THE EFFECT OF TEXT-TYPE MIXING ON UNIVERSITY AGE STUDENTS

by

Julianne Wellinger

B.A., University of British Columbia, 1983

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTERS OF ARTS

in the Faculty
of
Education

SIMON FRASER UNIVERSITY

March 1994

APPROVAL

Name: Julianne Wellinger

Degree: Master of Arts

Title of Thesis: The Effect of Text-type Mixing on University Age Students

Examining Committee:

Chair: Geoffrey Madoc-Jones

Gloria Sampson
Senior Supervisor

Leone Prock
Associate Professor

Mary Kooy
Limited Term Faculty
Faculty of Education
Simon Fraser University
External Examiner

Date Approved MAR. 17, 1994

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of Thesis/Project/Extended Essay

The Effect of Text-Type Mixing on University Age Students

Author:

(Signature)

Julianne Wellinger

(Name)

March 17, 1994

(Date)

Abstract

The present study examined text-type mixing and its effect on reading comprehension. First, a quantitative approach that identifies groups of co-occurring linguistic features within a text was used to linguistically analyze three texts (each 7266 to 13636 words in length) by hand. These texts with similiar content were taken from three commonly used undergraduate textbooks. The frequency counts of the linguistic features within the texts were normalized to a text length of 1000 words to allow a comparison of frequency counts.

Biber in *Variation Across Speech and Writing* analyzed the co-occurrence frequencies of 67 syntactic features in a one-million word corpus. A factor analysis revealed that the cluster of features consisting of nouns, word length, prepositional phrases, type/token ratio, and attributive adjectives was specific to texts that had a high informational focus. The three texts used in this study were analyzed following Biber's model. These texts too had high frequency counts within the above cluster of features.

To test the effects of text-type mixing on reading comprehension, three groups of 15 first and second year university students, a total of 45 students, were given one of three passages that were developed containing common content from the above analyzed texts. The propositional content of each passage (approximately 5 pages long) was identical and identically structured.

Passage I followed the linguistic conventions of the academic prose genre or text-type while both Passages II and III mixed linguistic dimensions of a more colloquial representation of English with those of academic prose. Passage II situated these occurrences of combined conventions of two genres within the topic sentence. Passage III, on the other hand, situated these occurrences within secondary sentences of the text. After the subjects read one of the passages they answered 20 multiple choice questions taken from a test item file specifically designed for one of the textbooks used in the linguistic analysis above.

Although the mean of the comprehension test results for the standardized text-type passage was higher than those for the mixed text-type passages, and the mean of the comprehension test results for those students having read Passage II was, in turn, higher than that for those having read Passage III; mixing text-types did not appear to have a statistically significant effect on reading comprehension.

A background questionnaire also revealed that, in all three text-type situations, the native speakers of English performed better than those students who could not speak English in Grade 1. In addition, those students in the standardized text-type group with no previous background knowledge of the material prior to the study generally obtained higher scores on the reading comprehension test than did the other students in either of the two mixed text-type groups.

Acknowledgments

I would like to thank my parents Tony and Ursula Wellinger and my twin brother John for their support and patience throughout my studies.

I would also like to acknowledge my friends for their support. In particular, I would like to recognize Donna Styles, a dear friend who provided constant encouragement and understanding.

And finally a very special thanks to Dr. Gloria Sampson, my senior supervisor, for giving me support, confidence and inspiration.

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

CHAPTER ONE

BACKGROUND TO THE STUDY

INTRODUCTION

Research on reading which has explored the relationship between the reader and the text and characteristics of texts that influence the reader's comprehension (Armbruster, and others 1989; Wagoner 1983; Smith 1991; Eckhoff 1983; Crafton 1982; Garner 1987) has led to debates in recent years which have focused on pedagogical implications of "genre" and its applications to writing (and for our purposes to reading). The issue of "genre" as a structuring device for language teaching has opponents likening it to a "...formulaic way of constructing (or aiding the construction of) particular texts..." (Swales, 1991) and of understanding of such texts. To better understand this issue, it is essential to examine concepts such as "genre" and "text-type" to clarify their relationship to the text and the reader.

In this chapter I will initially discuss the relationship between the concepts of "text-type" and "genre" and the research on analyzing text characteristics and focus, in particular, on Biber's model of textual analysis. The relationship between the reader and the text will then be examined and finally the purpose of the study and a definition of terms will be included.

GENRE AND TEXT-TYPE

The term "genre" has made a place for itself in various fields such as folklore, literary studies, linguistics, and rhetoric. Swales (1990) considers the use of this term in these fields in his book *Genre Analysis*.

In folklore studies, there appears to be three approaches to genre. The first sees genres as classificatory categories (Ben-Amos, 1976), where a story can be classified as a fable, myth, etc. The second approach sees genres as forms that through tradition are considered permanent. The final approach does not label genres according to their form but rather for how they are perceived by the community.

Swales notes that, in literary studies, unlike the folklorists, stability is not as an important feature to the literary critics, who are more concerned with showing how an author breaks the mould of convention and establishes originality. The value of considering genre when analyzing a piece of literature is due to the fact that genre provides an interpretive and evaluative structure for a work of art.

In rhetoric, there are two views, deductive and the inductive. The former view has a closed system of only four categories: expressive, persuasive, literary and referential. They are used to classify discourse into a particular type according to which component in the communicative process receives the primary focus. Each category matches sequentially with the sender, receiver, linguistic form, and finally the realities of the world. The

remaining inductive approach takes context more into account and gives genre a more central place. Rhetorical scholars who take a more inductive approach study the historical development of discourse in recurrent settings.

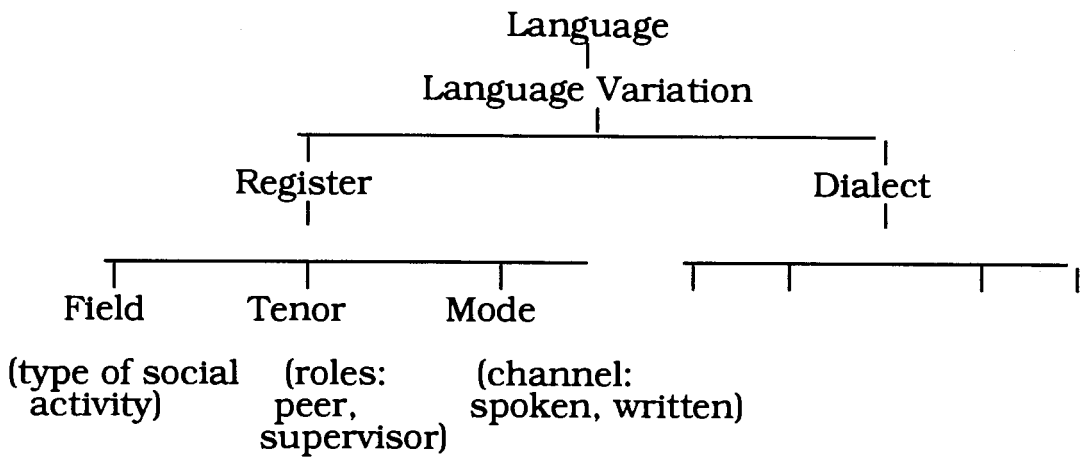
In the field of linguistics, Swales (1990) notes that the term genre is only found with any frequency among linguists of either ethnographic or systemic persuasions and that this is probably due to the well-established and central linguistic concept of register. The concept genre within ethnography tends to refer to the type of communicative event itself, for example, jokes, stories, lectures, etc.

Within systemic linguistics, the concept of genre has also been discussed although the relationship between it and the concept of register have tended to not always be clear. In fact, Frow (1980) does not make a distinction between genre and register.

Halliday (1978) makes use of the concept register (or functional language variation) to refer to the way in which language varies according to the situation in which it is spoken or written. Register, according to Halliday and others, has three dimensions to it: field, mode and tenor. The field (what is spoken or written about), the mode (how language is used), and tenor (the attitude of the speaker or writer to the listener or reader, and to the subject) are expressed through language (Butler, 1985). Jones, Gollin, Drury and Economou (1989) also refer to register as dealing with the style of language, how it describes the way people speak in a linguistic point of view, and the way language works (see Figure 1).

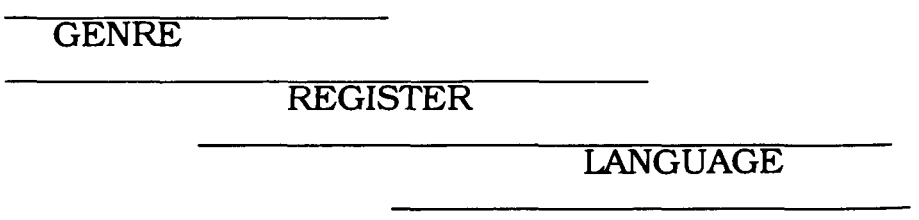
Jones, Gollin, Drury and Economou then proceed to define genre as being a staged, goal-oriented, purposeful activity in which speakers (and writers) engage as members of our culture.

Figure 1 Language, Language Variation and Register



Both Martin (1985) and Couture (1986) clarify the distinction between the two terms register and genre. According to Martin, text instances are considered to be structures generated by system choices on three semiotic communication planes, genre, register and language (Ventola, 1988). These planes are organized so that genre is a higher semiotic plane than register and register a higher plane than language (see Figure 2 below). Genre would appear to be realized by register and register realized by language. Martin's position is based on the fact that genre constrains the ways in which the register variables of field, tenor and mode can be combined in a particular society (Swales, 1990).

Figure 2 Genre, Register and Language as Seen by Martin



Couture, on the other hand, views genre and register slightly differently. Registers impose constraints at the linguistic levels of vocabulary and syntax, whereas genre constraints operate at the level of discourse. Further, genres are completeable structured texts, while registers represent more generalizable stylistic choices (Swales, 1990).

Swales shares Martin's (1985) viewpoint that genres are goal-oriented, social practices. Swales defines genre as

> "...a class of communicative events, the members of which share some set of communicative purpose. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre.... Exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content, and intended audience. If all high probability expectations are realized, the exemplar will be viewed as protypical by the parent discourse community..."
>
> (Swales, 1990)

Biber makes a distinction between "genre" and "text-type". He believes that genre does not adequately represent the underlying text types. Biber proposes that "text can differ by subject-matter, purpose, rhetorical structure, and style, in addition to situational

parameters such as the relation between the communicative participants, the relation of the participants to the external context, and the relation of the participants to the text itself" (Biber 1988, p. 70).

Biber goes on to distinguish between the concepts of "text-type" and "genre" by using "genre" to refer to categorizations assigned on the basis of external criteria relating to author/speaker purpose and by defining "text-type" on the basis of strictly linguistic criteria (similarities in the use of co-occurring linguistic features, irrespective of their genre classification). To illustrate this distinction, Biber uses the example of a science fiction text; it represents a genre of fiction (relating to the author's purpose), but it might represent an abstract and technical text type (in terms of its linguistic form).

Biber considers genres and text types as complementary text categorizations. For example, texts within particular genres can differ greatly in their linguistic characteristics, or on the other hand, different genres can be quite similar linguistically.

TEXT ANALYSIS

There have been a number of text typologies proposed within linguistics and related fields. Variations in typologies can be traced to different originating points of departure and the importance given to a particular element. Typically, however, a functional basis has been used by researchers in developing typologies.

Within Halliday's view of language as social semiotic (Couture 1986), separate domains, such as literature, linguistics and composition have, in their study of language, drawn closer in their investigations of the functions of written language. According to Couture (1986), an adequate functional theory of language must unite speakers, listeners, and situations, and seek the sources of sociosemantic congruence.

The researchers following a functional approach would first identify one or two particular functional dichotomies and then describe the "types" by the poles of those distinctions. For example, functional parameters such as formal/informal, literary/colloquial, etc. have been used. This polarity though does not take into consideration the degree to which a text is, for example, formal or informal.

Another typology proposed by Longacre (1976) distinguishes texts using the parameters of projected time and temporal succession: narrative, expository, procedural, and hortatory. Chafe (1982), on the other hand, employs a four-way classification of texts with respect to the parameters of "involvement-detachment" and "integration-fragmentation".

Within rhetorical theory, four basic "modes" of discourse are traditionally distinguished: narration, description, exposition, and argumentation. While the importance of these four discourse types is widely accepted, there seems to be, however, less agreement on the particular parameters distinguishing among them.

Biber (1989) sees much of the research on spoken/written differences to be based on mode differences between speech and

writing as the distinguishing factor between oral and literate text types. This seems to commonly occur when an oral and a literate text type is compared linguistically and it is assumed that the findings can be generalized to discourse as a whole. Besnier (1988) identifies another problem with early approaches to variation across modes, stating that it was often assumed that variation between spoken and written language overshadowed variation between different types of spoken language and different types of written language. Besnier also claims that researchers in their quest for physical and cognitive explanations for the structural patterns they uncovered have ignored sociolinguistic concerns such as how, why, where, and by whom the discourse is produced, and have not paid attention to the norms of communication at play in each context of production and to the sociocultural definition of the register in the range of communicative activities of the members of the society. Besnier also agrees with Biber's statement that a typology of texts is a research prerequisite to any comparative register analysis.

TEXT ANALYSIS AND CORPUS LINGUISTICS

Within the interdisciplinary field of corpus linguistics, computers are used in the analysis of extended naturally-occurring texts, spoken and written. The central goal of linguists in this field is to reach a better understanding of the workings of human language. The corpus-based approach is founded on the assumption that if a sufficiently large amount of language data or

text is analyzed, the computer's lack of sophisticated knowledge and powers of inference can be compensated to a certain extent. The approach relies on probabilistic predictions made on the basis of observed frequencies in texts to do so. The strength of the above approach is that it is able to deal with any kind of text presented it (Leech 1987).

