis, but gathering information from many sources as possible on the students' other needs (including physical, personal, familial, professional, cultural, societal, and so forth).

To avoid the dangers just described, needs analysis is defined rather as "the systematic collection and analysis of all relevant information that is necessary to satisfy the language learning needs of the students within the context of the particular institution(s)" (Brown,1996:272).

4.1.2 Goals and Objectives

If the purpose of doing needs analysis is to satisfy the language learning needs of the students, one outcome of analysing those needs might be the specification of formal *program goals*. Such goals are general statements of what must be accomplished in order to satisfy the students' needs. For example, the students in the ELI at UHM are taking ESL courses for the purpose of improving their ability to study English at the university level.

Objectives, on the other hand, are statements of the exact contents, knowledge, or skills that students must learn in order to achieve a given goal. For instance, taking the goal mentioned in the previous paragraph – being able to write a term paper – the teachers might realise that student's first need to develop essential library skills. One such skill might be the ability to find a book in the library. To do this, the student needs several sub skills: knowing the English alphabet, pinpointing a particular book in the catalogue, locating the call number for that book, and finding the book. Thus, objectives are derived from considering how to achieve the program goals best. Recall that the goals were in turn derived from perceptions of what the students needed to learn.

4.1.3 Language Testing

According to Brown (1996), the next logical step in curriculum development is the drafting of tests based on a program's objectives. The goals, objectives, and administrative necessities of a program may require extensive test development. This may be in turn necessitate adopting, developing, or adapting tests for a wide variety of decisions including the proficiency, placement, diagnostic, and achievement decisions.

The tests may have significance. The tests can then help teachers to investigate the degree to which the objectives are appropriate for the students in question *before* investing the time and energy needed to adopt, develop, or adapt the materials required to teach those objectives. While NRT proficiency and placement test results may be useful for determining approximately the level of the materials and teaching, only CRTs can directly measure students' abilities with regard to the course objectives. With CRT results in hand, teachers can then determine the degree to which the objectives are appropriate for the particular students involved in the particular course of study.

In short, we found that developing CRTs to diagnose the students' weaknesses and strengths also afforded us the opportunity to analyse the appropriateness our course objectives for their abilities. Unfortunately, our perceptions of their needs were pitched far too low.

4.1.4 Materials Development

Once tentative needs analysis, objectives, and tests have been put into place, materials can be developed in a rational manner to fit the specific needs and abilities of the participants in the program. With the appropriate testing information in hand, adopting, developing, or adapting materials become relatively easy because the course is fairly clearly defined. Indeed, the very decisions about whether to adopt, develop, or adapt materials become much easier. For instance, teachers can easily assess existing materials and decide whether these materials could be adopted to fill the needs of the students, or whether they will have to be adapted to meet the students' needs and program goals.

The purpose of this section is not to prescribe any particular syllabus or materials for any program but rather to argue that materials should be based on the needs analysis, objectives, and test results of the specific curriculum in question. Such decisions must be left to the teachers and administrators who are on site and know the situation best.

4.1.5 Language Teaching

Contrary to what many teachers may think, the type of curriculum development described here can allow teachers more freedom than usual in the classroom to teach in the ways that they judge to be correct. In such a curriculum, both the teachers and students are aware of the objectives for each course and are mindful of the fact that these objectives will be tested at the end of the course. In other words, the teachers must be involved in the process of curriculum development, feedback, and revision, and they must often be consulted along the way.

There is strength to be found in numbers, so curriculum planners will find it useful to involve teachers, administrators, and students in defining the needs within a particular program and establishing the course objectives and tests. Teachers are typically required to determine what the students need to learn, define, the course goals and objectives, select or develop course tests and materials, and do the teaching.

4.1.6 Program Evaluation

Often the terms *testing* and *evaluation* are linked to each other and sometimes are even used interchangeably. The evaluation may involve some testing as one source of information, but evaluation is not limited in any way to testing. Brown (1996) defines *evaluation* as the systematic collection and analysis of all relevant information necessary to promote the improvement of the curriculum and analyse its effectiveness within the context of the particular institution(s).

A needs analysis is usually conducted at the beginning of a curriculum development project and is focused on the linguistic needs of the participants. In the needs analysis phase, information is gathered using interviews, questionnaires, linguistic analyses, guesswork, and a good deal of professional judgment. In contrast, evaluation strategies can be broader, using all available information to analyse the effectiveness of the program. Thus, evaluation can use all the information gathered in (a) doing the initial needs analysis, (b) developing, listing, and refining objectives, (c) writing, piloting, and revising tests, (d) adopting, developing, or adapting materials, and (e) putting all the above in place through teaching (Brown, 1996).

In short, evaluation allows a language program to profit from ongoing information gathering, analysis, and synthesis, for purposes of improving each component of a curriculum based on information about all the other components separately and collectively. This ongoing process is what makes the systems approach to curriculum development so potentially powerful and efficient.

4.2 The Place of Tests in Curriculum Implementation

Curriculum implementation involves actually putting in place the elements developed in the curriculum planning and making them work and fit together within the existing program in a way that will help administrators, teachers, and students. In this section, the focus is on those issues related to the role of *tests* in curriculum implementation. Moreover, also the focus of this section is on how the various types of tests all feed

information into the program – information that has not only positive effects on the students' lives but also influences all the elements of the curriculum itself.

4.2.1 Initial Screening and Proficiency Procedures

Many teachers may find themselves in a position in which they need *proficiency procedures* to determine how much of a given language their students have learned during their lives. At first, they will only be concerned with knowing about the students' proficiency in general terms without reference to any particular program. This is likely to be the case when the students are brand new to a language program, and the teachers want to get a general notion of how much of the language they know. To do this, teachers will probably need tests that are general in nature, such as the TOEFL in the ELI example. These same teachers may also want to establish guidelines for which types of students are automatically exempt from training, for which students need to take the placement test, and for which students deserve an interview or further information gathering.

At the same time that they are using such initial screening measures, teachers may be able to get a tentative estimate of the general level of language proficiency among their students. Such information may aid in determining entrance standards (or exit) for a curriculum, in adjusting the level of goals and objectives to the true abilities of the students, or in making comparisons across programs. As a result, initial screening procedures are often based on proficiency tests that are general in nature but important and globally related to curriculum structure.

4.2.2 Placement Procedures

The duties of the Director of the English Language Institute (ELI) including placing the students into levels of study are as homogeneous as possible to facilitate the overall teaching and learning of ESL. To that end, the ELI has quite naturally developed its *placement procedures.* These procedures are not based on the placement test results, as is the case in some language programs. In addition to the test scores, we use the

information gained from the initial screening as well as the second-week diagnostic and achievement procedures.

Most teachers will find themselves having to make placement decisions. In most language programs, students are grouped according to ability levels. Such grouping is desirable so that teachers can focus in each class on the problems and learning points appropriate for students at a particular level. Placement tests can help teachers to make such decisions. Such tests are typically norm-referenced and therefore fairly general in purpose, but, unlike proficiency tests, placement tests should be designed to fit the abilities and levels of the students in the particular program. The purpose of such tests is to show how much ability, knowledge, or skill the students have. The resulting scores are then used to place students into levels of study, or at times to exempt them entirely.

To do this, teachers need tests that are general but designed specifically for the types and levels of their students as well as for the goals of their program. Teachers may also need to establish guidelines for using as many types of test information as possible along with other types of data. Also, they might want to conduct placement interviews wherein all available information is marshalled for making the placement decisions.

4.2.3 Second-Week Diagnostic Procedures

Many teachers may find themselves using such diagnostic procedures for purposes of checking if their placement decisions were correct, but also for purposes of identifying and diagnosing strengths and weaknesses that students may have with relation to the course objectives and the material to be covered in the course. These procedures may be based on test results, but other factors should probably also come into play. The teachers' observations of the students' classroom performances and attitudes may be one source of information.

These diagnostic procedures are clearly related to achievement procedures. After all, diagnosis and achievement decisions can be based on two administrations of the same test. However, while diagnostic decisions are usually designed to help identify students' strengths and weaknesses at

the beginning or during instruction, achievement procedures are typically focused on the degree to which each student has accomplished the course objectives at the end of instruction.

4.2.4 Achievement Procedures

In the ELI, the CRT post-tests are administered as part of the *achievement procedures.* The CRT achievement tests are administered during the students' regularly scheduled final examination periods, which are two hours long. The students are, of course, told all this at the beginning of the course. Since until recently our criterion-referenced tests were more or less experimental, we were cautious about treating them as minimal competency tests on which students must achieve a certain minimum score to pass the course.

Most teachers will probably agree that they would like to foster achievement, particularly in the form of language learning, in their course or program. In order to find out if their efforts have been successful and to help them maximise the possibilities for student learning, achievement procedures like our tests and performance reports may prove useful. The tests used to monitor such achievement should be developed to measure the particular objectives of a given course or program. Moreover, that they must be flexible in the sense they can be made to change readily in response to what is learned from them in terms of the tests themselves or other curriculum elements. In other words, carefully designed achievement procedures are most useful to a language program when they are flexible and responsive for affecting curriculum changes and continually analysing those changes concerning the program realities.

Summary

This chapter will discuss (1) the place of tests in curriculum planning that consists of need analysis, goals and objectives, language testing, material development, language teaching and program evaluation; (2) the place of tests in curriculum implementation that deals with ELI as a language program, initial screening and proficiency procedures, placement

procedures, second-week diagnostic procedures, achievement procedures, and testing as an integrated system.

Brown (1996) describes six broad types of activities that are often identified in the curriculum design in order to promote good teaching and learning: needs analysis, goals and objectives setting, testing (both NRT and CRT), materials development, teaching, and program evaluation.

Needs Analysis

Two dangers, however, may arise in taking needs analysis. Students are people and therefore have needs and concerns that are not solely linguistics. Goals and Objectives

If the purpose of doing needs analysis is to satisfy the language learning needs of the students, one outcome of analysing those needs might be the specification of formal *program goals*.

Language Testing

According to Brown (1996), the next logical step in curriculum development is the drafting of tests based on a program's objectives. The goals, objectives, and administrative necessities of a program may require extensive test development. The tests may have significance. Materials Development

Once tentative needs analysis, objectives, and tests have been put into place, materials can be developed in a rational manner to fit the specific needs and abilities of the participants in the program. With the appropriate testing information in hand, adopting, developing, or adapting materials become relatively easy because the course is fairly clearly defined. For instance, teachers can easily assess existing materials and decide whether these materials could be adopted to fill the needs of the students, or whether they will have to be adapted to meet the students' needs and program goals.