In the late 1970's when computer use in text analysis was mainly restricted to features that appeared on the typographical surface, Ellegard and students at Gothenburg University and Umea University carried out a text analysis made entirely "by hand" and then used the computer for statistical analyses. Sixty-four text samples were taken from four text categories of the Standard Corpus of Present-Day American English (compiled at Brown University during 1963 and 1964). Using a traditional parsing system for analysis purposes, the group looked at grammatical properties of the text on three different levels: the sentence level, the clause (and phrase) level, and the word level. The statistical results of the study indicated that many linguistic structures occur with remarkable regularity and consistency, both across genres and within individual genres (Ellegard 1978).

In contrast to the manual textual analysis performed by Ellegard and his students, other more sophisticated applications of the computer in textual analysis later became available. CLAWS (the first version), a probabilistic system developed between 1981 and 1983 at the Universities of Lancaster, Oslo and Bergen automatically carries out a grammatical analysis of texts (tagging and parsing) and is reputed to correctly tag words with a rate of

between 96% and 97%, depending on the text (Garside 1987; Marshall 1987).

The textual analysis of corpus linguistics has opened research in many new directions. For example, while Leech and Halliday both agree on viewing language as inherently probabilistic and stressing the need to investigate frequencies in texts to establish probabilities in the grammatical system, their purposes diverge somewhat. Leech's investigations aim towards the tagging and parsing of text items. Halliday, on the other hand, is interested in the "interaction between different subsystems and for a better understanding of historical and developmental change and the variation of language across registers" (Aijmer, Altenberg 1991, p. 3).

In summary, corpus linguistics has many advantages to it. Instead of linguistic analyses being restricted to contrived or invented bodies of text, corpus linguistics allows large amounts of naturally occurring text to be linguistically analyzed in a very short time.

BIBER'S MODEL OF TEXTUAL ANALYSIS

Unlike other studies where analysis begins with a functional distinction and then identification of linguistic features associated with that distinction is conducted as a second step, or those studies that analyze linguistic variations in terms of a single parameter, Biber uses quantitative techniques to identify groups of linguistic features that actually co-occur in texts. These clusters of

features within a particular group define a linguistic dimension, which is then subsequently interpreted in functional terms (Biber, 1988). Priority is therefore given to the linguistic dimension as opposed to the functional as a determinant of type of discourse.

Biber's multi-feature/multi-dimensional approach to linguistic variation was developed to describe the textual relations between spoken and written genres. The texts used in his study (1988) originated from two major text corpora: the Lancaster-Oslo-Bergen Corpus of British English; and the London-Lund Corpus of Spoken English. Although not all texts in the corpora were used, examples from all 23 genres were represented. The corpus contains approximately 1,500,000 words (Biber, p.66).

In Biber's study, a total of six parameters of variation are identified through a factor analysis of co-occurring linguistic forms and interpreted as the following underlying textual dimensions:

(1) Involved versus Informational
(2) Narrative versus Non-Narrative Concerns
(3) Explicit versus Situation-Dependent References
(4) Overt Expression of Persuasion
(5) Abstract versus Non-Abstract Information
(6) On-Line Informational Elaboration

Most of the dimensions consist of two groupings of features, which represent sets of features that occur in a complementary pattern. That is, when the features in one group occur together

frequently in a text, the features in the other group are markedly less frequent in that text, and vice versa. To interpret the dimensions, it is important to consider likely reasons for the complementary distribution of these two groups of features as well as the reasons for the co-occurrence pattern within each group. Figure 3 below presents the sets of features for each of the six dimensions identified. **Note: Because some features are included on more than one of the factors (eg. present tense on Dimensions 1 and 2), each feature was included in the computation of only one dimension score to ensure the experimental independence of the dimension scores. Thus, each linguistic feature was included in the dimension score of the dimension on which it had the highest loading.

Figure 3 Co-occurrence of Linguistic Features Associated With Each of Biber's Six Dimensions

Dimension 1
private verbs
THAT deletion
contractions
present tense verbs
2nd person pronouns
DO as pro-verb
analytic negation
demonstrative pronouns
general emphatics
1st person pronouns
pronoun IT
BE as main verb
causative subordination
discourse particles
indefinite pronouns
general hedges
amplifiers
sentence relatives

Dimension 2
past tense verbs
third person pronouns
perfect aspect verbs
public verbs
synthetic negation
present participial clauses

(present tense verbs)
(attributive adjectives)
(past participial WHIZ deletions)

Dimension 3
WH relative clauses on object positions
pied piping constructions

Dimension 1 (cont.)
WH questions
possibility modals
non-phrasal coordination
WH clauses
final prepositions
(adverbs)
(conditional subordination)
---
nouns
word length
prepositions
type/token ratio
attributive adjectives
(place adverbials)
(agentless passives)
(past participial WHIZ deletions)
(present participial WHIZ deletions)

Dimension 3 (cont.)
WH relative clauses on subject positions
phrasal coordination
nominalizations
---
time adverbials
place adverbials
adverbs

Dimension 4
infinitives
prediction modals
suasive modals
conditional subordination
necessity modals
split auxiliaries
(possibility modals)

None

Dimension 5
conjunctions
agentless passives
past participial clauses
BY passives
past participial WHIZ deletions
other adverbial sub-ordinators
(predicative adjectives)
---
type/token ratio

Dimension 6
THAT clauses as verb complements
demonstratives
THAT relative clauses on object positions
THAT clauses as adjective complements
(final prepositions)
(existential THERE)
(demonstrative pronouns)
(WH relative clauses on object positions)
---
phrasal coordination

According to Biber, the first dimension, Involved versus Informational, represents at one end of the pole discourse which is interactional, affective, demonstrates involved purposes, and is associated with strict real-time production and comprehension constraints. The prevalent co-occurrence of the linguistic features within the group of features above the dotted line of Dimension 1 (see Figure 3) is associated with an involved non-informational

focus. At the other end of the pole is discourse with highly informational purposes, which is carefully crafted and highly edited. High frequencies of features from below the dotted line are associated with a high informational focus and a careful integration of information in a text.

Dimension 2 distinguishes discourse with primarily narrative purposes, characterized by high frequencies of the upper group of features listed under Dimension 2, from non-narrative purposes (for example, expository, descriptive, etc.), which are characterized by high frequencies of the features located below the dotted line of Dimension 2.

The third dimension characterizes highly explicit, context-independent reference versus nonspecific, situation-dependent reference. In this case, the upper group of features listed under Dimension 3 are associated with the explicit reference whereas the remaining set of features correspond to the nonspecific reference.

Dimension 4 indicates the degree to which persuasion is marked overtly, whether marking the speaker's point of view, or the speaker's attempt to persuade the addressee.

The next dimension, Abstract versus Non-abstract Information marks informational discourse that is abstract, technical, and formal in style versus other types of discourse.

The last dimension, Dimension 6, distinguishes discourse that is informational but produced under real-time conditions versus discourse that is not produced under real-time constraints.

Dimensions have both linguistic and functional content. The linguistic content consists of a group of linguistic features that co-

occur with a high level of frequency in texts. The co-occurring patterns are interpreted in terms of the situational, social, and cognitive functions shared by the co-occurring linguistic features (Biber, 1989). In addition, unlike other studies, dimensions here permit a continuous range of texts to be characterized along each dimension.

Dimension scores can be computed to characterize each text with respect to each dimension. This can be done by first normalizing frequencies of all linguistic features to a text length of 1,000 words and standardizing to a mean of 0.0 and a standard deviation of 1.0. Dimension scores for each text are then computed by summing the frequencies of the positively loaded and subtracting the negatively loaded defining linguistic features of the dimension.

Using the dimension scores, the linguistic relations among texts can be considered by comparing their dimension scores, and the relation among text varieties can be considered by comparing the mean dimension scores of each variety.

Biber's work (1988) demonstrates that certain linguistic features that co-occur in English texts appear to be rule-governed and that there is an internal linguistic coherency of text types. The results of a similar study by Besnier (1988) that linguistically analyzed several registers of Nukulaelae, a Polynesian language of Central Island, are congruent with Biber's findings of internal linguistic coherency within text-types.

LANGUAGE AND GENRE

Language is more than communication and the conveying of information. Language is also a social practice and language use is an indicator of social structures and process.

Swales (1990) distinguishes between speech communities and discourse communities. He views membership in a speech community as achieved through birth, adoption or accident, while membership in a discourse community is achieved through persuasion, training or relevant qualification.

We are therefore born into a speech community, but become members of a variety of discourse communities. A discourse community can be identified as having its own genre and some specific lexis. For example, Marshall (1991) considers science learning as a process of initiation into a new culture, where linguistic and communicative competence need to be acquired to communicate within this specific discourse community. The number of discourse communities we belong to will depend on factors such as age, education, socioeconomic status.

Miller (1984) states that the number of genres in any society is indeterminate and depends upon the complexity and diversity of society. Ferrara, Brunner, and Whittemore (1991) also see language as ever changing and provide an example of this based on their investigation of an emerging register called "Interactive Written Discourse (IWD)". Other registers they mention include Ferguson's Baby Talk, Sport Announcer Talk, Foreign Talk, and Bureaucratic Language and state that these registers show

syntactic variation on every level of language: morphological; phonological, syntax, and discoursal as well as lexicon.

READING AND GENRE

Reading is a human behaviour which takes place in connection with written language. While spoken language may be natural to human beings, writing is really a technology that is artificial in the sense of it being governed by consciously contrived, articulable rules (Ong, 1982).

If we viewed written language as nothing more than ciphered speech, reading might be simply defined as the ability to decode or translate the script into its spoken equivalent. However, reading includes a higher level of processing, namely the ability to extract meaning from written text.

There are basically three models of reading, the bottom-up model, the top-down and the interactive model. The processes involved in reading in the bottom-up model appear to be organized hierarchically. The attainment of any given level presumes the execution of all subordinate or less complex levels. This dependency is unidirectional (from the individual letter to the sentence).

In contrast to the first model, the top-down model is based on the premise that skilled readers should rely as little as possible on graphemic details but should exploit the semantic and syntactic constraints of a text.

Adams (1982) summarizes the fundamental problems with the bottom-up and the top-down models as being their very one-sidedness. The former model fails to recognize the role of the higher order knowledge that even young readers bring to the text while the latter model fails to acknowledge the importance of lower level processes which the text requires of the reader.

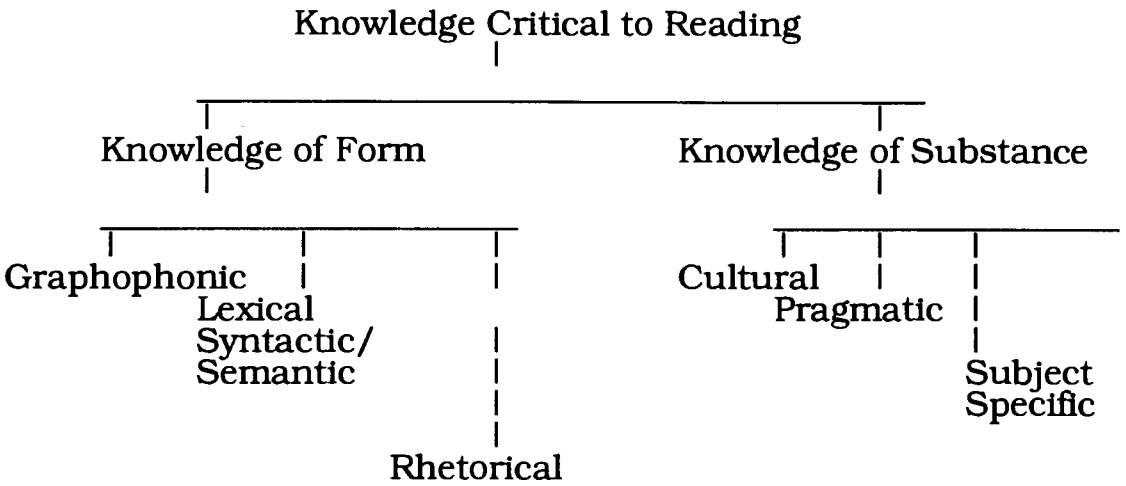
The interactive model of reading, in contrast, allows parallel processing of multiple types and levels of information to occur. This model maintains that readers use expectations based on world knowledge and the organization of text.

A skilled reader actively organises his experience of the text, using his knowledge of the world and of previously encountered structurally similar texts to facilitate comprehension. Dubin and Eskey (1986) present a model of the knowledge crucial to reading (Figure 4 below). Subsumed under the knowledge base critical to reading is the knowledge of form and substance. According to Dubin and Eskey, knowledge of substance includes three main areas: knowledge of the subject material itself; the reader's effective selection of the cultural information relevant to the comprehension of the text; and the reader's understanding of the situational context (ie. knowledge and beliefs of the writer and the relation between the writer and the reader).

Dubin and Eskey's knowledge of form also includes three areas of knowledge. The first area, "graphophonic", pertains to the knowledge of letter-sound correlations. The second area covers "lexical", and "syntactic/semantic" knowledge. Here the reader needs to know the vocabulary; the rules to combining words to

form grammatical sentences; and finally to understand the meaning communicated through language.

Figure 4 Knowledge Critical to Reading



In the reading process genre is an important factor. Genres follow certain conventions. These provide guidance in the form of background knowledge (prior schemata) to readers (Langer, 1990; Littlefair, 1989). Littlefair contends that experienced readers implicitly know that different genres of books have different linguistic patterns but that it is, however, a far more difficult task to be able to describe these linguistic differences explicitly. Specifically skilled readers are sensitive to the many linguistic variations linked to genres and use them as a vehicle to further understanding of the reading material.

Studies have indicated that exposure to written language helps children learn about language structures (Eckhoff, 1983).

Within the classroom environment, narration is the first main genre the student encounters. Over time he internalizes this structure and knows what to expect in a narration. Later, the student is introduced to other genres which he internalizes. It would therefore follow that an adult would be familiar with a large number of genres through exposure to them over his life time.

Perera (1986) however, still feels that structured teaching of reading after mastering the initial decoding state of reading is still necessary. She compares this need for instruction to teaching piano and practicing scales.

In the area of second language learning, Walsh (1982) focuses on the difficulties students studying in a second language have in reading scientific texts in English. The difficulties occur due to three separate yet closely connected variables: the linguistic, the rhetorical, and the conceptual variables. ESP (English for Specific Purposes) specialists have tried to address these concerns in order to initiate the non-native speaker into the desired discourse community. The interest of a genre-based approach to the teaching of English would fulfill part of this need by making explicit the knowledge about how the type of text will vary according to purpose, topic, audience and channel of communication.

PURPOSE OF THE STUDY

The purpose of this study is to examine the effects of genre mixing on reading comprehension. The hypothesis of interest is based on the premises that a) genre is rule-governed, that b)

combining two genres in one text breaks rules and c) that therefore a passage that combines conventions of two genres will be harder to understand by literate readers fully enculturated in many genres because it breaks or violates the genre rules that they have internalized.

TERMS

There are basically two important concepts that need to be defined although they have already been discussed earlier in this chapter. They are "genre" and "text-type".

The term "genre" is generally defined as "a category of... literary composition characterized by a particular style, form, or content" according to *Websters Dictionary*. In this study "genre" will refer to a social action that emerges as a conventional response to particular and recurrent situations (Nystrand, 1986) which include both the written and spoken language modes.

In this study "text-type" is a text that is defined strictly on the basis of linguistic criteria.

CHAPTER TWO

THE TEXTS

INTRODUCTION

This chapter will address the first premise of the hypothesis of interest, namely that genre is rule-governed. The texts used in the linguistic analysis and the procedures used and the results of the linguistic analysis will be presented in this chapter. Chapter Three will focus on the remaining premise that combining two different genres in one text breaks linguistic rules which would cause literate readers who have internalized these rules to encounter difficulties understanding a text that combines the linguistic conventions of two genres.

SELECTION OF TEXTS

Three popular texts commonly used in a general introductory course of Educational Psychology were selected for Part I of the study's purposes. These texts are *Applying Educational Psychology in the Classroom* by Myron Dembo, *Educational Psychology* by Anita Woolfolk, and *Educational Psychology* by N.L. Gage and David Berliner. The selection was based on the texts used in introductory Ed Psych classes and the popularity of the texts, indicated by the number of editions. In this case the first two texts used were in their third edition and the last text in its fourth. Another

consideration in using the *Educational Psychology* text by Woolfolk in particular was the availability of Katherine Cummings' *Test Item File* accompanying it. Test items contained within this test item file were later used in the study to assess reading comprehension.

A topic was selected from one of the Woolfolk text's "Table of Contents" and matched with chapters discussing the same content in the other two texts. The topic was testing and standardized testing in education. These chapters (Section F containing Chapters 22 and 23 in Gage' and Berliner's textbook, Chapter 14 in Woolfolk's textbook, and Part 4 Chapter 12 in Dembo's textbook) (Appendix A) were then linguistically analyzed according to Biber's model of internal textual relations as a means of characterizing the texts. Two points need to be noted here. Firstly, none of the headings, tables and figures within the chapters of the chosen textbooks was analyzed. Secondly, because of the varying text lengths, the frequency counts of all the linguistic features analyzed were normalized to a text length of 1,000 words in order to compare frequency counts across texts.

DESCRIPTION OF THE TEXTS ANALYZED

Initially the three texts *Educational Psychology* by Anita Woolfolk, *Educational Psychology* by N.L. Gage and David Berliner, and *Applying Educational Psychology in the Classroom* by Myron Dembo were analyzed by the software program "Correct Grammar". It was used to calculate the average number of words, sentences, and paragraphs, and to assess the readability level of each text. It

is to be noted here that headings were not included in the count nor later linguistically analyzed.

The program "Correct Grammar" uses three different scales for assessing the readability of texts, the Flesch Reading Ease Score, Flesch-Kincaid Grade Level and the Gunning Fog Index. The Flesch Reading Ease Score is based on the number of words in each sentence, and the average number of syllables per word. Alternatively the Flesch-Kincaid and Gunning Fog systems attempt to represent readability as a school grade level. Tables 1 and 2 summarize these properties.

Table 1 Properties of the Three Texts

Properties	Woolfolk text	Dembo text	Gage text
# of words	7266	7942	13636
# of letters per word	4.8	4.8	4.8
# of syllables per 100 words	164	164	163
# of sentences	441	452	832
# of words per sentence	16.4	17.5	16.3
# of paragraphs	137	178	280
# of sentences per paragraph	3.2	2.5	2.9

Table 2 Readability Scores for the Three Texts

Woolfolk Text		
Flesch Reading Ease Score	50.8	Fairly Difficult
Grade Level Required	12	
US Adults Who Can Understand	54%	
Flesch-Kincaid Grade Level	10.2	
Gunning Fog Index	9.4	
Dembo Text		
Flesch Reading Ease Score	50.2	Fairly Difficult
Grade Level Required	12	
US Adults Who Can Understand	54%	
Flesch-Kincaid Grade Level	10.5	
Gunning Fog Index	9.9	
Gage Text		
Flesch Reading Ease Score	51.5	Standard
Grade Level Required	11	
US Adults Who Can Understand	69%	
Flesch-Kincaid Grade Level	10.0	
Gunning Fog Index	9.5	

The properties of the three texts in Table 1 above provide a comparative overview of the structuring and the lengths of the Woolfolk, Dembo and Gage texts. With reference to the length of the three texts, the Gage text is almost double in length. Consequently, as expected, the number of sentences in the Gage text is again almost double in number. The totals for all three texts on both the number of letters per word and on the number of syllables per 100 words levels, are either exactly or almost identical.

On the paragraph level, the Woolfolk, Dembo and Gage texts differ. Unlike the almost consistent 2:1 ratio of text length to the

number of sentences above, the number of paragraphs within the texts vary. In this case, the Gage text has over double the number of paragraphs in relation to the Woolfolk text and the Dembo text approximately two-thirds the number of paragraphs in the Gage text. Although this might seem significant, it must be kept in mind that it is necessary to take into the account the length of the Gage text as being almost double either the Woolfolk and Dembo texts.

The next property, that of the number of sentences per paragraph, differs again in the relationship among the three texts. In this instance, the number of sentences per paragraph within the Gage text lies almost half way between the Woolfolk and Dembo texts.

Having considered the paragraph level, attention can now be directed to the sentence level and the average number of words located within a sentence. Table 1 shows that the Woolfolk and Gage texts contain almost the same number of words per sentence whereas the Dembo text contains longer sentences (17.5 words per sentence as opposed to 16.4 and 16.3 respectively for the Woolfolk and Gage texts).

The overview of the properties examined in the preceding paragraph leads the way into looking at the readability scores obtained for the three texts since two properties mentioned above, namely the number of syllables per 100 words and the number of words per sentence, are made use of to evaluate the readability of a text.

Both the Woolfolk and Dembo texts have, overall, very similar readability scores. The Gage text, however, has a higher Flesch

reading ease score making that particular text easier to read and elevating the reading difficulty level from fairly difficult to that of standard. The difference in percentage of US adults who can understand the Gage text is certainly much larger than in the comparison of any of the other scales or levels. For example, 69% of US adults are able to understand the Gage text, whereas only 54% of US adults would understand either the Woolfolk or Dembo texts. In comparison, the grade level required to understand any of the three texts and both the Flesch-Kincaid grade level and Gunning Fog index do not fluctuate as much. For each scale respectively, there are ranges of fluctuations of over one year and for both Flesch-Kincaid grade level and Gunning Fog index over .5 units.

LINGUISTIC ANALYSIS OF THE THREE TEXTS

The linguistic analysis of selections from the three texts *Educational Psychology* by Anita Woolfolk, *Educational Psychology* by N.L. Gage and David Berliner, and *Applying Educational Psychology in the Classroom* by Myron Dembo is based on Biber's model of internal relations. Following Biber's model, just those sixty-seven potentially important linguistic features that have been associated with particular communicative functions in previous research were focused on. These linguistic features included:

A. Tense and Aspect Markers
    (1) past tense
    (2) perfect aspect

(3) present tense

B. Place and Time Adverbials

(4) place adverbials

(5) time adverbials

C. Personal Pronouns

(6) first person pronouns

(7) second person pronouns

(8) third person pronouns

Impersonal Pronouns

(9) pronoun *it*

(10) demonstrative pronouns

(11) indefinite pronouns

Pro-verbs

(12) pro-verb *do*

D. Questions

(13) direct WH-questions

E. Nominal Forms

(14) nominalizations

(15) gerunds

(16) total other nouns

F. Passives

(17) agentless passives

(18) *by*-passives

G. Static Forms

(19) *be* as main verb

(20) existential *there*

H. Subordination

Complementation

(21) *that* verb complements

(22) *that* adjective complements

(23) WH-clauses

(24) infinitive

Participial Forms

(25) present participial clauses

(26) past participial clauses

(27) past participial WHIZ deletion relatives

(28) present participial WHIZ deletion relatives

Relatives

(29) *that* relative clauses on subject position

(30) *that* relative clauses on object position

(31) WH relative clauses on subject position

(32) WH relative clauses on object position

(33) pied-piping relative clauses

(34) sentence relatives

Adverbial clauses

(35) causative adverbial subordinators: *because*

(36) concessive adverbial subordinators:

although, though

(37) conditional adverbial subordinators: *if, unless*

(38) other adverbial subordinators: (having multiple functions

I. Prepositional Phrases

(39) total prepositional phrases

Adjectives and Adverbs

(40) attributive adjectives

(41) predicative adjectives

(42) total adverbs

J. Lexical Specificity

(43) type/token ratio

(44) word length

K. Lexical Classes

(45) conjuncts

(46) downtoners

(47) hedges

(48) amplifiers

(49) emphatics

(50) discourse particles

(51) demonstratives

L. Modals

(52) possibility modals

(53) necessity modals

(54) predictive modals

M. Specialized verb classes

(55) public verbs

(56) private verbs

(57) suasive verbs

(58) *seem/appear*

N. Reduced forms and dispreferred structures

(59) contractions

(60) subordinator-*that* deletion

(61) stranded prepositions

(62) split infinitives

(63) split auxiliaries

O. Coordination

(64) phrasal coordination

(65) independent clause coordination

P. Negation

(66) synthetic negation

(67) analytic negation: *not*

The frequency count of the 67 linguistic features mentioned above (excluding type/token ratio and word length) were normalized to texts of 1000 words to enable comparison among the three texts. This was performed so that a comparison of non-normalized counts would give a truly accurate assessment of the frequency distributions in texts of varying lengths. In order to normalize the counts to represent frequencies per 1000 words, the frequency count of each linguistic feature was divided by the length of the text being analyzed and then multiplied by 1000. Table 3 shows the frequency counts for all three texts and the results will be discussed shortly.

As noted in the previous paragraph, type/token ratio and word length were not calculated in the same way as the frequency counts. Type/token ratio is the measure of difficulty of a text and has been extensively used in child language research as an index of lexical diversity (Richards, 1987). It is assumed that an increase in

the number of different lexical items in a text increases textual difficulty. The ratio of the number of different words in a text (the word "types") to the total number of words in a text (the word "tokens") is computed, in this case, by counting the number of different lexical items that occur in the first 400 words of a text, and then dividing by four.

The final linguistic feature not calculated as part of the frequency counts, word length, is the mean length of the words in a text, in orthographic letters.

Table 3 Frequency Count of each Linguistic Feature for each Text (Normalized to a Text Length of 1000 words)

Linguistic Feature *	Woolfolk Text	Dembo Text	Gage Text
(1)	9.08	3.50	5.65
(2)	4.68	4.18	6.31
(3)	62.21	51.64	79.42
(4)	5.23	0.31	3.37
(5)	2.34	1.05	3.22
(6)	5.37	2.51	18.55
(7)	12.52	3.34	8.95
(8)	9.77	7.21	13.57
(9)	4.95	4.60	7.85
(10)	3.58	0.42	1.91
(11)	0.69	0.84	0.95
(12)	0.69	0.21	1.39
(13)	0.83	2.61	2.79
(14)	40.19	58.33	41.58
(15)	18.17	30.63	18.33
(16)	280.07	254.73	263.86
(17)	20.23	7.35	13.57
(18)	2.06	4.08	2.27
(19)	26.42	17.56	25.81
(20)	1.65	1.46	1.91
(21)	3.85	5.44	5.87
(22)	0.55	0.21	0.51
(23)	4.82	3.66	9.17
(24)	14.31	17.87	20.90

Linguistic Feature *	Woolfolk Text	Dembo Text	Gage Text
(25)	0.14	0.00	0.88
(26)	0.00	0.10	0.29
(27)	0.00	1.46	3.45
(28)	1.24	1.88	1.32
(29)	3.17	3.76	5.57
(30)	0.00	0.42	0.66
(31)	0.96	2.20	3.37
(32)	0.69	0.00	0.51
(33)	0.55	1.15	3.22
(34)	0.41	0.63	0.37
(35)	1.38	2.20	1.10
(36)	0.55	1.05	0.22
(37)	5.37	1.78	5.72
(38)	0.69	0.73	0.73
(39)	120.70	113.10	115.28
(40)	107.49	95.12	83.60
(41)	15.41	12.96	16.94
(42)	35.65	28.12	31.31
(43)	9.75	7.50	9.25
(44)	4.80	4.80	4.80
(45)	4.40	6.06	2.86
(46)	2.89	0.84	2.71
(47)	0.00	0.00	0.29
(48)	1.65	0.52	1.10
(49)	0.41	0.00	1.03
(50)	0.14	0.00	0.07
(51)	6.74	6.69	9.09
(52)	11.01	8.99	11.15
(53)	3.44	3.55	5.57
(54)	5.64	4.08	4.77
(55)	1.65	2.09	2.05
(56)	7.98	8.88	14.81
(57)	0.41	0.84	1.98
(58)	0.83	0.63	0.73
(59)	0.96	0.10	1.76
(60)	0.28	0.00	0.66
(61)	0.00	0.21	0.07
(62)	0.28	0.00	0.00
(63)	1.38	1.99	1.32
(64)	15.69	22.68	14.74
(65)	2.06	2.20	2.57
(66)	1.38	0.94	0.51
(67)	5.64	6.74	7.63

*Linguistic features are numbered according to those cited earlier

FREQUENCY COUNTS OF THE LINGUISTIC FEATURES

In all three texts, the five largest frequency counts of linguistic features present are the same, namely present tense (Woolfolk 62.21; Dembo 51.64; and Gage 79.42), total other nouns (Woolfolk 280.07; Dembo 254.73; and Gage 263.86), prepositional phrases (Woolfolk 120.70; Dembo 113.10; and Gage 115.28), type/token (Woolfolk 52.75; Dembo 54; and Gage 51.5) and attributive adjectives (Woolfolk 107.49; Dembo 95.12; and Gage 83.60). See Table 3 above.