Language Teaching

There is a strength to be found in numbers, so curriculum planners will find it useful to involve teachers, administrators, and students in defining the needs within a particular program and establishing the course

objectives and tests. Teachers are typically required to determine what the students need to learn, define, the course goals and objectives, select or develop course tests and materials, and do the teaching.

Program Evaluation

The evaluation may involve some testing as one source of information, but evaluation is not limited in any way to testing. The Place of Tests in Curriculum Implementation

Initial Screening and Proficiency Procedures

Many teachers may find themselves in a position in which they need *proficiency procedures* to determine how much of a given language their students have learned during their lives.

Placement Procedures

Most teachers will find themselves having to make placement decisions. In most language programs, students are grouped according to ability levels. Placement tests can help teachers to make such decisions. Such tests are typically norm-referenced and therefore fairly general in purpose, but, unlike proficiency tests, placement tests should be designed to fit the abilities and levels of the students in the particular program. The purpose of such tests is to show how much ability, knowledge, or skill the students have. To do this, teachers need tests that are general but designed specifically for the types and levels of their students as well as for the goals of their program. Second-Week Diagnostic Procedures

Achievement Procedures

The tests used to monitor such achievement should be developed to measure the particular objectives of a given course or program.

Comprehension Questions

1. Explain the relationship between testing and curriculum?
2. Where are the positions of tests in curriculum planning?
3. Where is the place of tests in curriculum implementation?

References

Bachman, L.F. 1990. *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

Brown, J.D. 1996. *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Djiwandono, M.Soenardi. 1986. *Kemampuan Berbahasa dan Penilaiannya dalam Pengajaran Bahasa:* Pidato Pengukuhan pada Penerimaan Jabatan Guru Besar IKIP Malang. IKIP Malang.

Djiwandono, M, Soenardi. 1996. *Tes Bahasa dalam Pengajaran.* Bandung; Penerbit ITB.

Ebel, R.L. and Frisbie, D.A. 1986. *Essentials of Educational Measurements.*New Jersey: Prentice Hall, Inc.

Groundlund, Norman E. 1986. *Measurement and Evaluation in Teaching.* New York: Macmillan Publishing Company.

Hasanuddin.2003. *Language Testing.* Gorontalo: IKIP Press.

CHAPTER 5
ADOPTING, DEVELOPING, AND ADAPTING LANGUAGE TESTS

Short Description

Testing or particularly language testing is far more complex than one might indicate. In fact, a number of efforts have been done to develop an effective testing program at their institution. Brown (1996) explores these perspectives as a series of testing *issues*, each of which can be described and thought about separately. Nonetheless, all these matters must be considered simultaneously when adopting, developing, or adapting proficiency, placement, achievement, and diagnostic tests for any language program. Each issue involves one way of characterizing language tests, and taken together; these issues must all be considered in classifying and describing language tests in a state of the art manner. All these issues can be categorised as either theoretical or practical, which will be discussed in the following section.

Basic Competence

Students are able to adopt, develop, and adapt language testing that covers.

- Theoretical issues
- Practical issues
- Interactions of Theoretical and Practical Issues
- Adopt, develop, and adapt language testing

5.1 Theoretical Issues

The theoretical issues that we address have to do with what tests should look like and what they should do. These issues have a great deal to

do with how a group of teachers feels that their course or program fits pedagogically within the overall field of language teaching. Theoretical issues may include (1) pedagogical beliefs in various language teaching methodologies, (2) beliefs in the relative importance of the skills in either receptive or productive skill, (3) linguistic distinction between competence and performance, and (4) test types range from what are called discrete-point to integrative tests and various combinations of the two.

5.1.1 The Language teaching Methodology Issue

Spolsky (1978) Hinofotis (1981) both have pointed out that language testing can be broken down into periods, or trends, of development. Hinofotis labeled them the pre-scientific period, the psychometric/structuralist period, and the integrative/sociolinguistic period. Brown (1996) used the term *movement* instead of periods to describe them because they overlap chronologically and can be said to co-exist today in different parts of the world.

Brown (1996) divided the history of language testing into (1) *the pre-scientific movement,* (2) *the psychometric-structuralist movement, and* (3) *the integrative-sociolinguistic movement.*

The *pre-scientific movement* in language testing is associated with the grammar-translation approaches to language teaching. The pre-scientific movement is characterised by translation and free composition tests developed by the classroom teachers. Nevertheless, this test requires a subjectivity scoring factor due to the difficulty of objective scoring.

The *psychometric-structuralist movement* of language testing worries about the objectivity, reliability, and validity of tests began to arise. Largely because of interaction between linguist and specialists in psychological and educational measurement, language tests became increasingly scientific, reliable, and precise. Psychometric- structuralist tests set out to measure the discrete structure points (Carroll, 1972). As with the language teaching methods, these tests were influenced by behavioral psychology. Psychometric-structuralist tests are usually in a multiple-choice format that is easy to administer and score.

The *integrative-sociolinguistic movement* has its roots in the arguments that language is creative. More precisely, language professionals began to believe that language is more than the sum of discrete parts being tested during the psychometric-structuralist movement. Hymes (1967) states that the development of communicative competence depended on more than simple grammatical control of the language; communicative competence also hinged on knowledge of the language appropriate for different situations. Based on the integrative-sociolinguistic movement, the tests of the future will focus on authentic and purposeful language situations where the student is attempting to communicate some real message. Savignon (1972) labeled this movement in term of the *communicative movement*. Cloze and dictation tests are the part of this movement to assess student's ability to manipulate language within a context of extended text rather than in a collection of discrete-point questions.

5.1.2 Two Skills-based Issues

The subtests on language tests are often separated into skill areas like reading, writing, listening, and speaking. An example of such skills-based subtests is the TOEFL, which currently reports subtest scores for (a) Listening Comprehension, (b) Structure and Written expression, and (c) Vocabulary and Reading Comprehension. Based on the subtests on language tests, Brown (1996) states two important issues namely the *channel issue* and the *mode issue*. He also explains that how skills-based issues interact in channels and modes.

In the channel issue, language teachers and testers can benefit from thinking about such subtests in terms of the channel used for communication – that is, written or oral. For instance, reading and writing subtests can be lumped together and referred to *written channel* subtests since they cover both skills on paper. Listening and speaking subtests, on the other hand, would be labeled *oral channel* subtests because they involve the use of sound to communicate.

In the mode issue, some tests also necessitate the simultaneous use of two skills within a single channel. For instance, an oral interview procedure like the *Interagency Language Roundtable Oral Interview* (ILR 1982) may require the students to understand and produce spoken the language. While the raters consider each skill separately, the net result is a single score that probably reflects some combined rating of both the listening and speaking skills. In such situations, a distinction between productive and receptive modes of communication can be useful. The *productive mode* includes those skills used to send information to others in the form of sound or light waves, those skills used to create the outward manifestations of language by writing or speaking. The *receptive mode* includes those skills that involve receiving and understanding the message from others – that is, reading and listening.

Interactions of skills-based issues in both receptive and productive skills, the same example tests were used to explain channels *and* modes. This is possible because a test necessarily taps at least one channel and one mode at any given time. Thus, reading comprehension tests are typically viewed as receptive mode tests of the written channel. Composition tests also involve the written channel, but they are in the productive mode.

Sometimes tests become even more complex, assessing two modes at the same time, or two channels simultaneously. The possible combinations are obviously numerous. Consider a composition test where students are required to read a two-page academic passage and then analyze it in their written composition. Moreover, a test like dictation is best described as a partially written channel and partially oral channel, as well as the partially receptive mode and partially productive mode. Hopefully, knowing about these issues will help teachers to understand better what they are testing.

5.1.3 The Competence/Performance Issues

Competence and performance in language testing are differently defined. According to Chomsky (1965), competence consists of the mental representation of linguistic rules, which constitute the speaker-hearer's

internalized grammar. The performance consists of the comprehension and production of language. This distinction has some interesting ramifications for language testing. If linguistic performance is viewed as imperfect and full of flaws, even in native speakers, such performance can only be taken as an outward manifestation of the underlying but unobservable linguistic competence.

This distinction can help teachers to realize that tests are at best fairly artificial observations of a student's performance, and performance is only an imperfect reflection of the underlying competence. Since both competence and performance are the interest of language teachers, they must be cautious in their interpretation of test results to remember that performance is only part of the picture – a part is a second-hand observation of competence.

In testing circles, the underlying competence is more often described in terms of a *psychological construct* (Brown, 1996:29). An example of a construct in our field is the notion of overall English as foreign language proficiency. Thus, a student's competence in EFL might more readily be discussed as overall proficiency, which is a psychological construct.

5.1.4 The Discrete-point/Integrative Issue

Traditionally, a discrete point test is one that attempts to focus attention on one point of grammar at a time. Each test item is pointed to one specific grammar component. In addition, Oller (1979) claims "a discrete point test purports to assess only one skill at a time (e.g., listening, or speaking, or reading, or writing). Brown (1996) also states that discrete-point tests are those which measure the small bits and pieces of a language, as in a multiple-choice test made up of questions constructed to measure students' knowledge of different structure. One question on such an ESL test might be written to measure whether the students know the distinction between a and an in English.

Integrative tests, on the other hand, are those designed to use several skills at one time, or more precisely, to employ different channels

and modes of the language simultaneously and in the context of extended text or discourse (Brown, 1996). Also, "integrative tests attempt to assess a learner's capacity to use many bits all at the same time, and possibly while exercising several presumes components of a grammatical system, and perhaps more than one of the traditionally recognized skills or aspects of skills" (Oller, 1979). Consider dictation as a test type. The student is usually asked to listen carefully and write down a short prose passage by the teacher. The skills involved are at least listening to comprehension and writing.

5.2 Practical Issues

The practical issues are related to what we have to do with putting tests into place in a program. Teachers may find themselves concerned with the degree to which tests are fair in terms of objectivity. Alternatively, they may have to decide whether to keep the tests cheap or fight for the resources necessary to do a quality job of testing. Brown (1996) explains three important issues practically. They are (1) the fairness issue, (2) the cost issue, and (3) relevant issue.