The co-occurrence of these particular linguistic features across the three texts does not appear to be random. Biber in his study also found co-occurrences of these linguistic features whose weightings along with those of a few other linguistic features to comprise the "Involved versus Informational Production" dimension. He found high frequencies of four of the linguistic features above in particular to be associated with a high informational focus and a careful integration of information in a text. These linguistic features include nouns, prepositional phrases, type/token ratio and attributive adjectives.

To understand this pattern of co-occurrence better, let's consider the five linguistic features cited above and their functions. The first item, nouns, is generally considered the primary bearer of referential meaning in a text. A high frequency of nouns would therefore indicate a great density of information. Prepositional phrases likewise serve to integrate high amounts of information into a text. While type/token ratio (number of different lexical items occurring in the first 400 words of a text, calculated as a

percentage) similarly marks high density of information, it also marks very precise lexical choice resulting in an exact presentation of informational content. Attributive adjectives are used to further elaborate nominal information since, unlike less integrated linguistic forms such as predicative adjectives or relative clauses, attributive adjectives pack information into relatively few words and structures.

The fifth linguistic feature with a high frequency count mentioned above, the present tense, is also mentioned by Biber along the "Involved versus Informational Production" dimension. However, the present tense as a linguistic feature is not given as high a priority because of its complementary relationship to the other co-occurring linguistic features.

The linguistic analysis of the Woolfolk, Dembo and Gage texts present evidence that genre is rule-governed. The three texts can be grouped with respect to their linguistic form. Certain patterns of co-occurrence can be observed using Biber's model. A particular grouping across the three texts include such linguistic features such as nouns, attributive adjectives, prepositional phrases, type/token ratio and the present tense.

INTERRATER RELIABILITY

Interrater reliability of the linguistic analysis was established by comparing the linguistic analyses of random selections in each of the three texts performed by a graduate student in the Department of Linguistics and that done by me. Initially

paragraphs from the three texts were linguistically analyzed and compared (approximately 500 words total).

The number of linguistic features actually tagged by both of us though did not match. There was a difference of about 34 items. Upon closer inspection, it was determined that except for one of the 34 items, the graduate student had included a group of determiners (quantifiers such as all, every, some, etc) that were found to not be essential in the linguistic analysis for distinguishing between and among different text-types (based on Biber's review of previous research in which linguistic features occurred in particular types of texts). In addition, one linguistic feature had been inadvertently mislabeled since the marker did not disagree on the type of linguistic feature the item was.

Another selection of passages randomly chosen from the three texts was again analyzed and compared. The overall correlation was 1 using the Pearson correlation. While this might be perceived as being unlikely to occur in a correlation, in this case it does have validity. This can be attributed to the fact that the underlying grammatical categories in a linguistic analysis are fundamental to all linguists and have been for many decades. Where there are discrepancies, however, lie in the perceptions of the various linguists of what function these individual categories fulfill and their degree of importance. Therefore the very scientific nature of the linguistic analysis would indicate that a perfect correlation would be acceptable.

STATISTICAL ANALYSIS OF THE TEXTS

The descriptive statistics for all three texts were then calculated and are shown in Table 4. In comparing the mean frequency count of each linguistic feature for all three texts to the mean frequency counts of Biber's academic prose, it is interesting to observe that most of the frequency counts are in line with what Biber found in his linguistic analyses, specifically that of academic prose.

This present finding reinforces the position that overall academic prose and other genres in general do have patterns of co-occurring linguistic features that mark underlying functional dimensions.

Table 4 Mean, Minimum and Maximum Values, Range, and Standard Deviation for the Frequency Count of Each Linguistic Feature of All Three Texts Combined

Ling. Feat.	Mean	Min	Max	Range	SD
(1)	6.08	3.50	9.08	5.58	2.81
(2)	5.06	4.18	6.31	2.13	1.11
(3)	64.42	51.64	79.42	27.78	14.02
(4)	2.97	0.31	5.23	4.92	2.48
(5)	2.20	1.05	3.22	2.17	1.09
(6)	8.81	2.51	18.55	16.04	8.56
(7)	8.27	3.34	12.52	9.18	4.63
(8)	10.18	7.21	13.57	6.36	3.20
(9)	5.80	4.60	7.85	3.25	1.78
(10)	1.97	0.42	3.58	3.16	1.58
(11)	0.83	0.69	0.95	0.26	0.13
(12)	0.76	0.21	1.39	1.18	0.59
(13)	2.08	0.83	2.79	1.96	1.08
(14)	46.70	40.19	58.33	18.14	10.10
(15)	22.38	18.17	30.63	12.46	7.15
(16)	266.22	254.73	280.07	25.34	12.83
(17)	13.72	7.35	20.23	12.88	6.44
(18)	2.80	2.06	4.08	2.02	1.11
(19)	23.26	17.56	26.42	8.86	4.95

Ling. Feat.	Mean	Min	Max	Range	SD
(20)	1.67	1.46	1.91	0.45	0.23
(21)	5.05	3.85	5.87	2.02	1.06
(22)	0.42	0.21	0.55	0.34	0.19
(23)	5.88	3.66	9.17	5.51	2.90
(24)	17.69	14.31	20.90	6.59	3.30
(25)	0.34	0.00	0.88	0.00	0.47
(26)	0.13	0.00	0.29	0.29	0.15
(27)	1.64	0.00	3.45	3.45	1.73
(28)	1.48	1.24	1.88	0.64	0.35
(29)	4.17	3.17	5.57	2.40	1.25
(30)	0.36	0.00	0.66	0.66	0.33
(31)	2.18	0.96	3.37	2.41	1.21
(32)	0.40	0.00	0.69	0.69	0.36
(33)	1.64	0.55	3.22	2.67	1.40
(34)	0.47	0.37	0.63	0.26	0.14
(35)	1.56	1.10	2.20	1.10	0.57
(36)	0.61	0.22	1.05	0.83	0.42
(37)	4.29	1.78	5.72	3.94	2.18
(38)	0.72	0.69	0.73	0.04	0.02
(39)	116.36	113.10	120.70	7.60	3.19
(40)	95.40	83.60	107.49	23.89	11.95
(41)	15.10	12.96	16.94	3.98	2.01
(42)	31.69	28.12	35.65	7.53	3.78
(43)	8.83	7.50	9.75	2.25	1.18
(44)	4.80	n/a	n/a	n/a	n/a
(45)	4.44	2.86	6.06	3.20	1.60
(46)	2.15	0.84	2.89	2.05	1.14
(47)	0.10	0.00	0.29	0.29	0.17
(48)	1.09	0.52	1.65	1.13	0.57
(49)	0.48	0.00	1.03	1.03	0.52
(50)	0.07	0.00	0.14	0.14	0.07
(51)	7.51	6.69	9.09	2.40	1.37
(52)	10.38	8.99	11.15	2.16	1.21
(53)	4.19	3.44	5.57	2.13	1.20
(54)	4.83	4.08	5.64	1.56	0.78
(55)	1.93	1.65	2.09	0.44	0.24
(56)	10.56	7.98	14.81	6.83	3.71
(57)	1.08	0.41	1.98	1.57	0.81
(58)	0.73	0.63	0.83	0.20	0.10
(59)	0.94	0.10	1.76	1.66	0.83
(60)	0.31	0.00	0.66	0.66	0.33
(61)	0.09	0.00	0.21	0.21	0.11
(62)	0.09	0.00	0.28	0.28	0.16
(63)	1.56	1.32	1.99	0.67	0.37
(64)	17.70	14.74	22.68	7.94	4.34
(65)	2.28	2.06	2.57	0.51	0.26
(66)	0.94	0.51	1.38	0.87	0.44
(67)	6.67	5.64	7.63	1.99	1.00

CHAPTER THREE

THE EXPERIMENTAL ASSESSMENT OF COMPREHENSION

INTRODUCTION

Chapter three will focus on the final premise of this study which postulates that combining two different genres in one text breaks linguistic rules which would cause literate readers who have internalized these rules to have difficulties understanding a text that combines the linguistic conventions of two genres. This chapter will present the experimental design used and Chapter four will present the results.

CONSTRUCTION OF THE READING PASSAGES AND READING COMPREHENSION TEST

My aim in creating the reading passages and developing the reading comprehension test is to investigate if the reading passages containing conventions of two different genres will be harder for literate readers enculturated in many genres to understand than a passage which contains only the linguistic conventions typical of a single genre.

The major problem in designing the three reading passages was the need to ascertain which linguistic features needed to be included in developing the mixed passages. Since the texts

analyzed in Chapter 2 fell into the genre of academic prose, I examined the means of the six descriptive dimension statistics of the academic prose genre and compared them with the means of those of the other 22 genres and sub-genres Biber had analyzed. Based on this comparison, it appeared the conversational face-to-face and telephone sub-genres varied the most over Dimensions 1, 3 and 5.

Referring back to the co-occurring linguistic features on Factors 1, 3 and 5, on which the dimensions are based, a number of linguistic features prevalent in the conversational genre were selected to be included in the two passages that combined linguistic features of both the academic prose and conversational genres. These two passages differed in that the linguistic elements associated with the conversational genre were placed in the topic sentence in one of the mixed text-type passage and placed in secondary sentences in the second mixed text-type passage. The third passage followed the linguistic conventions of the academic prose genre. See Appendix B for the three reading passages used in the study and Appendix D for the two mixed text-type passages in which the altered sentences are highlighted.

Propositional content was also taken into consideration when constructing the passages because of its importance within the comprehension process. A proposition refers to a simple linguistic description, a "unit of meaning" that has been most commonly used in work on memorizing and comprehending text. In essence a proposition is an abstract statement about an entity (i.e. a person or an object) or about the relationship between two or more such

entities (Mitchell, 1982). For example, a proposition might state a property or state of affairs is true of a person or object (e.g. Lucy is pretty) or it might state that a certain action or activity is taking place between two entities (e.g. Lucy hit the ball).

Studies have shown that in memory tasks, sentences are analyzed and stored in terms of its propositional structure. Mitchell cites research conducted by Anderson and Bower (1973) and Ratcliff and McKoon (1978) that supports this particular view. A study by Sachs (1974) found that target sentences appear to be stored in a form which preserves the overall meaning of the sentences, but not necessarily the details of the wording. For example, readers in Sach's study tended to report that formal and lexical sentences were identical to the target sentences.

Propositions, however, can not be examined solely in isolation since most texts are made up of a large number of different propositions and the way in which separate units are integrated is extremely important in the comprehension process. Research conducted by Kintsch and his colleagues (Kintsch and Keenan, 1973; Kintsch 1974; Kintsch, Kozminsky, Streby, McKoon and Keenan, 1975) investigated this field of study. They analyzed texts into hierarchical structures consisting of a few superordinate propositions and a much larger number of different levels of subordinate propositions where the former propositions corresponded to the major themes of the text while the latter expressed more peripheral information. Their research found that the probability of recalling a particular proposition was closely related to its position in the hierarchy. Readers tended to recall

higher-level propositions rather than low-level details in the hierarchical plot structures.

In this study, the propositional content of all three reading passages was identical and identically structured. The propositions in all three passages were not semantically changed in any way and the hierarchical structure of the propositions was retained. Two examples taken from the reading passages are presented below to demonstrate this.

Figure 5 Prepositional Content Examples from the Three Passages

---

Example 1:Passages A and B (Standardized Tests section, pg. 2, final paragraph)

This type of test is used in the selection of a limited number of candidates for admission to certain programs.

Passage C (same location as above)

You use this type of test to select a limited number of candidates for admission to certain programs.

Example 2:Passages A and C (Interpreting Standardized Tests section, pg. 3, paragraph 5)

Percentile rank scores are another form of ranking used in comparing a student's raw score to that of the norming sample.

Passage B (same location as Passages A and C)

You can use percentile rank scores to compare a student's raw score to that of the norming sample.

The passages for Groups B and C differed in the placement of the mixed conventions within either the topic sentence or a

secondary sentence. Group B read the passage where the topic sentences had been altered linguistically, whereas Group C read the passage where the secondary sentences had been affected by the deliberate mixing of generic conventions.