The first, *fairness* can be defined as the degree to which a test treats every student the same or the degree to which it is impartial. Teachers would ensure that their personal feelings do not interfere with the fair assessment of the students or bias the assignment of scores. The aim of maximising objectivity is to give each student an equal chance to do well. So teachers and testers often do everything in their power to find test questions, administration procedures, scoring methods, and reporting policies that optimise the chances that each student will receive equal and fair treatment.

The second practical issue is the *cost issue*. In the best of all possible worlds, unlimited time and funds would be available for teaching and testing languages. Unfortunately, this is rarely true. Most teachers are to some degree underpaid and overworked and must continually make decisions that are based on how high some aspect of teaching, or testing, may turn out to be. This issue affects all the self-evident. Lack of funds

can cause the abandonment of otherwise well thought out theoretical and practical positions that teachers have taken.

The third is logistical issues that consist of (1) ease of test construction, (2) ease of test administration, and (3) ease of test scoring. Test construction issue involves the degree to which different types of tests are easy or difficult to produce. Special considerations with regard to test construction can range from deciding how long the test should be to consider what types of questions to use. Ease of test administration is a critical issue because testing is a human activity that is very prone to mix-ups and confusion. The degree to which a test is easy to administer will depend on the amount of time it takes, the number of subtests involved, the amount of equipment and material required to administer it, and the amount of guidance that the students need during the test. Ease of test scoring is an important issue because a test that is easy to score is cheaper and is less likely to result in scorers making simple tallying, counting, and copying mistakes that may affect the students' scores.

5.3 Interactions of Theoretical and Practical Issues

We must stress the importance of recognizing that each of the theoretical and practical issues discussed above can and will interact with all the others – sometimes in predictable patterns and at the other times in unpredictable ways. For instance, if a group of high-school language teachers wants to develop a test that, from a theoretical point of view, is communicative yet integrative and measures productive skills, they may have to accept that the test will be relatively subjective, expensive, and hard to administer and score. If on the other hand, they decide they want a test that is very objective and easy to administer and score, they may have to accept the fact that the questions must be relatively discrete-point so that the answer sheets can be machine-scorable. This decision will naturally result in a test that is less communicative and that focuses mostly on receptive skills.

5.4 Adopt, Develop, or Adapt?

Once a consensus has been achieved as to the purpose and type of test to employ, a strategy must be worked out that maximise the quality and effectiveness of the test that is eventually put into place. Each program would possibly have a testing expert to develop tests especially tailored for that program. In many cases, any rational approach to testing will be a vast improvement over the conditions. Between these two extremes (developing tests from scratch or adopting them from commercial sources on pure faith) is the notion of adapting existing tests and materials so that they better serve the purposes of the program. This section aims to give the teacher a rational basis for adopting, developing, or adapting language tests so that the tests will be maximally useful in their specific language programs.

5.4.1 Adopting Language Tests

The tests that are used in language programs are often adopted from sources outside of the program. This may mean that the tests are bought from commercial publishing houses, adopted from other language programs, or merely taken from the book sources. Given differences that exist among the participants in the various language programs around the world, it is possible that many of the tests acquired from external sources are being used with students that is not the same as expected when the tests were initially developed and standardized. Using tests with the wrong types of students can result in mismatches between the tests and the purposes of the program.

Brown (1996) moreover states other approaches that teachers might want to use to improve their ability to select quality tests for their programs. They would include: (a) informing themselves about language testing through taking a course or reading up on it, (b) hiring a new teacher, who also happens to have an interest in, or already knows about, the subject of testing, and (c) giving one member of the faculty release time to become informed on the topic. The checklist provided in table 2.1

(Brown, 1996:39) should aid in selecting tests that more or less match the purposes of a language program.

In short, there are many factors that must be considered even in adopting an already published test for a particular program. Many of these issues can be addressed by any thoughtful language teacher, but others, such as examining the degree to which the test is reliable and valid, will take more knowledge and experience with language tests.

5.4.2 Developing Language Tests

In an ideal situation, teachers will have enough resources and expertise available in their program that the developed tests of proficiency, placement, achievement and diagnostic are adjusted with the goals of the program and the ability levels and needs of the students.

If a group of teachers decides to develop their tests, they will need to begin by deciding which tests to develop first that would mean developing tests of achievement and diagnosis first because they will tend to be based on the objectives of the particular program. In the interim, while developing these achievement and diagnostic tests, previously published proficiency and placement tests could be adopted as needed. Later, these teachers may wish to develop their own placement test so that the test questions being used to separate students are learning in the program.

5.4.3 Adapting Language Tests

The process of adapting a test to specific situations involves some variant of the following steps:

Administer the test in the particular program, using the appropriate teachers and their students;

Select those test questions that work well at spreading out the students (for NRTs) or that are efficient at measuring the learning of objectives (for CRTs) in the particular program;

Develop a shorter, more efficient revisions of the test – one that fits the program's purposes and works well with its students (some new

questions may be necessary, one similar to those that worked well, in order to have a long enough test); and

Evaluate the quality of the newly revised test (see Table 2.1, Brown, 1996: 39).

With the basic knowledge is provided by Brown (1996) any language teacher can accomplish all these steps. In fact, following the guideline is given in developing and improving test items will enable any teacher to adapt a test to a specific set of program goals and decision-making purposes.

Summary

Testing or particularly language testing is far more complex than that one tradition might indicate. In fact, numerous to develop an effective testing program at their institution. Each issue involves one way of characterizing language tests, and taken together; these issues must all be considered in classifying and describing language tests in a state of the art manner. Theoretical Issues

The Language teaching Methodology Issue

Spolsky (1978) Hinofotis (1981) both have pointed out that language testing can be broken into periods, cr trends, of development. Brown (1996) divided the history of language testing into (1) *the pre-scientific movement,* (2) *the psychometric-structuralist movement, and* (3) *the integrative-sociolinguistic movement.*

The *pre-scientific movement* in language testing is associated with the grammar-translation approaches to language teaching. The pre-scientific movement is characterised by translation and free composition tests developed by the classroom teachers, who are on their own when it comes to developing tests. The *psychometric-structuralist movement* of language testing worries about the objectivity, reliability, and validity of tests began to arise. Psychometric- structuralist tests set out to measure the discrete structure points (Carroll, 1972). As with the language teaching methods, these tests were influenced by behavioral psychology.

Psychometric-structuralist tests are usually in a multiple-choice format that is easy to administer and score.

More precisely, language professionals began to believe that language is more than the sum of discrete parts being tested during the psychometric-structuralist movement. Cloze and dictation tests are the part of this movement to assess the student's ability to manipulate language within a context of extended text rather than in a collection of discrete-point questions.

Two Skills-based Issues

The subtests on language tests are often separated into skill areas like reading, writing, listening, and speaking. Based on the subtests on language tests, Brown (1996) states two important issues namely the *channel issue* and the *mode issue*. In the mode issue, some tests also necessitate the simultaneous use of two skills within a single channel. Interactions of skills-based issues in both receptive and productive skills, the same example tests were used to explain channels *and* modes. In the productive mode, composition tests are also in the written channel.

Sometimes tests become even more complex, assessing two modes at the same time, or two channels simultaneously. Hopefully, knowing about these issues will help teachers to understand better what they are testing.

The Competence/Performance Issues

Competence and performance in language testing are differently defined. This distinction has some interesting ramifications for language testing.

The Discrete-point/Integrative Issue

Consider dictation as a test type. Practical Issues

In the best of all possible worlds, unlimited time and funds would be available for teaching and testing languages. The third is logistical issues that consist of (1) ease of test construction, (2) ease of test administration, and (3) ease of test scoring. Test construction issue involves the degree to which different types of tests are easy or difficult to

produce. Ease of test administration is a significant issue because testing is a human activity that is very prone to mix-ups and confusion. Each program will possibly have a testing expert to develop tests especially tailored for that program. The goal of this section is to provide teachers with a rational basis for adopting, developing, or adapting language tests so that the tests will be maximally useful in their specific language programs.

Adopting Language Tests

The tests that are used in language programs are often adopted from sources outside of the program. Therefore, if the tests given to the students do not match, there will be a disagreement between the tests and purposes of the program.

Developing Language Tests

In the interim, while developing these achievement and diagnostic tests, previously published proficiency and placement tests could be adopted as needed. Later, these teachers may wish to develop their own placement test so that the test questions being used to separate students are learning in the program.

Adapting Language Tests

Administer the test in the particular program, using the appropriate teachers and their students; evaluate the quality of the newly revised test.

Comprehension Questions

1. How to adopt, develop, and adapt language testing in English?
2. What are differences between theoretical issues and practical issues?
3. What are interactions of theoretical and practical issues?
4. Develop and try out and then adopt, develop, and adapt language testing.

References

Bachman, L.F. 1990. *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

Brown, J.D. 1996. *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Djiwandono, M.Soenardi. 1986. *Kemampuan Berbahasa dan Penilaiannya dalam Pengajaran Bahasa:* Pidato Pengukuhan pada Penerimaan Jabatan Guru Besar IKIP Malang. IKIP Malang.

Djiwandono, M, Soenardi. 1996. *Tes Bahasa dalam Pengajaran.* Bandung; Penerbit ITB.

Ebel, R.L. and Frisbie, D.A. 1986. *Essentials of Educational Measurements.*New Jersey: Prentice Hall, Inc.

Groundlund, Norman E. 1986. *Measurement and Evaluation in Teaching.* New York: Macmillan Publishing Company.

Hasanuddin.2003. *Language Testing.* Gorontalo: IKIP Press

CHAPTER 6
DEVELOPING AND IMPROVING TEST ITEMS

Short Description

This chapter deals with (1) what is an item? (2) developing norm-referenced language tests, (3) norm-referenced item statistics, (4) developing criterion-referenced language tests. The information supplied in this chapter will enable teachers to develop, analyze, select, and refine those items most suitable for testing their students – whether their purpose is to develop an NRT for proficiency or placement decisions or a CRT for diagnostic or achievement decisions.

Basic Competence

Students are able to develop and improve test items that cover

- Developing NRT
- Developing CRT
- Norm-referenced item statistics

6.1 What is an Item?

An item is the basic unit of language testing. The item is sometimes difficult to define. Some types of items, like multiple-choice or true-false items, are relatively easy to identify the individual test questions that anyone can recognize as discrete units. Brown (1996) defines the term item very broadly "as the smallest unit that produces distinctive and meaningful information on a test or rating scale."

Item analysis is the systematic evaluation of the effectiveness of the individual items on a test. This is usually done for purposes of selecting the "best" items that remain on a revised and improve version of

the test. Sometimes, however, item analysis is performed simply to investigate how well the items on a test are working with a particular group of students.

6.2 Developing Norm-Referenced Language Tests

6.2.1 Item Format Analysis

In *item format analysis*, testers focus on the degree to which each item is properly written so that it measures all and only the desired content. Such analyses often involve making judgments about the adequacy of item formats. Brown (1996) in table 3.1 shows some *general guidelines*, which apply to most language testing formats. They are in the form of questions that teachers can ask themselves when writing or critiquing any type of item formats. The purpose of asking these questions is to ensure that the students answer the item correctly or incorrectly for the right reasons. Let consider each question in table 3.1 in turn (Brown, 1996:51).

1. Teachers will, of course, want their item formats to match the purpose and content of the item. In part, this means matching the right type of item to what is being tested in terms of channels and modes. For instance, teachers may want to avoid using a multiple-choice format, which is basically receptive (students read and select, but they produce nothing), for testing productive skills like writing and speaking.
2. The issue of making sure that each question has only one correct answer is not as obvious as it might at first seem. Correctness is often a matter of degrees rather than an absolute. An option that is correct to one person may be less so to another, and an option that seems incorrect to the teacher may appear to be correct to some of the students.
3. Each item should be written at approximately the level of proficiency of the students who will take the test.
4. Ambiguous terms and tricky language should be avoided unless the purpose of the item is to test ambiguity. The problem is that

ambiguous language may cause students to answer incorrectly even though they know the correct answer.

5. Likewise, the use of negatives and double negatives may needlessly be confusing and should be avoided unless the purpose of the item is to test negatives.
6. Teachers should also avoid giving clues in one item that will help students to answer another item.
7. All the parts of each item should be on one page.
8. Teachers should also avoid including extra information that is irrelevant to the concept or skill being tested.
9. All teachers should also be on the bias that may have crept into their test items. Race, gender, religion, nationality, and other biases must be avoided at all costs, not only because they are morally wrong illegal in many countries but also they affect the fairness and objectivity of the test.
10. They should always have at least one or more colleagues (who are native speakers of the language being tested) look over and perhaps take the test so that any additional problems may be spotted before the test is used to make decisions about students' lives.

Brown in table 3.2 (1996:54) includes other questions that are specifically designed for *receptive response items*. A receptive response item requires the student to select a response rather than produce one. In other words, the responses involve receptive language in the sense that the item responses from which students must choose are heard or read, receptively. Receptive response item formats include true-false, multiple choice, and matching items.

True-false items are typically written as statements, and students must decide whether the statements are true or false. There are two potential problems that teachers should consider in developing items in this format as follows:

1. The statement should be carefully worded to avoid any ambiguities that might cause the students to miss it for the wrong reasons.

2. Teachers should also avoid absoluteness clues. Absoluteness clues include terms like *all, always, absolutely, never, rarely, most, often* and so forth.

Multiple choice items are made up of an *item stem* or the main part of the item at the top, a *correct answer*, which obviously the choice that will be counted as correct, and the *distractors*, which are those choices that will be counted as incorrect. The term *options* refer collectively to all the choices presented to the students and include the correct answer and the distractors. Five potential pitfalls for multiple choice items appear in table 3.2 (Brown, 1996:54).

Teachers should avoid unintentional clues (grammatical, phonological, morphological, and so forth) that help students to answer an item without having knowledge or skill being tested.

Teachers should also make sure that all the distractors are plausible. If one distractor is ridiculous, that distractor is not helping to test the students. Instead, those students who are guessing will be able to dismiss that distractor and improve their chances of answering the item correctly without actually knowing the correct answer.

To make a test reasonably efficient, teachers should double-check that items contain no needless redundancy.

Any test writer may unconsciously introduce a pattern into the test that will help the students who are guessing to increase the probability of answering an item correctly.

Teachers can also be tempted to use options like "all of the above," "none of the above," and "a. and b. only."

Matching items present the students with two columns of information; finding and classifying matches between the two sets of information are further to be conducted by the students. The information given in the left-hand column will be called the premises and that shown in the right-hand column will be labeled options. Thus, in a matching test, students must match correct option to each premise. There are three guidelines that teachers should apply to matching items as in the following.

1. More options should be supplied than premises so that, students cannot narrow down the choices as they go along by simply keeping track of the options that they have already used.
2. The options should usually be shorter than the premises because most students will read a premise and then search through the options for the correct match.
3. The premises and options should be logically related to one central theme that is evident to the students.

In *productive response items* in table 3.3 includes additional questions that should be applied. Productive response items require the students actually to produce responses rather than just select them respectively. In other words, the responses involve productive language in the sense that the answers must either be written or spoken. Productive item formats include fill-in, short response, and task types of items.

Fill-in-items are those wherein a word or phrase is replaced by a blank in a sentence or longer text, and the student's job is to fill in that missing word or phrase. There are five sets of issues that teachers should consider when using fill-in items.

In answering fill-in items, students will often write alternative correct answers that the teacher did not anticipate when the items were written. To guard against this possibility, teachers should check to make sure that each item has one very concise correct answer.

In deciding how much context to provide for each blank, teachers should make sure that enough contexts has been given that the purpose, or intent, of the item is clear to those students who know the answer.

Generally speaking, all the blanks in a fill-in test should be the same length – that is, if the first blank is twelve spaces long, then all the items should have blanks with twelve spaces.

Teachers should also consider putting the main body of the item before the blank in most of the items so that the students have the information necessary to answer the item once they encounter the blank.

In situations where the blanks may be very challenging and frustrating for the students, teachers might consider supplying a list of responses from which students can choose in filling in the blanks.

Short-response items are usually questions that the students can answer in a few phrases or sentences. This type of question should conform to at least the following two guidelines.

1. Teachers should make sure that the item is formatted so that there is one, and only one, concise answer or set of answers that they are looking for in the responses to each item. The parameters for what will be considered an acceptable answer must be thought through carefully and delineated before correcting such questions.
2. Short-response items should be phrased as clear and direct questions.

Task items are defined here as any of a group of fairly open-ended item types; students are expected to work on a task in the language that is being tested. A task test might include a series of communicative tasks, a set of problem-solving tasks, and a written task (Brown, 1996:60).

While task items are appealing to many language teachers, some complications may arise in trying to use them. To avoid such difficulties, consider at least the following points.

The directions for the task should be so clear that both the tester and the student know exactly what the students must do.

The task should be sufficiently narrow in scope so that it fits logistically into the time allotted for its performance and yet broad enough so that an adequate sample of the student's language use is to obtain for scoring the item properly.

Teachers must carefully work out the scoring procedures for task items for the same reasons listed in discussing the other types of productive response items. The scoring procedures can be an *analytic approach* or a *holistic approach*. A task cab scored using an analytic approach, in which the teachers rate various aspects of each student's language production separately, or a task can be scored using a holistic

approach, in which the teachers use a single general scale to give a global rating for each student's language production.

If teachers decide to use an analytic approach, they must then decide which categories of language to judge in rating the students' performances. Naturally, these decisions that must also occur before the scoring process begins.

Having worked out the approach and categories of language to rate, it is still necessary to define the points on the scales for each category (see analytic scale for Rating Composition tasks in Brown, 1996: 62).

6.3 Norm-Referenced Item Statistics

Two statistical analyses can help in analyzing a set of norm-referenced items: item difficulty analysis and item discrimination analysis. These statistical analyses are only useful insofar as they help teachers to understand and improve the effectiveness of item formats and content.

6.3.1 Item Facility Analysis

Item facility (IF) (also called item difficulty or item easiness) is a statistical index "used to examine the percentage of students who correctly answer a given item" (Brown, 1996:64) Heaton (1988: 178) called item facility in term of facility value or index of difficulty). Locating the IF index is done by adding up the number of students who were able to answer a particular item correctly, and by dividing that sum by the total number of the test-takers. As a formula, it looks like this:

$$\text{IF} = \frac{N_{correct}}{N_{total}}$$

Where

$N_{correct}$ = number of students answering correctly."

N_{total} = number of students taking the test (Brown, 1996:65).

The result of this formula is an item facility value that can range from 0.00 to 1.00 for different items. Teachers can interpret this value as the percentage of correct answers for a given item. For example, the

correct interpretation for an IF index of .27 would be that 27% of the students correctly answered the item. In most cases, an item with an IF of .27 would be a very difficult question because many more students missed it then answered it correctly. On the other hand, an IF of .96 would indicate that 96% of the students answered correctly – a straightforward item because almost everyone responded accurately.

6.3.2 Item Discrimination Analysis

Item discrimination (ID) indicates the degree to which an item separates the students who performed well from those who performed poorly. These two groups are sometimes referred to as the high and low scorers or upper two and lower-proficiency students (Brown, 1996). The reason for identifying these two groups is that ID allows teachers to contrast the performance of the upper-group students on the test with that of the lower-group students. This begins by listing the name of students, their item responses, as well as overall scores in descending order based on the total scores.

The upper and lower groups are sometimes defined as upper and lower third, or 33%. Some test developers will use the upper and lower 27%. Brown (1996) defines the upper and lower groups as some whole number that is roughly 33%.

Once data are sorted into groups of students, calculation of the discrimination indexes is easy. To do this, calculate the facility for the upper and lower groups separately for each item. This is done by dividing the number of students in the upper group; then divide the number who answered correctly in the lower group by the total the IF for the lower group is subtracted from the IF the upper group on each item as follows:

$$ID = IF_{upper} - IF_{lower}$$

Where

ID = item discrimination for an individual item

IF_{upper} = item facility for the upper group on the whole test

IF_{lower} = item facility for the lower group on the whole test

For example, the IF for the upper group on item 4 is 1.00, because everyone in that group answered it correctly. At the same time, the IF for the lower group on that item is.00 because everyone in the lower group answered it incorrectly. Calculating the item discrimination index is done by subtracting the IF for the lower group from the IF for the upper group and got an index of the contrasting performance of students with the high score in the whole test with those who scored low.

6.3.3 NRT Development and Improvement Projects

The development or improvement of a norm-referenced language test is a major undertaking like many other aspects of language curriculum development. Such projects are usually designed to:

1. "pilot a relatively large number of test items on a group of students similar to the group that will ultimately be assessed with the test.
2. analyze the items using format analysis and statistical techniques, and
3. Select the best items to make up a shorter, more efficient revised version of the test". (Adapted from Brown, 1996:69).