The length of the three passages was kept to approximately four pages due to time requirements on the volunteer subjects. The passages were also controlled for level of proposition in the hierarchy and for reading level.

METHODOLOGY USED TO TRANSFORM TEXTS

The methodology used in this study to transform the conventional passage into the mixed text-type passages involved either altering the topic sentences in one passage and the auxiliary sentences in the other mixed text-type passage. I selected this methodology because of my concern for the influence of the position within a paragraph of the mixed text-type sentences.

Although the above methodology was used, there are other methodologies that could have been looked at and should definitely be included in future studies investigating the effect of text-type mixing on reading comprehension.

Other methodologies could focus on creating passages by mixing text-types on the paragraph level. For example, one methodology could mix text-types on the paragraph level with an ABABAB ordering of paragraphs where each entire paragraph follows the linguistic conventions of a particular genre. Another possible methodology could incorporate groupings of paragraphs.

Here, a control group could be compared to two experimental groups that have either been given passages containing AAABBB or BBBAAA orderings of paragraphs. These are just some of the other methodologies that could be considered in any further studies regarding text-type mixing.

DESCRIPTION OF THE THREE PASSAGES

As with the three texts analyzed in Chapter Two, the computer software program "Correct Grammar" was also used in assessing the three constructed passages according to the Flesch-Kincaid Grade Level, Gunning Fox Index, Grade Level Required and the Flesch Reading Ease Score.

Table 5 Properties of the Three Comprehension Passages

Properties	Passage A	Passage B	Passage C
# of words	1882	1810	1899
# of letters per word	5.0	4.9	4.9
# of syllables per 100 words	175	170	170
# of sentences	103	102	103
# of words per sentence	18.6	17.8	18.7
# of paragraphs	27	27	27
# of sentences per paragraph	3.9	4.0	4.1

Table 6 Readability Scores for the Three Passages

Passage A		
Flesch Reading Ease Score	39.9	Difficult
Grade Level Required	14	
US Adults Who Can Understand	33%	
Flesch-Kincaid Grade Level	12.2	
Gunning Fog Index	11.3	
Passage B		
Flesch Reading Ease Score	44.8	Fairly Difficult
Grade Level Required	13	
US Adults Who Can Understand	43%	
Flesch-Kincaid Grade Level	11.3	
Gunning Fog Index	10.5	
Passage C		
Flesch Reading Ease Score	43.7	Fairly Difficult
Grade Level Required	13	
US Adults Who Can Understand	43%	
Flesch-Kincaid Grade Level	11.8	
Gunning Fog Index	11.0	

The properties and readability scores of the three reading comprehension passages (see Tables 6 and 7 above) are similar to those of the three texts analyzed earlier. They are still within the fairly difficult range on the Flesch Reading Ease Score. Even though Passage II and III had been manipulated mixing text-types, the properties or factors assessed by the computer software program "Correct Grammar" do not vary overall across the passages. Basically, this can be attributed to the factors themselves since the focus of the software program is more on the raw quantitative aspect of a text (ie. number of sentences, number of paragraphs, etc) rather than on the linguistic patterns occurring within a text.

SUBJECTS

For the purposes of the study, participants who were considered literate readers and enculturated in many genres were desired. University students were drawn upon to make up the sample for this study since their presence at university is assumed to be indicative of their exposure to and familiarity to a variety of different genres.

The subjects of this study were first and second year students at Simon Fraser University. The 45 subjects were all volunteers, ranging in age from 17 to 22. There were 32 females and 13 males. There were 3 subjects who did not speak English when they started elementary school and of those three only one spoke a language other than English to his/her parent(s) at home.

Of the sample population, one subject from Group A, the group that was presented with the pure passage following the genre conventions of academic prose was not included in the final analysis. This was done because the subject's background (information taken from the subject's personal questionnaire) appeared to indicate that she was not representative of a literate reader of English who had over time internalized generic rules. She was the only subject who not only did not speak English when starting elementary school but did not currently speak English to at least one of her parents at home. Even though the subject had some prior knowledge (having taken a psychology course) that might have influenced her score, her score of 4 out of 20 on the reading comprehension test seemed to substantiate the conclusion

that although she was a second year university student, she was not considered a literate reader enculturated in many genres.

PROCEDURES

The subjects were tested individually or in small groups. They were initially asked to complete a personal questionnaire and a participation form (see Appendix E and F respectively). In addition to the general questions regarding age, sex, year of study at the university, each subject was questioned about his or her English language background (ie. Did you speak English when you started elementary school? Do you and at least one of your parents speak English when you converse at home?). Subjects were also asked about their background knowledge about the topics included in the passages used in the comprehension study. This was done to ensure that the presence of factors that could possibly influence the results obtained on the comprehension test was included in the final analysis.

Upon completion of the questionnaire, each subject was randomly presented with one of the three developed reading passages and given 20 to 25 minutes to read the text. At the conclusion of this time or when the subject felt he was ready (whichever occurred first), the reading passage was removed and the subject asked to complete a 20 multiple choice test based on the reading passage within 10-15 minutes although most participants did not require more than 10 minutes.

The items on the 20 multiple choice test originated from Cumming's *Test Item File* for the Woolfolk text. The inclusion of certain questions was based on the presence of information presented in the three passages.

STATISTICAL ANALYSIS

The data were analyzed in several steps. The data from the personal questionnaire and the results of the reading comprehension test were statistically analyzed to determine whether linguistic coherency within text types has an effect on reading comprehension.

CHAPTER FOUR

ANALYSIS OF RESULTS

INTRODUCTION

The results of this study's assessment of comprehension will be presented. The linguistic analysis of the three reading comprehension passages and the results of the comprehension test given to 45 university students after their having read one of three passages identical in content but differing in style (mixing text-types) will be the basis of the discussion.

LINGUISTIC ANALYSIS OF THE READING PASSAGES

The three reading comprehension passages were also linguistically analyzed according to Biber's model. (See Table 8). The mixed passages II and III have certain different frequency counts of linguistic features when compared to Passage I. These three specific features are first person (6), second person (7) and contractions (59). They have high frequencies in conversational genres.

Table 7 Frequency Count of each Linguistic Feature for each Comprehension Passage (Normalized to a Text Length of 1000 words)

Linguistic Feature*	Passage A	Passage B	Passage C
(1)	4.74	4.97	3.69
(2)	2.10	4.97	6.32
(3)	48.44	48.06	58.98
(4)	5.27	5.52	5.26
(5)	0.52	1.10	1.05
(6)	0.00	7.18	10.53
(7)	0.00	9.94	13.69
(8)	2.11	3.87	4.74
(9)	3.69	3.87	4.21
(10)	2.63	3.31	4.21
(11)	0.53	0.55	1.05
(12)	0.00	0.00	1.05
(13)	0.00	0.00	0.00
(14)	50.03	44.75	41.60
(15)	22.64	20.44	15.80
(16)	237.49	243.65	232.23
(17)	16.32	13.81	8.95
(18)	2.11	1.66	2.11
(19)	25.28	24.31	24.75
(20)	3.16	2.87	2.63
(21)	2.63	2.21	2.11
(22)	0.00	0.00	0.00
(23)	5.27	7.18	7.37
(24)	13.69	19.34	16.85
(25)	0.00	0.00	0.52
(26)	0.00	0.00	0.00
(27)	7.89	6.08	5.27
(28)	0.00	0.00	0.00
(29)	2.63	1.10	2.63
(30)	0.00	0.00	0.00
(31)	2.63	2.76	2.11
(32)	0.52	0.55	0.52
(33)	1.57	1.66	1.57
(34)	0.53	0.55	0.00
(35)	1.05	1.66	1.57
(36)	1.05	1.10	0.52
(37)	2.63	3.31	3.16
(38)	2.11	2.21	1.57
(39)	113.74	109.39	100.58
(40)	87.41	104.39	82.15
(41)	12.11	13.81	13.16
(42)	21.06	22.65	21.59
(43)	58.25	59.25	57.25

Linguistic Feature*	Passage A	Passage B	Passage C
(44)	5.00	4.90	4.90
(45)	2.11	1.66	1.05
(46)	1.57	1.10	2.11
(47)	0.00	0.00	0.00
(48)	0.00	0.55	1.05
(49)	0.00	0.00	0.00
(50)	0.00	1.66	1.57
(51)	9.00	7.18	7.90
(52)	5.27	6.08	5.27
(53)	1.57	1.10	0.52
(54)	1.05	1.10	1.05
(55)	1.05	1.66	1.57
(56)	0.00	2.76	4.74
(57)	1.57	1.10	1.05
(58)	0.00	0.00	0.00
(59)	0.52	8.8	13.70
(60)	0.00	0.00	0.00
(61)	0.00	0.6	1.05
(62)	0.00	0.00	0.00
(63)	0.52	0.00	0.52
(64)	13.16	13.8	12.11
(65)	3.16	3.3	6.85
(66)	0.52	0.6	0.52
(67)	3.69	3.9	5.26

*Linguistic features are numbered according to those cited earlier

COMPREHENSION TEST RESULTS

Forty-five first and second year university students were given one of three passages to read and subsequently twenty multiple choice questions to test their comprehension of the material presented them. Group A was given a passage that followed the conventions of academic prose. In contrast, groups B and C were given mixed text-type passages that mixed conventions particular to academic prose and those of the more informal genre of face-to-face conversations.

Results of the reading comprehension test are presented below in terms of the descriptive statistics for the groups. In the table below, the maximum score obtainable on the comprehension test is 20.

Table 8 Descriptive Statistics for the Groups of Subjects Participating in the Reading Comprehension Study

Group	# of Subjects	Mean	Median	Min	Max	Range
A	14	13.14	13	9	16	7
B	15	12.53	12	8	18	10
C	15	12.33	11	6	17	11
AB & C	44	12.57	12	6	18	12

Non-parametric tests were carried out on the data because the ranges for the scores for both Groups B and C reading the mixed-type passages were much wider than that for Group A. The extreme scores would have an undue influence on any analysis performed.

The sign test presents the number of scores above, below and equal to the median. In Group A, when only the 14 subjects are included in the analysis, there are 5 scores below, 3 equal to and 6 above the median 13. In group B, there are 6 above and below the median and one equal to the median of 12. And in Group C, there are 6 above the median 11, 7 below and 2 equal to the median.

The Wilcoxon signed-ranks test is used to assess the significance of the imbalance. The test statistic is calculated by first finding the difference between each pair of scores of Group A and B. The absolute values of the differences are then ranked, ignoring the sign. If two scores in a pair are the same (that is, if the difference is zero), that pair is ignored altogether. If two values of the difference are tied, they are given the mean of the ranks they would have had if they had been different in value. Each rank is then given the sign of the difference it corresponds to. The sum of the negative and the positive values are added and the smaller of these two sums is the test statistic W. For significance to occur, the calculated value must be smaller than or equal to the critical value. A non-directional .05 significance level is used in the study. The results of the Wilcoxon signed-ranks test are listed below in Table 10. In all cases, there is no significance.

Table 9 Wilcoxon signed-ranks test for Groups A, B and C

Group Pairs	N	W	Critical Value
Group A & B	14	36	21
Group A & C	15	54.5	25
Group B & C	15	44.5	25

An auxiliary study of the data available within the groups was also made. Within each group of the study, test scores and variables such as prior knowledge (the subject having previously taken any psychology courses), year of study at the university,

being able to speak English when entering elementary school, and speaking English at home to at least one parent were taken into account.

Table 10 Scores Obtained for Subjects in Group A

Subj. #	Year	Psych Courses	Spoke Engl. in Grade 1	Speaks Engl. to Parent(s)	Score Obtained
1	2	Yes	Yes	Yes	16
2	2	Yes	Yes	Yes	14
3	2	Yes	Yes	Yes	11
4	1	No	Yes	Yes	17
5	2	Yes	Yes	Yes	14
6	2	Yes	Yes	Yes	13
7	1	No	Yes	Yes	16
8	1	Yes	Yes	Yes	10
9	1	No	No	Yes	10
10	2	Yes	Yes	Yes	12
11	1	No	Yes	Yes	15
12	2	Yes	Yes	Yes	9
13	1	Yes	Yes	Yes	10
14	1	Yes	Yes	Yes	13

Within Group A, scores do not vary greatly in regards to year of study at the university. Even prior knowledge still does not greatly influence the overall test results of the subjects within the group. Little can be said about the relationship between scores and not knowing English when entering elementary school since there is only subject in this situation. The score obtained (10) does not really allow for any firm conclusions to be drawn although the score is quite good when considering the fact that the subject had previously taken no psychology courses.

Different inferences can be made within Group B (See Table below). For example, year of study at the university tends to positively influence the results of the comprehension test. In addition, prior knowledge also does, in general, positively influence the scores obtained. In the two cases where the subjects did not speak English when entering elementary school, the fact that there was prior knowledge and that both were second year students suggests that the scores were affected by these variables. The number of cases though tend to not permit one to unqualifibly state this however.

Table 11 Scores Obtained for Subjects in Group B

Subj. #	Year	Psych Courses	Spoke Engl. in Grade 1	Speaks Engl. to Parent(s)	Score Obtained
21	2	No	Yes	Yes	8
22	2	Yes	Yes	Yes	14
23	2	Yes	Yes	Yes	18
24	1	Yes	Yes	Yes	11
25	2	Yes	No	Yes	9
26	2	Yes	No	Yes	14
27	1	No	Yes	Yes	12
28	1	No	Yes	Yes	11
29	1	No	Yes	Yes	10
30	1	No	Yes	Yes	14
31	1	Yes	Yes	Yes	10
32	2	Yes	Yes	Yes	17
33	1	No	Yes	Yes	9
34	1	No	Yes	Yes	15
35	1	No	Yes	Yes	16

Within the final group, Group C, the first two interrelationships discussed in the last paragraph pertaining to Group B also pertain to Group C (ie. year of study and prior knowledge) (See Table 11). As to those three cases where the subjects did not speak English when entering elementary school, much lower scores such as 6 and 9 are obtained.