Ideal items in an NRT development project have an average IF of .50 and the highest available ID. These ideal items would be considered well-centered – that is, 50% answer correctly and 50% incorrectly. In reality, however, items rarely have an IF of exactly .50. In other words, items in the range between .30 and .70 are typically acceptable. Djiwandono (1996:144) and Ebel (179:267) have the same idea and have suggested the following guidelines for making decisions based on ID:

.40 and up Very good items
.30 to .39 Reasonably good but possibly subject to improvement
.20 to .29 Marginal items, usually needing and being subject to improvement
Below .19 Poor items, to be rejected or improved by revision

Of course, Ebel's guidelines should not be used as hard and fast rules. Instead, these should be used as aids in the processes of decision-

making regarding which items to keep and which to discard. Ones may employ these guidelines until a sufficient number of items has been found to make up whatever norm-referenced test is under development.

6.3.4 Distractor Efficiency Analysis

Further statistical analysis of the different part of each item may help to ensure that they are all functioning well. Recall that the parts of a multiple-choice item include the *item stem*, or the main part of the item at the top, the *options*, which are the alternative choices presented to the student, the *correct answer*, which is option that will be counted as correct, and the *distractors*, which are options that will be counted as incorrect. Also recall that these incorrect options are called distractors because they should be diverted, or pull away, the students from the correct answer if they do not know which is correct. The primary goal of *distractor efficiency analysis* is to examine the degree to which the distractors are attracting students who do not know the correct answer (Brown, 1996). To do this for an item, the percentages of students who chose each option are analyzed. If this analysis can also give the percentages choosing each option in the upper, middle, and lower groups, the information will be even more interesting and useful.

Some other insights can be gained from distractor efficiency statistics which might never have been perceived without them. In item 2, for instance, option c. is the correct answer, with the majority (60%) of the high group choosing that answer. Oddly, the other 40% of the high group selected a wrong answer, option a. In a situation like this, it is important to revert to the original item and scrutinize it from both format and the perspective of the content.

6.4 Developing Criterion-Referenced Language Tests

A central difference between NRTs and CRTs is that NRTs typically produce normal distributions, while CRTs do not necessarily do so. In addition, CRTs may not necessarily produce scores that are normally distributed. In fact, a CRT that is designed to measure students and the

teacher was marvelous; the students would all score 100% on any end-of-course achievement test that was criterion-referenced to measuring that material. Of course, a teacher could create the same effect by writing a final examination that is far too easy for the students.

In fact, the distributions of scores on a CRT may not be normal for either the pretest or the posttest. On ideal CRT designed to test course objectives, all the students would score 0% at the beginning of the course (indicating that they need to learn the material) and 100% at the end of the course (noting that they have all learned the materials). However, in reality, human beings are never perfectly ignorant at the beginning of a course or perfectly knowledgeable at the end.

The CRT development is in much the same sense that they should aim for the normal distribution when they are developing NRTs. Therefore, CRTs deal with (1) item quality analysis, the role of item facility, difference index, the B-index, and CRT item selection (Brown, 1996:75-83).

6.4.1 Item Quality Analysis

The CRT of tests is commonly used for testing achievement and diagnosis, both of which are relatively specific to a particular program. One result of the program-specific nature of CRTs is that the analysis of individual item quality is often crucial. *Item quality analysis* for CRTs ultimately means that judgments must be made about the degree to which the items are valid for the purposes and content of the course or program involved. The first concern in analyzing CRT item quality is with the content of each item. A second consideration is whether the form of each item adequately assesses the desired content. Therefore, the goal of *item content analysis* for a CRT is to answer the question to what extent each item measures the content that is was designed to measure as well as the degree to which that content should be measured at all.

In the end, content analysis inevitably involves some "expert" (for example language teacher or a colleague) who must judge the items. This involves each teacher looking at the test and having some input as to which

item should be kept in the revised version of the test and which should be reworked or thrown out. In some situations, strategies similar to those advocated by Popham (1981) are employed. These strategies include the writing of item specifications based on clearly defined objectives that are judged by teachers as well as by outside and independent reviewers and by examinees.

Item specifications, in Popham's (1981:121-122) terms, are clear item descriptions that include a general description, a sample item, stimulus attributes, response attributes, and specification supplements, which will be defined as follows:

1. *General description*: A brief general description of the knowledge or skills being measured by the item.
2. *Sample item*: An example item that demonstrates the desirable item characteristics.
3. *Stimulus attributes:* A clear description of the stimulus materials – that is, the material that will be encountered by the student.
4. *Response attributes:* A clear description of the types of (a) options from which students will be expected to select their receptive language choices (responses), or (b) standards by which their productive language responses will be judged.
5. *Specification of the supplement*: For some items, the supplemental material will be necessary for clarifying the four previous elements; for example, the specification supplement might include a list of vocabulary items from which the item writer should draw, or a list of grammatical forms, or a list of functions of the language.

The goal of such item specifications is to provide a clear enough description so that any trained item writer using them will be able to generate items very similar to those written by any other item writer.

6.4.2 CRT Development and Improvement Projects

The revision process for NRTs was described earlier as being based on a single administration of the test, which fine because the purpose

of an NRT is usually a single determination of the proficiency or placement of the students in a single population. The piloting of items in a CRT development project is quite different because the purpose of selecting those items is fundamentally different. Since a central purpose of a CRT is to assess how much of an objective or set of objectives has been learned by each student, CRT assessment has to occur before and after instruction in the concepts or skills being taught to determine whether there were any gain in scores. As a result, the piloting of a CRT often involves administering it as a pretest and posttest and comparing results.

6.4.2.1 Role of Item Facility

The resulting CRTs can be administered, and statistical analysis can proceed. As in NRT item analysis, item facility plays an important role. However, two possible item facilities exist for each item – one for pretest and one for the posttest. In CRT development, the goal is to find items that reflect what is being learned, if anything. Hence, an ideal item for CRT purposes in one that has an IF (for the whole group) of .00 at the beginning of instruction and an IF of 1.00 at the end of instruction (Brown, 1996).

Two different strategies can be used to make such a comparison of the performance on the item of those students who have studied the content (posttest) with those who have not (pretest). The first approach is called an *intervention strategy*, which begins by testing the students before instruction in a pretest. The second strategy is the *differential group's strategy*, which begins by finding two groups of students. One group has the knowledge or skills that are assessed on the test and another group that lacks them. The test developer then can compare the item facility indexes of the first group with the item facility indexes for the second group. In either case, the item statistic that the tester calculates to estimate the degree of contrast between the two administrations of the test is called the *difference index*.

6.4.2.2 Difference Index

The difference index (DI not to be confused with ID) indicates the degree to which an item is reflecting a gain in knowledge or skill. The difference indicates the degree to which a CRT item is distinguishing between the students with an in-depth understanding regarding the material and the lesson that is taught and those who do not. To calculate the difference index, the IF for the pretest results is subtracted from the IF for the posttest results. For example, if the post-test IF for item 10 on a test was .77 and the pretest IF was .22, the teacher would know that only 22% knew the concept or skill at the beginning of instruction while 77% knew it by the end. The relatively high DI for that of .77 - .22 = .55 would indicate 55% gain.

6.4.2.3 The B-index

The most straightforward of the indexes is called the B-index. The index refers to a statistical item in comparing the IFs of students who passed a test with the those who are not. IN other words, the masters and non-masters on the test are identified by whether or not they passed the test, and then the B-index indicates the degree to which the masters (students who passed the test in this case) outperformed the non-masters (students who failed the test) on each item.

The B-index can be calculated by using the following simple formula:

$$\text{B-index} = IF_{pass} - IF_{fail}$$

Where

B-index = difference in IF between students who passed and failed a test

IF_{pass} = item facility for students who passed the test

IF_{fail} = item facility for students who failed the test (Brown, 1996:82).

Interpretation of the B-index is similar to that for the difference index (DI). However, the B-index indicates the degree to which an item distinguishes between the students who passed the test and those who

failed rather than contrasting the performance of the students before and after instruction, as is the case with the difference index.

6.4.2.4 CRT Item Selection

Having analyzed the items on a CRT, teachers will ultimately want to revise the tests by selecting and keeping those items that are functioning well for achievement or diagnostic decisions. The item quality analysis can help with the selection process by providing information about how well each item fit the objective being measured and the degree to which that objective fit the course or program involved. Calculating difference indexes and B-indexes provides information about how sensitive each item was instruction and how effective each item was in making the decision about who passed the test and who failed.

In short, the difference index and B-index can help teachers to select that subset of CRT items associated with the instruction and learning in a subject and that subset related to the difference between students who passed or failed the test.

Summary

This chapter deals with (1) what is an item?, (2) developing norm-referenced language tests, (3) norm-referenced item statistics, (4) developing criterion-referenced language tests.

An item is the basic unit of language testing. The item is sometimes difficult to define. *Item analysis* is the systematic evaluation of the effectiveness of the individual items on a test. Sometimes, however, item analysis is performed simply to investigate how well the items on a test are working with a particular group of students.

Developing Norm-Referenced Language Tests

Item Format Analysis

Teachers will, of course, want their item formats to match the purpose and content of the item. Each item should be written at

approximately the level of proficiency of the students who will take the test.

Ambiguous terms and tricky language should be avoided unless the purpose of the item is to test ambiguity. Teachers should also avoid giving clues in one item that will help students to answer another item.

All teachers should also be on the bias that may have crept into their test items. A receptive response item requires the student to select a response rather than produce one. Receptive response item formats include true-false, multiple choice, and matching items.

Five potential pitfalls for multiple choice items appear in table 3.2 (Brown, 1996:54).

Teachers should avoid unintentional clues (grammatical, phonological, morphological, and so forth) that help students to answer an item without having knowledge or skill being tested.

If one distractor is ridiculous, that distractor is not contributing to test the students. Any test writer may unconsciously introduce a pattern into the test that will help the students who are guessing to increase the probability of answering an item correctly.

Matching items present the students with two columns of information; the matches between the sets of information are to be found and identified by the students.

Productive response items require the students actually to produce responses rather than just select them respectively. Productive item formats include fill-in, short response, and task types of items.

In answering fill-in items, students will often write alternative correct answers that the teacher did not anticipate when the items were written. Teachers should also consider putting the main body of the item before the blank in most of the items so that the students have the information necessary to answer the item once they encounter the blank.

Short-response items are usually questions that the students can answer in a few phrases or sentences. Short-response items should be phrased as clear and direct questions.

Task items are defined here as any of a group of fairly open-ended item types; furthermore, students are required to work on a task the test language. Teachers must carefully work out the scoring procedures for task items for the same reasons listed in discussing the other types of productive response items. Norm-Referenced Item Statistics

Two statistical analyses can help in analyzing a set of norm-referenced items: item difficulty analysis and item discrimination analysis.

Item Facility Analysis

Item facility (IF) (also called item difficulty or item easiness) is a statistical index applied to scrutinize the percentage of students who can answer a given item correctly (Brown, 1996:64) Heaton (1988: 178) called item facility in term of facility value or index of difficulty). Adding up the number of students who can answer a particular item correctly, and dividing this sum by the total test takers are done to locate the IF index.

$$\text{IF} = \frac{N_{correct}}{N_{total}}$$

Where

$N_{correct}$ = number of students answering correctly

N_{total} = number of students taking the test (Brown, 1996:65).

The result of this formula is an item facility value that can range from 0.00 to 1.00 for different items. Teachers can interpret this value as the percentage of correct answers for a given item. For example, the correct interpretation for an IF index of .27 would be that 27% of the students correctly answered the item. Item Discrimination Analysis

Item discrimination (ID) indicates the degree to which an item separates the students who performed well from those who performed poorly. The reason for identifying these two groups is that ID allows teachers to contrast the performance of the upper-group students on the test with that of the lower-group students.

Where

ID = item discrimination for an individual item

IF_{upper} = item facility for the upper group on the whole test

IF_{lower} = item facility for the lower group on the whole test

For example, the IF for the upper group on item 4 is 1.00, because everyone in that group answered it correctly. Pilot a relatively large number of test items on a group of students similar to the group that will ultimately be assessed with the test, analyze the items using format analysis and statistical techniques, and .40 and up Very good items.

.20 to .29 Marginal items, usually needing and being subject to improvement

To do this for an item, the percentages of students who chose each option are analyzed.

In item 2, for instance, option c. is the correct answer, with the majority (60%) of the high group choosing that answer.

Developing Criterion-Referenced Language Tests

Item Quality Analysis

The first concern in analyzing CRT item quality is with the content of each item. Item specifications, in Popham's (1981:121-122) terms, are clear item descriptions that include a general description, a sample item, stimulus attributes, response attributes, and specification supplements, which will be defined as follows:

Sample item: An example item that demonstrates the desirable item characteristics.

The goal of such item specifications is to provide a clear enough description so that any trained item writer using them will be able to generate items very similar to those written by any other item writer.

The piloting of items in a CRT development project is quite different because the purpose of selecting those items is fundamentally different.

As in NRT item analysis, item facility plays an important role. However, two possible item facilities exist for each item – one for pre-test and one for the posttest.

The test developer then can compare the item facility indexes of the first group with the item facility indexes for the second group. Difference Index

The *B-index* refers to a statistical item that compares the IFs of students who ace a test with the IFs of students who unable to pass the test. Where B-index = difference in IF between students who passed and failed a test

IF_{pass} = item facility for students who passed the test
IF_{fail} = item facility for students who failed the test (Brown, 1996:82).

CRT Item Selection. Having analyzed the items on a CRT, teachers will ultimately want to revise the tests by selecting and keeping those items that are functioning well for achievement or diagnostic decisions.

Comprehension Questions and Tasks

1. How to develop and improve NRT?
2. How to develop and improve CRT?
3. Find out the result of tests at junior and senior high school and analyze their item facility, item discrimination, and their distractor efficiency.

References

Bachman, L.F. 1990. *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

Brown, J.D. 1996. *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Djiwandono, M.Soenardi. 1986. *Kemampuan Berbahasa dan Penilaiannya dalam Pengajaran Bahasa:* Pidato Pengukuhan pada Penerimaan Jabatan Guru Besar IKIP Malang. IKIP Malang.

Djiwandono, M, Soenardi. 1996. *Tes Bahasa dalam Pengajaran.* Bandung; Penerbit ITB.

Ebel, R.L. and Frisbie, D.A. 1986. *Essentials of Educational Measurements.*New Jersey: Prentice Hall, Inc.

Groundlund, Norman E. 1986. *Measurement and Evaluation in Teaching.* New York: Macmillan Publishing Company.

Hasanuddin.2003. *Language Testing.* Gorontalo: IKIP Press

CHAPTER 7
DESCRIBING TEST RESULTS

Short Description

The purpose of describing the results of a test is to provide test developers and test users with a picture of students' performance. This chapter deals with (1) four different scales of measurement, (2) displaying data, (3) statistics for describing the central tendency, (4) dispersion, and reporting descriptive statistics.

Basic Competence

Students are able to describe test results that cover

- Four different scales of measurement
- Displaying data
- Statistics for describing the central tendency and dispersion

7.1 Scales of Measurement

Brown (1996) displays four types of scales that appear in the language teaching literature. The four scales represent four different ways of observing, organising, and quantifying language data. The four scales are the nominal, ordinal, interval, and ratio scales.

A *nominal scale* is used for categorising and naming groups. Most language teaching professionals will be interested in identifying groups into which language students might fall. Some of the most common categories or groupings are according to gender, nationality, native language, educational background, socioeconomic status, the level of language study, membership in a particular language class, and even whether or not the students say that they enjoy language study. The

essence of the nominal scale is that it names independent categories which people (or other living things or objects) can be classified.

An *ordinal scale* names a group of observations, but, as its label implies, an ordinal scale also orders, or ranks, the data. For instance, if we want to rank the students from best to worst in some ability based on a test, arrange the students' scores from high to low using ordinal numbers than simply rank the students.

An *interval scale* also represents the ordering of a named group of data, but it provides additional information. As its name implies, an interval scale also shows the intervals, or distances, between the points in the rankings. Also, the distances between some of the average scores are only one point each. In short, interval scales contain information about the distances between students' scores, which is missing on ordinal and nominal scales.

A *ratio scale* also represents the ordering of a named group of data and shows the distances between the points in the rankings, but it provides additional information. First, a ratio scale has a zero value; and second, as the name implies, the point on the scale is precise multiples, or ratios, of other points on the scale. For instance, if the lights in a room are turned off, there is zero electricity flowing through the wires.

7.2 Displaying Data

One way of displaying data is frequency. Frequency is the term that is used to describe this very common-sense sort of tallying procedure. Frequency can be used to indicate how many people did the same thing on a certain task, or how many people have a certain characteristic, or how many people fall into a certain set of categories. This frequency is particularly useful when dealing with a nominal scale. Frequencies are valuable because they can summarise data and thereby reveal patterns that might not otherwise be noticed. For instance, frequency distribution displays the frequency of each score value arranged from high to low scores.

Frequency data can be shown in far more graphic and appealing ways than the plain, ordinary frequency distribution shown in table 4.4 (Brown, 1996:100). Such graphic displays of scores generally come in one of three forms: a histogram, a bar graph, or a frequency polygon. A *histogram* of the frequencies of a set of scores is normally displayed by assigning score values to the horizontal line. If bars are drawn instead of Xs to represent the score frequencies, the result is a *bar graph*. Likewise, when dots are placed where the top X would be at each score value and are then connected by lines, the result is a *frequency polygon*. Thus, understanding how graphs work can help teachers to defend their program successfully against serious external misrepresentations about enrollments, budgets, teaching loads, and so forth.

At a minimum, teachers should examine the descriptive statistics whenever they administer a test. Descriptive statistics are numerical representations of how a group of students performed on a test (Brown, 1996:102). Two aspects of group behaviour are considered in descriptive statistics: the middle of the group and individuals, which are in statistical terms that are called *central tendency* and *dispersion.*

7.3 Central Tendency

Central tendency is the first aspect of a test to consider. Central tendency describes the most typical behaviour of a group. Four statistics are used for estimating central tendency: the mean, the mode, the median, and the midpoint (Brown, 1996: 102-105).

The *mean* score of any test is the arithmetical average: i.e. the sum of separate scores divided by the total number of testees (Heaton, 1988). The mean is symbolised in writing by X said 'ex-bar". Another way to define a statistical concept is to give its formula as in the following.

$$\bar{X} = \frac{\sum X}{N}$$

Where $\bar{X}$ = mean, X = scores, N = number of scores, and Σ = sum.

Example, if we know, the sum of scores is 702, and the number of students is 26 so that 702 divided by 26 is 27.

The mode is another indicator of central tendency. The mode is that score which occurs most frequently. In Table 4.5 (Brown, 1996:103) shows that the mode would be 69, the only score received by four students. Thus, the mode would be that score, which is most fashionable, or the one received by the most students.

The *median* is that point below which 50% of the scores fall and above which 50% fall. Thus, in the set of scores 100, 95, 83, 71, 61, 57, 30, the median is 71, because 71 has three scores above it (100, 95, and 83) and three scores below it (61, 57, and 30). What is the median of the following set of scores: 11, 23, 40, 50, 57, 63, 86? Fifty, right?

In a real data, cases arise that are not so clear. For example, what is the median for these scores: 9, 12, 15, 16, 17, 27? In such a situation when there is an even number of scores, the median is taken to be midway between the two middle scores. In this example, the two middle scores are 15 and 16, so the median is 15.5.

The midpoint in a set of the score is that point halfway between the highest score and the lowest score on the test. The formula for calculating the midpoint is:

$$\text{Midpoint} = \frac{High+Low}{2}$$

For example, if the lowest score on a test was 30 and the highest was 100, the midpoint would be halfway between these two scores. To use the formula: (a) identify the high and the low score (100 and 30 here), (b) add the low score to the high one (100 + 30), and (c) divide the result by 2 as follows:

$$\text{Midpoint} = \frac{100+30}{2} = \frac{130}{2} = 65$$

7.4 Dispersion

With a clear understanding of how to examine the central tendency of a set of scores in hand, the next step is to consider dispersion, or how the individual performances vary from the central tendency. Three indicators of the dispersion are commonly used for describing distributions of test scores: the range, the standard deviation, and the variance.

7.4.1 Range

Most teachers are already familiar with the concept of range from tests that they have given in class. The range is the number of points between the highest score on a measure and the lowest score plus one (one is added because the range should include the scores at both ends). For instance, if the highest score is 77 and the lowest is 61, the range 17 points (77 – 61 + 1 = 17). The range provides some idea of how individuals vary from the central tendency.