Table 12 Scores Obtained for Subjects in Group C

Subj. #	Year	Psych Courses	Spoke Engl. in Grade 1	Speaks Engl. to Parent(s)	Score Obtained
41	2	No	No	Yes	9
42	1	No	Yes	Yes	13
43	2	Yes	Yes	Yes	18
44	1	Yes	Yes	Yes	11
45	2	Yes	Yes	Yes	17
46	2	Yes	Yes	Yes	16
47	1	No	Yes	Yes	9
48	1	Yes	Yes	Yes	17
49	1	No	No	Yes	6
50	1	No	Yes	Yes	10
51	1	No	Yes	Yes	12
52	2	No	Yes	Yes	12
53	2	Yes	No	Yes	9
54	2	No	Yes	Yes	15
55	1	No	Yes	Yes	16

What is interesting to note is that when comparing within Group B those subjects having either no prior knowledge or not being able to speak English when entering elementary school respectively to those in Group C, there are in every case lower scores obtained within Group C.

If one had to conjecture the reason for this, the reason might lie in the comprehension test questions used and the location of the answers within the passages. Of the 20 questions asked the subjects, only five traced their correct responses within topic sentences of the passages. It would appear that subjects in Group C would therefore find the reading comprehension test more difficult because of this.

CHAPTER FIVE

THE RELEVANCE OF TEXT ANALYSIS FOR FIRST AND SECOND LANGUAGE TEACHING

INTRODUCTION

In this chapter some of the main points of the study are reviewed. The methodology used to transform texts and other possible methodologies are noted. Limitations of this study are also described, as well as implications of this study and implications for curriculum design and implementation. Finally, suggestions are made for further research.

DISCUSSION

The goal of this study was to investigate the effect of text-type mixing on first and second year university age students between the ages of 17 and 22. The research design itself was divided into two parts: initially addressing the premise that genre is rule-governed, and then administering a comprehension test after having given 45 university students one of three reading passages. One of the passages followed the conventions of the academic prose genre while the other two combined conventions of both the academic prose and the face-to-face conversation genres. It was hypothesized that comprehension would be affected by violating the rules by combining different conventions of genres.

The initial part of the study compared three texts taken from commonly used university texts in Educational Psychology based on a linguistic analysis using Biber's model of comparing texts along dimensions of linguistic variation. The results of the linguistic analysis substantiated the premise that genre is indeed rule-governed; texts can be defined based on patterns of co-occurring linguistic features. All three texts analyzed did have major similarities in the highest frequencies of certain linguistic features, namely nouns, prepositional phrases, attributive adjectives, type/token ratio, and present tense. The co-occurrence of these particular linguistic features comprise the "involved versus informational" linguistic dimension, one of the six linguistic dimensions Biber found in his linguistic analysis of both written and spoken texts.

The results of the final part of the study addressing the effect of text-type mixing on reading comprehension indicated that the reading comprehension levels of the subjects in Groups B or C who were exposed to mixed text-types were not significantly affected when compared to those subjects in Group A who were exposed to the pure or conventional passage.

The auxiliary focus on the results obtained within Groups B and C and the background of the subjects within the groups leads one to believe that the subjects seem to rely on prior knowledge and/or year of study at the university to help them comprehend text when expectations of certain internalized generic rules are violated. And finally, the position of the examples of the mixed text-type sentences within the two passages seems to have an effect

on the comprehension levels of the subjects. Where the pertinent information necessary to answer the multiple choice questions is located, being either in the topic sentence or secondary sentence, appears to influence comprehension.

STUDY LIMITATIONS

The major limitation not only of this study but of the study of text-type mixing in general is the lack of any previous experimental or explanatory research to date (to my knowledge) on text-type mixing. Most research appears to have focussed on the description of registers and the linguistic features present within these registers. It is however Biber's multi-feature/multi-dimensional approach to linguistic variation that provides a means of demonstrating that co-occurring syntactic forms are found clustered in various genres.

Another closely related limitation concerns the direction taken in the present study (ie. methodology used in creating the mixed passages). Because the particular methodology adopted in this study is just one attempt in investigating the effects of text-type mixing, it is recommended that other methodologies be considered before any final conclusions can be drawn regarding text-type mixing and its effect on reading comprehension. (See "Methodology Used to Transform Texts" in Chapter Three for discussion of other possible methodologies).

A final limitation, also specific to this study, is the small number of students involved. The strength and generalizability of

the statistical calculations were consequently reduced by this limitation. Were this study to be duplicated, a larger group should be tested. Not only should a larger group be tested but it would be beneficial to have more groups. For example, both native English speaking groups and second language speaking groups could be included across all (conventional and mixed) text-type situations.

IMPLICATIONS OF THIS STUDY

There are a number of important implications of this study for both text analysis within the field of linguistics and for curriculum design in the field of education. At this time, I will direct my attention to the field of linguistics and possible contributions.

The knowledge that genres are rule-governed can be applied to the analysis of naturally-produced oral or written texts. Biber's multi-feature/multi-dimensional approach to linguistic variation would allow linguists to quantitively evaluate and compare texts. Historical changes of text-types over time could also be noted just as there are records of spelling changes that have occurred over time (ie. Middle English versus current English). Changes in the communicative needs of a discourse community could result in new genres evolving. They could also be classified according to the co-occurring linguistic features. In addition, comparisons would not necessarily be restricted to within text-types since textual relations among different types of genres could also be investigated.

In addition, not only could comparisons within text-types be made, but textual relations among different types of genres could also be investigated.

The flexibility of Biber's model would accommodate for any variability since, unlike previous studies that treated linguistic variation in terms of dichotomous distinctions, Biber's model treats the variation in terms of continuous scales. These continuous quantifiable parameters of variation (dimensions) comprise those features that actually co-occur rather than what we expect to occur.

IMPLICATIONS FOR CURRICULUM DESIGN AND IMPLEMENTATION

Text analysis and the knowledge that genre is rule-governed have tremendous potential in education for first and second language teaching. Initially first language instruction will be addressed followed by a discussion of second language teaching and possible implications of text analysis and rule-governed genre.

In first language teaching, Littlefair (1989) believes that while the emphasis in reading has been on the cognitive aspect of learning to read, it is also important to consider the linguistic aspect involved. It is necessary to understand the linguistic way in which an author has constructed meaning. Lemke (1988) also suggests this when he states that the mastery of a subject in terms of both comprehension and active use of linguistic forms is basically a mastery of how to separate and combine its

characteristic semantic patterns with its specialized genre structures.

Another concern other researchers such as Painter (1989) and Martin and Rothery (1986) have expressed regards the preponderance of the narrative genre within the classroom. They feel that other genres that would be beneficial to later activities are being neglected by overemphasis on the narrative genre.

Students should be presented with examples of various genres, even at an early age. With the aid of the teacher, certain linguistic features that can be identified with particular genres can be highlighted. This would not have to be done in a dry manner but could be implemented by having students search for certain types of words. For example, if a teacher wanted the class to become aware of the presence of passives within a particular text, he could ask the class to go through the text and underline the occurrences of this linguistic feature. An understanding of the technical linguistic background would not be necessary. In this way even elementary students could learn to distinguish between different text-types.

In order to understand how rule-governed genres might influence curriculum design in second language instruction, a brief overview of some of the language acquisition theories and resulting methodologies should be examined.

The earliest period in second language instruction, which lasted until World War II, focused solely on translating text from and into the target language. The particular method of instruction

used during this period is commonly referred to as the grammar-translation or traditional method.

The second period coincided with and was influenced by the need for highly fluent speakers during the war. Swaffar, Arens, and Byrnes (1991) refer to the paradigm of this time as the "normative input/language replication" paradigm. It defined foreign language learning as the ability to meet an absolute standard of grammatical correctness and measured student performance strictly based on vocabulary and grammar mastery. This language-centered approach required learners to perform tasks utilizing preselected grammatical structures and vocabulary items to fill in slots or engage in pattern drills. Dominant theories of this paradigm are the linguistic theory of structuralism and the learning theory of behaviourism.

Developments in the foreign language research of the seventies caused a major paradigm shift from the prevailing "normative input/language replication" paradigm of the fifties and the early sixties. Research in various disciplines supported the feasibility of the "authentic input/language creation" paradigm. Work in discourse analysis and artificial intelligence confirmed that the need for knowledge of particular vocabulary and syntax changes greatly with subject matter or social demands. Cognitive sciences supplied evidence that creativity --rethinking and reformulating language-- promoted deeper processing. Swaffar, Arens, and Byrnes (1991) mention Bartlett and his schemata theory that proposes that humans learn by using cognitive strategies to integrate prior knowledge and information.

Unlike the old paradigm that ignored learner cognition and affect in its definition of language, the "authentic input/language creation" paradigm viewed language as being a creative process which occurs within a social context. Consequently, second language learning was redefined as the ability to perceive and operate within real-world situations, in order to perform real-world tasks. It was recognized that knowledge of grammatical forms and structures alone did not adequately prepare learners for effective and appropriate use of the language they were learning. The communicative approach of this time evolved into a basis for culturally and socially responsive language teaching that did not prescribe a particular teaching methodology.

Humanistic approaches focused not on just language teaching but also on helping students develop themselves as people. This led to such methodologies such as Community Language Learning, Suggestopaedia, The Silent Way and Total Physical Response.

Having looked at some of the existing language acquisition theories and resulting methodologies, how might the results of the present study influence language teaching?

A genre-based approach would differ from other approaches. While language-centered approaches would likely require learners to perform tasks by utilizing preselected grammatical structures and vocabulary items; and learner-centered approaches to also draw the learners' attention to functional or notional properties of languages, and in the case of learning-centered approaches to focus the learners' attention on negotiation of meaning, the genre-

based approach would focus the learners' attention on rhetorical action and on the organizational and, in particular, for purposes of this study on the linguistic means of its accomplishment (Swales, 1990). The goal of the learner within a genre-based approach would be to become a member of a chosen discourse community via effective use of established genres within that community. This is a significant change from other approaches since the actual teaching of English is not focused upon; nor is the focus on the learner. It is instead on illustrative texts of a particular genre and their rhetorical effects. Activities within the genre-based approach could include the analysis and critiquing of existing texts and later the composition of similar texts.

This approach to enhancing reading comprehension could also be directed to the field of writing. The student could appreciate the effectiveness of using varying text-types for their own specific communicative purposes.

What is important to note here, is that, for the second language learner, it is not the lack of the vocabulary but usually the lack of knowledge of the appropriate formal schemata which leads to poor comprehension.

In conclusion, encouraging the understanding of the way in which different genres are constructed linguistically is not a plea for a restricted, prescriptive teaching of reading or writing. As Smith and Hillocks (1988) argue, the function of genre conventions is essentially to establish a contract between the writer and the reader so as to make certain relevant expectations operative and thus to permit both compliance and deviation from accepted modes

of intelligibility. Knowledge of genre conventions facilitates the student's ability to become enculturated in the various genres much sooner than he might be able to at this time. It would allow him access to the different genres and allow him to move between them.

SUGGESTIONS FOR FURTHER RESEARCH

The present study revealed a potentially rewarding area for further research. Firstly, because the area of text-type mixing is so new, more studies need to examine the effect various other methodologies that could be used to transform texts into mixed text-type passages has on reading comprehension.

A follow-up study to this thesis could focus on whether reading comprehension improves through awareness of or promoting awareness of the presence of co-occurring linguistic features within texts. It would be interesting to explore the effect of focussing the student's attention to the linguistic means by which rhetorical action is accomplished has on reading comprehension. In addition, the closely related field of writing instruction could also investigate the effect awareness of certain linguistic constructs being present in certain texts has on the writer's ability to develop appropriate texts. This could be accomplished by comparing examples of various genres generated by the learner prior to instruction and after instruction.

Results from these studies could eventually lead to new methodologies for language teaching. Of course, these studies

would not be restricted to first language learning/teaching but could prove extremely beneficial to second language learning/teaching.

CONCLUSION

This study has endeavored to investigate genres defined strictly on the basis of linguistic criteria and the effect genre or "text-type" mixing would have on reading comprehension.

The results of the multi-feature/multi-dimensional linguistic analysis corroborated Biber's own finding that certain co-occurring linguistic features appear to be rule-governed and that there is an internal linguistic coherency of text-types.

The question regarding text-type mixing and its effect on reading comprehension was similarly addressed. The results of the reading comprehension test did not support the hypothesis that text-type mixing would affect comprehension. Although the findings of no significant effect are specific to the research design used in this study, additional research is definitely required before we can conclusively state that text-type mixing has no effect on reading comprehension.

This study can be compared to a child's first tentative step in the area of text-type and text-type mixing and hopefully others will continue investigation into this area.

Certainly Biber's multi-feature/multi-dimensional model of linguistic analysis should prove exciting to ESL (English as a Second Language) as a foundation for further research into how

learners can more easily access their desired discourse communities.

The following saying by Wittgenstein clearly sums up the relationship that exists between a person, language and the genres available to him.

> "The limits of my language mean the limits of my world". (Clark, Eschholz, Rosa, 1981)

APPENDIX A

STANDARDIZED AND TEACHER-MADE MEASUREMENT INSTRUMENTS*

Orientation

Throughout the next two chapters the terms **test**, **measurement**, and **evaluation** will be used repeatedly. A test is a measure containing a series of questions, each of which has a certain correct answer. The term measurement is a broader concept because educators and psychologists can measure characteristics in ways other than giving tests (e.g., observations and rating scales). Measurement is the process of assigning numbers to behavior according to certain rules or specifications to determine differences among individuals on the behavioral characteristics being measured. Finally, evaluation is the process of obtaining information to form judgments so that educational decisions can be made. Evaluation is the most comprehensive of the three terms and always included value judgments regarding the desirability of the data collected.

The next two chapters focus on formal measurement instruments and evaluation procedures. In previous chapters I have pointed out numerous occasions when teachers make instructional decisions using informal classroom evaluation procedures. Teachers form judgments and make decisions during such activities as oral questioning, class discussions, and observations of student social interaction and work habits.

Gronlund (1985a) provides examples of the types of instructional decisions teachers face and the type of evaluation information (in parentheses) that might be helpful in answering the questions:

1. How realistic are my teaching plans for this particular group of pupils? (scholastic aptitude tests, past records of achievement)

2. Should students be grouped for more effective learning? (range of scholastic aptitude and achievement scores, past records of achievement)

3. To what extent are the pupils ready for the next learning experience? (readiness tests, pretests of needed skills, past records of achievement)

4. To what extent are pupils attaining the course's minimum essentials? (mastery tests, observation)

---

5. To what extent are pupils progressing beyond the minimum essentials? (periodic quizzes, general achievement tests, observation)

6. At what point would a review be most helpful? (periodic quizzes, observation)

7. What type of learning difficulties are the pupils encountering? (diagnostic tests, observation, pupil conferences)

8. Which pupils should be referred for counseling, special classes, or remedial programs? (scholastic aptitude tests, achievement tests, diagnostic tests, observation)

9. Which school mark should be assigned to each pupil? (review of all evaluation data)

10. How effective was my teaching? (achievement test, pupils' ratings, supervisors' ratings) (p. 4)

These questions are not meant to be an all-inclusive list of evaluation issues that teachers encounter during the year. However, the questions do cover a broad range of concerns that need to be addressed during the instructional process. They can be helpful for you to keep in mind as you read the next two chapters. Also, remember that these questions do not occur in any particular order but are interrelated in the teaching-learning process.

After reading this chapter, you will be able to identify characteristics of good tests, distinguish between different types of tests and measurements, analyze controversial issues regarding standardized testing, and write classroom tests to evaluate instructional objectives. In Chapter 13 you will learn how to analyze, interpret, and report test scores.

Before going further in this chapter, let us think about possible student perspectives on the measurement and evaluation process. When students hear the term evaluation, the first thought that they often have concerns tests and or grades. By the time students reach college, they will have taken hundreds of examinations--some welcomed, some feared. Unfortunately, most of us probably remember more of the anxiety engendered by examinations than we do the joys. Can you remember how you felt as a young child waiting for the teacher to return a test or homework assignment? What about the sleepless night before an important examination in college or the wait for a postcard reporting a final grade?

Students usually do not understand why they have to be involved in so much testing. It is often useful to take the time to explain the purpose of measurement activities so that students better understand how you will use the information. Finally,

consider how you policies and procedures regarding measurement and evaluation may influence students' achievement, motivation, and classroom behavior.

THE USE OF EVALUATION IN CLASSROOM INSTRUCTION

In Chapter 1, I stated that teachers are involved in evaluation throughout the teaching-learning process. The examples of instructional decisions identified in the preceding section illustrate this involvement: Teachers must make some decisions before beginning instruction (e.g., How realistic are my teaching plans for this particular group of pupils?), other decisions, during instruction (e.g., To what extent are the pupils progressing beyond the minimum essentials?), and still others, after instruction (e.g., Which school mark should be assigned to each pupil?)

Airasian and Madams (1972) developed a classification system for describing evaluation procedures for various instructional purposes that helps to organize the types of question that concern teachers in planning, implementing, and evaluating the outcomes of instruction. The systems includes four functions of evaluation--placement (to determine student performance at the beginning of instruction), formative (to monitor learning progress during instruction), diagnostic (to diagnose learning difficulties during instruction), and summative (to evaluate achievement at the end of instruction). (I first introduced formative and summative evaluation in Chapter 8 when describing mastery learning.) Gronlund (1985a) provides a summary of each type of evaluation.

Placement Evaluation

Placement evaluation is concerned with the student's entry behavior before the beginning of instruction. A number of important questions need to be answered at this stage of instruction: Does the student have the needed knowledge and skills to begin instruction? Has the student already mastered the objectives of the particular lesson or unit? Does the student's ability, learning style, attitude, or interest indicate that the child would benefit from a particular method of instruction? To answer these questions, teachers use a variety of instruments such as readiness tests, aptitude tests, pretests on course objectives, and observational techniques. In summary, placement evaluation concerns relate to placing students in the proper position in the instructional sequence and providing the most beneficial method or mode of instruction for each student.

Formative Evaluation

Formative evaluation is used to provide on-going feedback to the teacher and student during instruction regarding success and failure. This feedback is helpful in deciding whether changes in subsequent learning experiences are needed and in determining specific learning errors that need correction. Formative evaluation depends on the development of specific tests to measure the

particular aspect of instruction that is covered. For example, if the teacher spends two weeks covering World War I in a history class, it may be useful to determine students' knowledge after the first week in order to decide what should be reviewed and who the second week of instruction should be approached. In this case, a specially designed test on the material in the first week is necessary to provide the needed information. Observational methods, in addition to paper-and-pencil tests, are often useful in monitoring student progress.

Diagnostic Evaluation

Diagnostic evaluation is used when formative evaluation does not answer all the questions regarding problems students have with certain instructional objectives. For example, why is it that Steve cannot divide by two digits? Why does Susan confuse certain letters in reading? Diagnostic evaluation searches for the underlying causes of learning problems in order to formulate a specific plan for remedial action. It involves special diagnostic instruments as well as observational techniques.

Summative Evaluation

Summative evaluation generally comes at the end of instruction to determine how well students have attained the instructional objectives, to provide information to grade students, and/or to evaluate teacher effectiveness. This type of evaluation usually includes achievements tests, rating scales, and evaluations of student products. Some educators believe that teachers overemphasize this category of evaluation while neglecting the importance of the other three categories in improving student achievement. It is important to remember that giving students grades is only a small part of the measurement and evaluation process.

JUDGING STUDENT LEARNING: NORM-REFERENCED AND CRITERION-REFERENCED MEASUREMENT

In evaluating achievement, you can interpret student learning in one or both of two ways: (1) in terms of performance relative to a group, or (2) in terms of performance relative to a behavioral criterion of proficiency. The first method is called **norm-referenced measurement**; the second, **criterion-referenced measurement**.

Norm-referenced measurement is the most common method of testing used by teachers and thus receives the most attention in this book. Students' scores on most tests reflect their performance as compared to that of their classmates. Grading on a curve is an example of norm-referenced measurement.

In recent years, *criterion-referenced measurement* has gained in use. With this method, the performance of a student is measured in terms of the learning outcomes or objectives of the course. The statement, "Lisa did better than 80 percent of the

students on a test of biological terminology" is a norm-referenced measurement. In criterion-referenced measurement, we might say that Lisa correctly answered 80 percent of the items on a test of biological terminology. Eighty percent means something quite different in each situation.

In criterion-referenced testing, the teacher concentrates on a limited number of specific objectives. Explicit instructional objectives are necessary because each test item must correspond to a particular objective or criterion. The following objectives and corresponding test items reflect this relationship:

In criterion-referenced measurement, the teacher is primarily concerned with how many items of a set of specific objectives a particular student has mastered. In norm-referenced measurement the test items are written to reflect the objectives and content in a more diffuse manner and result in a large spread of scores, which is necessary to rank students reliably in order of achievement. Criterion-referenced measurement does not aim for a wide range of scores because the purpose is to have all students master the objectives.

The actual construction of norm-referenced and criterion-referenced tests is similar, both essay and objective items being used. The difference lies in the purpose of the tests.

Criterion-referenced testing tends to be used more in individualized programs like Project PLAN and mastery learning programs (discussed in Chapter 8) when the instructional intent is to raise almost all students to a specified level of achievement. Presently, classroom instruction uses this testing to greatest advantage when the learning outcomes are cumulative and progressively more complex, as in mathematics, reading, and foreign language, and when minimum levels of mastery can be established. When the subject matter is not cumulative, when the student does not need to reach some specified level of competence, and when tests measure success in comparative steps, norm-referenced testing is preferred.

CHARACTERISTICS OF A GOOD TEST

Anyone can develop an instrument and state that it serves a specific function or purpose. For example, we often find pamphlets in drug stores describing how for a small fee individuals can determine their personality profiles by completing a small questionnaire. In most cases, these questionnaires cannot accurately measure personality or any other trait. Fortunately, evaluation specialists have developed specific criteria for judging the quality of various measurement instruments. Test specialists are expected to provide detailed information concerning these criteria. Two of the most important criteria for evaluating instruments are **validity** and **reliability**. Each criterion is discussed in terms of norm-referenced measurement.

Test *validity* refers to the appropriateness of the interpretation of the test scores with regard to a particular use. For example, if a test is to be used to measure reading comprehension, it should measure reading comprehension and not measure other irrelevant factors not associated with comprehension. The more a test can be shown to accomplish this goal, the more valid is the test.

It is important to realize that validity refers to the appropriate use of test scores and not to the test itself. Also, validity is a matter of degree because validity does not exist as an all-or-none characteristic. Finally, validity is specific to some particular use because a test is not valid for all purposes (Gronlund, 1985a). For example, a mathematics test may have high validity for measuring mathematical reasoning but have low validity for measuring computational skills.

In the past, validity was separated into three types--content, criterion-related, and construct. However, the most recent revision of the *Standards for Educational and Psychological Testing* (1985) takes the position that validity is unitary and that each of what used to be considered a separate types of validity should be thought of as approaches to establishing validity. Messick (1981) points out that this change was made because the notion of separate types of validity gives the impression that the types are equal and that any one type could be used by test developers to establish validity.

Content validity aims for an adequate sampling of a specified universe of content. In constructing a test, the teacher should include test items that sufficiently represent the instructional objectives of the unit and the subject matter. Many students, after analyzing the test questions in their courses, often remark that some material was not even tested on the examination. Such a test may have low content validity. Later in this chapter, some guidelines to help you ensure content validity when constructing your own tests will be identified.

Criterion-related validity is concerned with two questions: How well does the test judge present ability? How well does the test judge future ability? Let us take two examples of tests: A typing test determines an applicant's ability as a secretary; the Scholastic Aptitude Test (SAT) predicts a student's future academic performance in college. In the first example, we want to estimate present ability, and we obtain the relationship between the test score and performance at the same time. A high correlation in this case would indicate that the typing test is a good indicator of typing skills on the job. This type of criterion-related validity is often called *concurrent validity*. In the second example, we are interested in prediction, and we extend the criterion related to the test scores over a period of time (usually a student's grade-point average during the first semester in college). This type of criterion-related validity is called predictive validity. Time is the variable in the two types of criterion-related validity. The typical procedure for reporting criterion-related validity is by the use of a validity coefficient, which reveals the correlation, or relationship, or relationship, between the test and the criterion. Although the criterion could be another test, it usually is some other type of performance indicator.

The third approach to validity is **construct validity**. Psychologists develop tests to measure traits and abilities such as intelligence, anxiety, creativity, and social adjustment. These traits are also called *constructs*. The tests determine the "amount" of the trait or construct a person possesses.

Although a test may exhibit more than one approach to validity, some tests are constructed with one major source of validity in mind. Content validity is an earmark of teacher-made tests and standardized achievement tests. Criterion-related validity is the main factor in the scholastic aptitude tests used to predict school or college success. Construct validity justifies the use of a test for measuring specific psychological traits or abilities.

Reliability

Reliability is a qualitative judgment about the *consistency* of test scores or other evaluation results from one measure to another. It is difficult to make good judgments on the basis of unreliable test scores. The methods available for determining reliability are evidence of the recognition that there are different types of consistency. Two important kinds of reliability are consistency over time and consistency over different forms of an evaluation instrument.

Consistency over time is often referred to as **test-retest reliability**. In this procedure, individuals take the same test at two different times, and the results of the tests are compared by using a correlation coefficient. If the results are stable--students who score high on the first administration score high on the second administration, and low achievers score low both times--this consistency will be indicated by a high correlation coefficient (see Figure 12.1).

Consistency over different but **equivalent forms** of the *same* test is determined by administering these forms to the same students in close succession and then correlating the resulting scores. The correlation coefficient so obtained provides a measure of equivalence indicating the degree to which both forms of the test measure the same aspects of student behavior. A high correlation coefficient would indicate that either test could be used to measure students' knowledge of the material. A low correlation coefficient would indicate that the two forms of the test *do not* measure the same material or that they differ in the degree of difficult. This type of reliability is important when using different forms of a test to measure the growth of students over a period. For example, students can be given form A of an achievement test in September and form B at the end of the year. If a history teacher has many sections of the same class, the teacher may want to sue two forms of the same test. However, unless the two tests cover the same material with a similar level of difficulty, the procedure will be unfair to some students. The teacher should evaluate the equivalent form reliability of the tests before using them.

In general, teachers can improve the reliability of their tests by (1) including items that discriminate among students, so that almost no one gets all the items correct or incorrect, (2) using objective scoring procedures, and (3) including a sufficient number

of test items. For example, it is generally recommended that multiple-choice tests include at least 35 to 40 items.

In thinking about reliability and validity, remember that a test can be reliable without being valid. A valid test must be reliable. For example, your instructor could give a spelling test for the final examination in this course. The instructor could show that the test produces relatively consistent scores on test-retest reliability. However, does the test measure knowledge of educational psychology? Is the test valid? Obviously not! If you can determine that a test has high degree of validity, then you can be assured of a reasonable level of reliability.