However, the range only reflects the magnitude of the outer edges (high and low) of all the variation in scores and therefore can be strongly affected by any test performance which is not really representative of the group of the students as a whole. For instance, if we add student named Ani, who scored 26, the range will be much larger than 17. With Ani included, the range is 52 (77 – 26 + 1 =52). However, her performance on the test is so different from the performances of the other students that she does not appear to belong to the group.

7.4.2 Standard Deviation

The standard deviation is an averaging process; as such, it is not affected as much by outliers as the range. Consequently, the standard deviation is generally considered a stronger estimate of the dispersion of scores. Brown (1996:107) defines standard deviation as a sort of average of the differences of all scores from the mean. The formula for the standard deviation (S, s or S.D,) is:

$$S = \frac{\sqrt{\Sigma(X-X)}}{N}$$

Starting from the inside and working outward, subtract the mean from each score (X – X), square each of these values (X – X), and add them up $\Sigma\,(X-X)$. This sum is then divided by the number of scores $\Sigma\,(X-X)/N$ and the square root of the result of that operation is the standard deviation.

Using the same scores and mean, table 4.6 (Brown, 1996:108) illustrates the steps required to calculate the standard deviation: (a) line up each score with the mean; (b) subtract the mean from each score; (c) each of the differences from the mean is squared; (d) the squared values are added up; and (e) the appropriate values can be inserted into the formula.

7.4.3 Variance

The variance is another descriptive statistic for dispersion. As indicated by its symbol, S, the test variance is equal to the squared value of the standard deviation. Thus the formula for the test variance looks very much like the one for the standard deviation except that both sides of the equation are squared that can be formulated as follows:

$$S^2 = \frac{\Sigma\,(X-\overline{X})^2}{N}$$

Hence, test variance can be defined as the average of the squared differences of students' scores from the mean. Test variance can also be defined as the square of the standard deviation, or as an intermediary steps in the calculation of the standard deviation.

Summary

The four scales are the nominal, ordinal, interval, and ratio scales.

A *nominal scale* is used for categorising and naming groups. In short, interval scales contain information about the distances between students' scores, which is missing on ordinal and nominal scales.

Displaying Data

One way of displaying data is frequency. For instance, frequency distribution shows the frequency of each score value arranged from high to low scores.

A *histogram* of the frequencies of a set of scores is normally displayed by assigning score values to the horizontal line. If bars are drawn instead of Xs to represent the score frequencies, the result is a *bar graph*. Descriptive statistics are numerical representations of how a group of students performed on a test (Brown, 1996:102). 7.3 Central Tendency

The *mean* score of any test is the arithmetical average: i.e. the sum of separate scores divided by the total number of testees (Heaton, 1988). Where $\bar{X}$ = mean, X = scores, N = number of scores, and Σ = sum.

The mode is that score which occurs most frequently. In table 4.5 (Brown, 1996:103) shows that the mode would be 69, the only score received by four students. For example, what is the median for these scores: 9, 12, 15, 16, 17, 27?

The midpoint in a set of the score is that point halfway between the highest score and the lowest score on the test.

$$\text{Midpoint} = \frac{High+Low}{2}$$

For example, if the lowest score on a test was 30 and the highest was 100, the midpoint would be halfway between these two scores. 7.4 Dispersion

Three indicators of the dispersion are commonly used for describing distributions of test scores: the range, the standard deviation, and the variance.

Range

The range is the number of points between the highest score on a measure and the lowest score plus one (one is added because the range should include the scores at both ends). For instance, if the highest score is 77 and the lowest is 61, the range 17 points (77 – 61 + 1 = 17).

For instance, if we add student named Ani, who scored 26, the range will be much larger than 17.

Standard Deviation

Consequently, the standard deviation is considered a stronger estimate of the dispersion of scores. Brown (1996:107) defines standard deviation as a sort of average of the differences of all scores from the mean.

Variance

This chapter deals with (1) four different scales of measurement, (2) displaying data, (3) statistics for describing the central tendency, (4) dispersion, and reporting descriptive statistics.

Basic Competence

Students are able to describe test results that cover

- Four different scales of measurement
- Displaying data
- Statistics for describing the central tendency and dispersion

Comprehension Questions

1. Explain four different scales of measurement.
2. How to display and describe the central tendency?
3. How to use statistics for describing the central tendency and dispersion?

References

Bachman, L.F. 1990. *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

Brown, J.D. 1996. *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Djiwandono, M.Soenardi. 1986. *Kemampuan Berbahasa dan Penilaiannya dalam Pengajaran Bahasa:* Pidato Pengukuhan pada Penerimaan Jabatan Guru Besar IKIP Malang. IKIP Malang.

Djiwandono, M, Soenardi. 1996. *Tes Bahasa dalam Pengajaran.* Bandung; Penerbit ITB.

Ebel, R.L. and Frisbie, D.A. 1986. *Essentials of Educational Measurements.*New Jersey: Prentice Hall, Inc.

Groundlund, Norman E. 1986. *Measurement and Evaluation in Teaching.* New York: Macmillan Publishing Company.

Hasanuddin.2003. *Language Testing*. Gorontalo: IKIP Press

ELT Assessment

CHAPTER 8
INTERPRETING TEST SCORES

Short Description

The purpose of developing language tests, administering and sorting them through the resulting scores is to make decisions about the students' performance. The sorting process is sometimes called *test score interpretation*. This chapter is about interpreting the performances of students on both norm-referenced and criterion-referenced tests. As a foundation, the discussion begins with the three concepts: (1) probability distributions, (2) the normal distribution, and (3) standardised scores.

Basic Competence

Students are able to interpret test score that covers

- Probability distribution
- Normal distribution
- Standardized scores

8.1 Probability Distributions

Probability is determined by a dividing the number of expected outcomes (one -head of this case) by the number of possible outcomes (two – both heads and tails are possibilities). In the coin of a coin flip, one expected outcome is divided by the number of possibilities to yield ½, or .50, which indicates a 50% probability of getting heads on any particular flip of a coin (Brown, 1996).

Since probability is a function of expected outcomes and possible outcomes. Expected outcomes represent those events for which a person is trying to determine the probability. The possible outcomes account for the

number of potentially different events that might occur as the events, is the ratio of the expected outcomes of the possible outcomes. This ranges from 0 to 1.0 and is commonly discussed in percentage terms. Thus, a ratio of .50 is also referred to as a 50% chance of getting heads.

Another way of keeping track of probabilities is to plot them out as they occur, perhaps in the form of a histogram. Typically, a histogram is designed so that the number of actual outcomes is on the ordinate and possible outcomes are on the abscissa. The result of plotting the coin flips as they occurred is a graph of the distribution. This distribution helps us to picture the events that took place more vividly than simply knowing the numbers (three heads and five tails).

8.2 Normal Distribution

The normal distribution does occur. The graphs of the coin flip distributions demonstrate that. Moreover, as the number of possible events gets larger, plots of those events increasingly take the shape of the bell curve. Additional evidence comes from the biological sciences, where the repeated observations show that living organisms grow, multiply, and behave in relatively predictable patterns. Many of these patterns take and shape of the *normal distribution*. For example, scores measuring the language performance of students, perhaps on a 100-point test, as shown in Figure 5.3 (brown, 1996:125) describes that their scores look reasonably normal, a distribution that is quite common among language students. Similar distributions would likely occur in graphs of their ages, their height, and their IQ scores as well.

The criterion-referenced decision making may be almost entirely independent of the normal distribution. Nonetheless, plotting the CRT scores of a group of students can never hurt. While CRT distributions are often quite different from NRT distributions, inspecting them can provide as much information about the CRT involved as the normal distribution does about NRTs.

So, to the surprise of many teachers, the normal distribution of scores, or something close to it, actually does occur if the purpose of the

test is norm-referenced and the number of students is sufficiently large. Hence, teachers should never dismiss out of hand the idea of the normal distribution. With a group of, say, 160 students taking the Hypothetical Language Test that Brown (on Figure 5.4: 128) illustrates a normal distribution pattern that occurs and recurs in nature as well as in human behaviour. More importantly, this pattern can aid in sorting out the test performance of language students.

8.3 Characteristics of Normal Distributions

The two most important features of a normal distribution are *central tendency* and *dispersion*. A third useful characteristic is the notion of a percent in the distribution. One way this concept can be helpful is in exploring the percent of students who fall within different score ranges on a test.

Recall that *central tendency* indicates the typical behaviour of a group and that four different estimates can be used: the mean, mode, median, and midpoint. All four of these estimate should be somewhere near the centre or middle if a distribution is normal. In fact, in a perfectly normal distribution, all four indicators of central tendency would fall on the same score value, right in the middle of the distribution.

Dispersion is predictable in a normal distribution. Dispersion describes how individual scores disperse, or vary, around the central tendency. This concept is commonly estimated statistically by using the range and standard deviation. The standard deviation is a normal distance measured in score points that mark off certain portions of the distribution, each of which is equal in length along the abscissa.

One central tendency and dispersion are understood as they apply to the normal distribution, some inferences can be made about the *percent* of students who likely to fall within certain score ranges in the distribution. First, recall that the mean, mode, and midpoint should all be the same in a normal distribution. Also recall that the median is the score below which 50% of the cases should fall, and of which 50% should be. Given these facts, teachers can predict with fair assurance that 50% of their students'

scores will be above the median (or mean, or mode, or midpoint) in a normal distribution. In like manner, researchers have repeatedly shown that approximately 34% of the scores will fall within one standard deviation above the mean, as shown in Figure 5.5 (Brown, 1996:130). That means that about 34% of the students scored 41 and 51 points on this particular test.

8.4 NRT and CRT Distribution

The discussion of the normal distribution and standardised scores applies to interpreting the results of norm-referenced proficiency or placement tests. The decisions based on NRTs are called ***relative decisions*** and that the interpretation of the scores focuses on the relative position of each student vis-a-vis the rest of the students about some general ability. Thus, the normal distribution and each student's position in that distribution, as reflected by his or her percentile or standardised score, make sense as viable tools for score interpretation.

Recall also that interpreting the results of criterion-referenced diagnostic and achievement tests is entirely different. CRT decisions are labelled ***absolute*** because they focus not on the student's position relative to other students but rather on the percent of the material that each student knows, largely without reference to the other students. Thus, at the beginning of a course, the distribution of scores on a CRT is likely to be positively skewed if the students need to learn the material covered in the course.

Item selection for CRTs involves retaining those items that students answer poorly at the beginning of the course and answer well at the end of instruction. This pattern shows up in the Ifs on the pretest and post-test as well as in the difference index (DI). The result of revising the CRTs on the basis of these item statistics is usually a magnification of any existing differences between the pretest and post test distribution. So certain conditions exist under which a skewed distribution is not only desirable but also something that testers may aim for revising their CRTs.

Summary

The purpose of developing language tests, administering them, and sorting through the resulting scores is to make decisions about the students. The sorting process is sometimes called *test score interpretation*. As a foundation, the discussion begins with the three concepts: (1) probability distributions, (2) the normal distribution, and (3) standardised scores.

Normal Distribution

The normal distribution does occur. The graphs of the coin flip distributions demonstrate that. For example, scores measuring the language performance of students, perhaps on a 100-point test, as shown in Figure 5.3 (brown, 1996:125) describes that their scores look reasonably normal, a distribution that is quite common among language students. Similar distributions would likely occur in graphs of their ages, their height, and their IQ scores as well.

The criterion-referenced decision making may be almost entirely independent of the normal distribution. Nonetheless, plotting the CRT scores of a group of students can never hurt. While CRT distributions are often quite different from NRT distributions, inspecting them can provide as much information about the CRT involved as the normal distribution does about NRTs.

So, to the surprise of many teachers, the normal distribution of scores, or something close to it, actually does occur if the purpose of the test is norm-referenced and the number of students is sufficiently large. Characteristics of Normal Distributions

The two most important characteristics of a normal distribution are *central tendency* and *dispersion*. *Dispersion* is predictable in a normal distribution. One central tendency and dispersion are understood as they apply to the normal distribution, some inferences can be made about the *percent* of students who likely to fall within certain score ranges in the distribution. NRT and CRT Distribution

The discussion of the normal distribution and standardised scores applies to interpreting the results of norm-referenced proficiency or placement tests

Comprehension Questions and Tasks

1. What are differences between probability distribution? Give an example of each distribution.
2. What are the characteristics of normal distribution?
3. Find out the result of tests English at Senior High School and analyse their probability distribution and normal distribution.

References

Bachman, L.F. 1990. *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

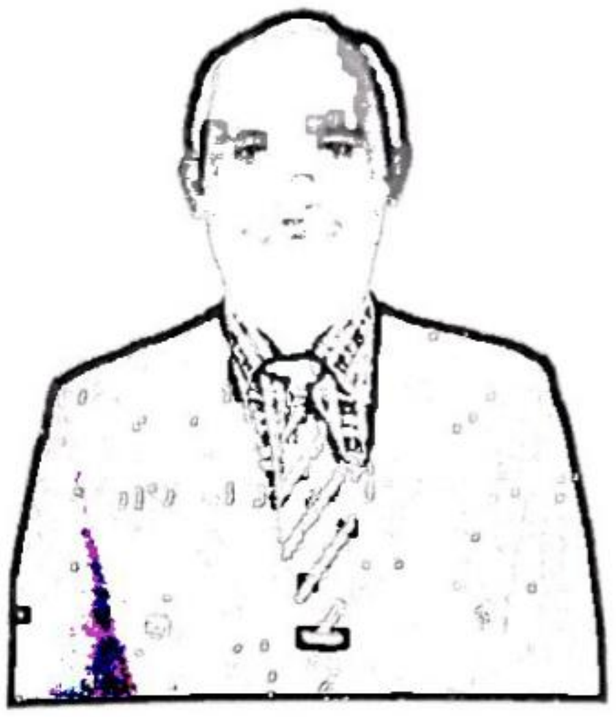
Brown, J.D. 1996. *Testing in Language Programs.* New Jersey: Prentice Hall Regents.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Djiwandono, M.Soenardi. 1986. *Kemampuan Berbahasa dan Penilaiannya dalam Pengajaran Bahasa:* Pidato Pengukuhan pada Penerimaan Jabatan Guru Besar IKIP Malang. IKIP Malang

Hasanuddin.2003. *Language Testing.* Gorontalo:IKIP Press.

ABOUT THE AUTHORS



Prof. Dr. Hasanuddin is a Professor in English Education Faculty of Letters and Culture State University of Gorontalo Gorontalo Province Sulawesi Indonesia. He became a lecturer at English Education since 1988. He enrolled his undergraduate studies at English Department Faculty of Letters Hasanuddin University Makassar and gained Sarjana Degree (Drs) in 1986. He continued his study at Master Degree at English Language Studies at Hasanuddin University in 1993 and finished his study in 1995. For the next opportunity, in 1997 he continued his study at Doctoral program (Ph.D) at State University of Malang East Java Indonesia in 1997 and graduated in 2001. He has followed elearning training and management in Bonn Germany in 2002. He has become university management, 2002 to 2004 he was the vice coordinator of academic affair post graduate program at State University of Gorontalo, 2005-2009 the chairman of Language Department Post Graduate program State University of Gorontalo, 2010 - 2012 Vice Director of Academic Affair Post Graduate Program and 2012 until today he is the vice Rector of Planning, Cooperation and Information System Sate University of Gorontalo. He has become a presenter in some national and international seminar and conferences in linguistics, English language teaching and ICT in education. He has already published book on titles: Language Testing, 2005, Teaching English as a Foreign Language, 2014. Applied Linguistics, 2014, some other books and articles related to Linguistics and English language teaching.

Dr. Hj. Rasuna Talib is a lecturer in English Education Faculty of Letters and Culture State University of Gorontalo Gorontalo Province Indonesia. She became a lecturer at English Education since 1994. He enrolled his undergraduate studies at Language and Art Department, English Study Program, FKIP UNSRAT Manado and gained Sarjana Degree (Dra) in 1992. She continued her study at Master Degree at English Language Studies at Hasanuddin University in 1997 and finished her study in 2000. She continued her study at Doctoral program (Ph.D) at State University of Jakarta Indonesia in 2009 and graduated in 2012. She followed Short Training lesson Study (STOLS) for ITTEPT in Japan, collaboration with JICA and DIKTI(2015), Up-Grading Adjudicator of English Debate 2009; participant of TOT Active Learning Order Thinking (ALIHE) at Makasar and Surabaya (2012). She was the secretary of English Department (2000-2002); the head of English Department Faculty of Letters and Culture the State University of Gorontalo (2002-2005); The vice dean of students coordinator at Faculty of Letters and Culture, the State University of Gorontalo, (2015-2010); the chairman of Post Graduate English Study program, State University of Gorontalo (2015-now). She has become a presenter in some national and international seminar and conferences in applied linguistics, English language teaching and Education. She has already published book related to English Language Teaching and Education: English for University Students (A Handbook of Activities & Classroom Teaching Language Testing (2015); Bahasa Sastra dan pembelajarannya, (part of chapter) Evaluasi pembelajaran; Bahasa Model Smart Regional, Nasional, dan Global -Pengembangan Karakter Akademika Berbasis Disiplin Ilmu, (part of chapter) Model Pendikar Berbasis Ilmu Sastra dan Budaya.

Prof. Dr. Hasanuddin is a Professor in English Education Faculty of Letters and Culture State University of Gorontalo Gorontalo Province Sulawesi Indonesia. He became a lecturer at English Education since 1988. He enrolled his undergraduate studies at English Department Faculty of Letters Hasanuddin University Makassar and gained Sarjana Degree (Drs) in 1986. He continued his study at Master Degree at English Language Studies at Hasanuddin University in 1993 and finished his study in 1995. For the next opportunity, in 1997 he continued his study at Doctoral program (Ph.D) at State University of Malang East Java Indonesia in 1997 and graduated in 2001. He has followed elearning training and management in Bonn Germany in 2002. He has become university management, 2002 to 2004 he was the vice coordinator of academic affair post graduate program at State University of Gorontalo, 2005-2009 the chairman of Language Department Post Graduate program State University of Gorontalo, 2010 -2012 Vice Director of Academic Affair Post Graduate Program and 2012 until today he is the vice Rector of Planning, Cooperation and Information System Sate University of Gorontalo. He has become a presenter in some national and international seminar and conferences in linguistics, English language teaching and ICT in education. He has already published book on titles: Language Testing, 2005, Teaching English as a Foreign Language, 2014. Applied Linguistics, 2014, some other books and articles related to Linguistics and English language teaching.

Dr. Hj. Rasuna Talib is a lecturer in English Education Faculty of Letters and Culture State University of Gorontalo Gorontalo Province Indonesia. She became a lecturer at English Education since 1994. He enrolled his undergraduate studies at Language and Art Department, English Study Program, FKIP UNSRAT Manado and gained Sarjana Degree (Dra) in 1992. She continued her study at Master Degree at English Language Studies at Hasanuddin University in 1997 and finished her study in 2000. She continued her study at Doctoral program (Ph.D) at State University of Jakarta Indonesia in 2009 and graduated in 2012. She followed Short Training lesson Study (STOLS) for ITTEPT in Japan, collaboration with JICA and DIKTI(2015), Up-Grading Adjudicator of English Debate 2009; participant of TOT Active Learning Order Thinking (ALIHE) at Makasar and Surabaya (2012). She was the secretary of English Department (2000-2002); the head of English Department Faculty of Letters and Culture the State University of Gorontalo (2002-2005); The vice dean of students coordinator at Faculty of Letters and Culture, the State University of Gorontalo, (2015-2010); the chairman of Post Graduate English Study program, State University of Gorontalo (2015-now). She has become a presenter in some national and international seminar and conferences in applied linguistics, English language teaching and Education. She has already published book related to English Language Teaching and Education: English for University Students (A Handbook of Activities & Classroom Teaching Language Testing (2015); Bahasa Sastra dan pembelajarannya, (part of chapter) Evaluasi pembelajaran; Bahasa Model Smart Regional, Nasional, dan Global -Pengembangan Karakter Akademika Berbasis Disiplin Ilmu, (part of chapter) Model Pendikar Berbasis Ilmu Sastra dan Budaya.

Penerbit Deepublish (CV BUDI UTAMA)
Jl. Rajawali, Gang Elang 6 No.3, Drono, Sardonoharjo, Ngaglik, Sleman
Jl. Kaliurang Km 9,3 Yogyakarta 55581
Telp/Fax : (0274) 4533427
Anggota IKAPI (076/DIY/2012)
cs@deepublish.co.id @penerbitbuku_deepublish
Penerbit Deepublish www.penerbitbukudeepublish.com

Kategori : English

ISBN 602453559-9