We are now ready to explore the types of measurement instrument used by teachers in making placement, formative, diagnostic, and summative evaluations. Remember, no matter what the function or purpose is of the instrument being used, the goal is to select or develop the most valid and reliable instruments possible.

STANDARDIZED TESTS

One way to improve the validity and reliability of tests is to construct the test adequately and to make the test situation as similar as possible for all students. This process is called standardization. **Standardized tests** include the following characteristics:

* They are commercially prepared by measurement experts who have carefully prepared and studied all test items.
* They measure various aspects of human behavior under uniform procedures.
* They include a fixed set of questions with the same directions, timing, constraints, and scoring procedures.

Test publishers provide a reference, or *norm*, group of students who have already taken the test so that teachers can compare their students' performances with those of other students in the state or nation. Thus, test norms provide a standard for comparing an individual's relative level of performance on a specific test. Test norms usually are provided in tables in the test manual.

Norms are established as part of the standardization process by administering the test to representative groups of students for whom the test was constructed. For example, if a test is designed to measure the arithmetic achievement of seventh-and eighth-graders, the test specialists would obtain scores of representative seventh-and eighth-graders from many regions and schools throughout the country. The size and sampling procedures for selecting schools and students differ from one test to another. As a result, it is important for the educational committee responsible for selection tests in a school district to evaluate the appropriateness of the norm or comparison group provided by the test publishers before selecting a test. In some situations, a test may have as a norm a group (e.g., students in urban areas) with which it would be inappropriate to compare the performance of students in another group (e.g., students in a rural school). Most test publishers who distribute achievement tests nationally undertake large-scale

norming procedures that involve representative groups from all geographical regions of the country and from school of different sizes and cultural and ethnic composition.

The two types of standardized tests most used in school evaluation programs are the aptitude test and the achievement test. An aptitude test measures a student's potential for success in learning. Individual intelligence tests and group scholastic aptitude tests are the most widely used aptitude test. They are useful in comparing a students' actual achievement, in the form of school grades, with their potential achievement. The results of an aptitude test often indicate that a student has the potential t attain a higher achievement level than is presently being achieved. Aptitude tests are often used in placement evaluation.

An achievement test measures how much of the academic content a student has learned in a particular grade level or course. Various forms of achievement tests are used in placement evaluation (to determine whether students have the needed knowledge to begin instruction), in diagnostic evaluation (to determine the underlying causes of learning problems), and in summative evaluation (to determine how well students have attained the instructional objectives). A school district can evaluate its instructional program by comparing students' achievement test results with those of other students in the state or nation. For example, if the students in a particular high school score consistently lower than the norm in mathematics, the school district should evaluate its curriculum and instructional methods.

Other standardized tests used to a lesser extent in school are interest, attitude, and personality inventories. Guidance counselors use interest inventories to help them in vocational counseling. Attitude and personality inventories identify factors that may influence study habits, motivation, and adjustment in school, and they can be helpful in furthering teachers' and counselors' understanding of their students.

Because most of the standardized tests used in school are aptitude and achievement test, this chapter focuses on them. If you are interested in learning more about other types of test, see the list of suggested readings at the end of this chapter for basic textbooks on measurement and evaluation that include discussions on a wide variety of standardized tests.

APTITUDE TESTS

Individual Tests of Intelligence

The theory and issues of intelligence were discussed under the subject of individual difference in Chapter 2. When educators refer to an **individual intelligence test** they usually mean the Stanford-Binet or Wechsler test, which are administered to a single student at a time by a *trained* examiner, usually a school psychologist.

The Binet, as it is often called, yields a single intelligence test score. It employs separate sets of items for different age levels. Examples of tasks at the 6-year level follow (Biehler, 1974a):

* First test: Defining at least six out of a list of forty-five vocabulary words. (These are arranged in order of difficulty. Testing stops after six consecutive failures.)
* Second test: Explaining the differences between tow objects, such as wood and glass. Must get two out of three.
* Third test: Telling which feature is missing in a series of "mutilated pictures." Must get four out of five.
* Fourth test: Picking out three, ten, six, nine, and the seven cubes form a pile of twelve, to demonstrate the ability to count. Must count at least four of the five trials correctly.
* Fifth test: Completing analogies such as "A bird flies, a fish...." Must get three out of four.
* Sixth test: Solving two out of three simple pencil mazes by tracing the proper pathway. (p. 585)

From these examples it is evident that verbal skills play an important role in success on this test. Because schools emphasize verbal abilities, we find that the Binet correlates well with academic achievement.

Other individually administered intelligence tests are the Wechsler Intelligence Scales: Wechsler Preschool and Primary Scale of Intelligence (WISC-R), and the Wechsler Adult Intelligence Scale (WAIS). They differ from the Stanford-Binet test on two counts. First, the tests are not constructed by specific age levels. Second, the tests yield scores from two sections--verbal and performance--as well as a combined IQ for the whole test.

The WISC-R consists of 12 subtests. The verbal score is measured by the following subtests: General Information, General Comprehension, Arithmetic, Similarities, and Vocabulary, with Digit Span included as an alternate test. The performance score is determined by Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Coding, with Mazes as an alternate test. The total IQ is computed from the combination of the verbal and performance subtests.

Students who take both the Stanford-Binet and the WISC-R intelligence tests will probably come out with a slightly different score for each. As you might expect, the verbal section of the WISC-R relates more closely to the Stanford-Binet is more heavily weighted with questions that measure verbal ability. Bilingual children, children from homes in which English is seldom spoken, and poor readers may have trouble with the Stanford-Binet. The WISC-R covers a broader range of abilities than does the Stanford-Binet and is more useful in clinical evaluation, especially in the diagnosis of brain and neurological disorders.

The Test Report. After an individual intelligence test had been completed, the examiner who has administered it scores the responses in the test booklet, computes the subject's intelligence score, and prepares a report on the subject's test-taking behavior. This last bit of information is sometimes more helpful to the teacher than the reported intelligence score. The subject's attitude toward the testing situation is telling: Was attention paid to the examiner's directions? Did the child answer the questions immediately or ponder over them? Did the child become easily

frustrated when difficult questions were asked? Did the child appear to be anxious during the testing session? The answers to these questions may answer the teacher's questions about academic placement and remedial help. At the least, they may help the teacher to understand classroom learning problems.

Cautions in interpreting intelligence tests. Many problems in intelligence testing have arisen from the false belief that test scores are static. Even in the most perfect testing situations, test scores on the same individual can fluctuate many points. The **standard error measurement**--discussed in the next chapter--is helpful in understanding this phenomenon. Some guidelines for interpreting individual test scores (based on Gronlund, 1985a) follow:

1. An intelligence test score from the same test can be expected to vary from 5 to 10 points. This, a score of 100 can be interpreted as a band of scores ranging from 95 to 105.
2. When intelligence test scores from different tests are compared, it is not unusual for the scores to differ among the tests. Also, since each test measures slightly different aspects of mental ability and is developed using different populations, the scores are not directly comparable. For this reason, it is important to know which test was used to measure intelligence.
3. The intelligence test scores of elementary students vary more than the scores of high school students. This is because mental abilities tend to be more stable as children grow older. Also, variation in group tests is generally greater than individual tests because it is more difficult to control the factors necessary for maximum performance (e.g., attention).

The foregoing variations occur in normal testing situations. However, greater problems in interpreting and using intelligence test scores occur when outside factors introduce additional possibilities of error into scores. In general, scholastic aptitude and intelligence test scores are less dependable for the following types of students (Gronlund, 1985a):

1. Those whose home environment does not provide the opportunity to learn the types of task included in the test.
2. Those who are little motivated by school tasks.
3. Those who are weak in reading skills or have a language handicap.
4. Those who have poor emotional adjustment. (p. 308)

Scholastic Aptitude Tests

Most of the testing in school is done on a group basis. *Scholastic aptitude tests*, like achievement test, are usually administered to a large number of students at one time by persons with relatively little training in test administration.

The use of group tests above the first two to three grades requires that the students be able to read the questions and the

choice of responses and indicate their responses on a special answer sheet provided. The types of test item and their level of difficulty are appropriate to various ages and grade levels.

Some group aptitude tests provide a single score similar to that of the Stanford-Binet, whereas others provide two or more scores based on separate types of mental ability. Examples of single-score tests are the Henmon-Nelson Tests of Mental Ability (grades 3-12) and the Otis-Lennon Mental Ability Test (grades K-12).

Tests representative of a two-factor approach are the Lorge-Thorndike Intelligence Tests (grades 3-13), the California Test of Mental Maturity (CTMM) (grades K-16), and the Kuhlman-Anderson Measure of Academic Potential (grades K-12). Figures 12.2 and 12.3 include sample test items from the Lorge-Thorndike Intelligence Tests. The verbal battery of tests is composed of five subtests: (1) vocabulary, (2) sentence completion, (3) arithmetic reasoning, (4) verbal classification, and (5) verbal analogy. The nonverbal battery consists of three subtests: (1) pictorial classification, (2) numerical relationships, and (3) pictorial analogy.

Tests of scholastic aptitude, or academic aptitude, do not measure inherited capacity but learned abilities or school success. The names of group tests often cause confusion over what they measure. Do not make a decision to use a test on its name alone, for it may turn out to measure something entirely different from what you intended it to.

Although group tests are economical and easy to administer, they have some disadvantages. First, the uninterested and unmotivated student may score lower on a group test than on an individually administered test because there is no individual examiner to focus specifically on that student's responses. ** Second, group tests rely heavily on paper-and-pencil pencil items and emphasize speed, verbal ability, and reading comprehension to a much greater extent than do individual tests. Last, the scoring is completely objective and allows no room for judgment of test-taking behavior. It is not uncommon for a teacher to ask for the administration of an individual intelligence test when information about a student appears to be inconsistent with the group test result.

ACHIEVEMENT TESTS

The measurement of educational achievement is an important part of developing effective educational programs because not all methods and procedures are uniformly successful. As a result, students, teachers, parents, and school officials need to know how successful their efforts have been, so that decisions can be made regarding which practices to continue and which to change. Two types of achievement test provide information about student progress--*standardized* and **teacher-made tests.**

Standardized achievement tests are constructed by test publishers for wide distribution in schools throughout the country. As a result, the content coverage must be broad enough to include

the basic content taught in many schools. As mentioned earlier, scores are interpreted with reference to how well a given student achieved in comparison to a state or national sample of students at the same age or grade level. Teacher-made achievement tests are constructed by the classroom teacher, or sometimes by several teachers, to measure the specific curriculum of a particular course in a particular school. Scores are interpreted with reference to a student's classmates. I first focus on standardized achievement tests and discuss teacher-made tests later in the chapter.

Standardized Achievement Tests

The standardized achievement test can be classified by content area and function (brown, 1983). Some achievement tests measure knowledge of arithmetic; others, history, physics, and other school subjects. Many achievement tests are batteries, measuring many content areas rather than only one area. The teacher can then compare test scores on the separate subtests to determine the relative strengths and weaknesses of students in the areas covered by the test.

Elementary school achievement tests focus on the basic skills taught in school: reading, language, mathematics, and study skills. Some tests also include measures of achievement in social studies and science. High school curricula vary more than do elementary school curricula, and as a result, it is more difficult to develop achievement test batteries that have high content validity in all high schools. Therefore, in addition to using test batteries, high school teachers can select separate tests from specific content-oriented tests to measure achievement in English, social studies, science, or mathematics. Another approach is to use tests that measure general educational development and do not depend on any particular courses for their questions. One such test battery is the Sequential Tests of Educational Progress (STEP II). The unique aspect of this test is that it emphasizes application, interpretation, and evaluation of academic content to a greater extent than do tests that measure basic skills. Because the questions are not derived from specific course content, this test provides a fairer measure of achievement for students with different educational experiences.

Select your achievement tests with care. First, although the tests have similar titles, they often differ in how much emphasis they place on the various skills measured. For example, the specific arithmetic skills tested on two achievement tests may differ. One might emphasize computational skills and the other, problem-solving abilities. Second, standardized achievement tests measure only a portion of the knowledge and skills taught in school, so be sure to choose the most appropriate test for your school's curriculum.

You can locate information about any standardized test from Bourses *Tests in Print and Mental Measurements Yearbooks*. The latter provides reviews of tests by measurement experts. You can order specimen sets of the examinations that appear in Buros from test publishers. These sets include a copy of the actual examination booklet and test manuals that evaluate the validity

and reliability of the test. Check the test items against the course content in the areas to be tested to determine whether the test adequately measures the objectives of the school.

The following are some achievement test batteries commonly used in schools:

- California Achievement Tests (grades K-12)
- Iowa Tests of Basic Skills (grades K-9)
- Comprehensive Tests of Basic Skills (grades K-12)
- Metropolitan Achievement Tests (grades K-12)
- SRA Achievement Series (grades K-12)
- Stanford Achievement Tests (grades 1-12)

A number of special types of achievement tests are particularly useful in making placement and diagnostic evaluations. These instruments include the diagnostic, readiness, and individual achievement tests.

Diagnostic Tests. A special type of achievement test used in schools is the diagnostic test, which is designed to assess those skills and abilities that are important in learning a particular subject, usually reading or mathematics. The Durrell Analysis of Reading Difficulty, for example, covers student performance in silent and oral reading, listening, comprehension, word analysis, phonetics, pronunciation, writing, and spelling. The teacher can recommend remedial instruction to overcome the learning problems identified by the test.

There are two important differences between a diagnostic test and a general achievement test. First, the diagnostic test analyzes knowledge in a single subject area in depth, whereas the general achievement test collects information on the distribution of knowledge across many subject areas. Second, the diagnostic test includes a larger proportion of items that are relatively easy to answer, the better to assess below-average performance, whereas the achievement test includes a wider range of items from very easy to very difficult, the better to assess a greater ability span for different grade levels.

Readiness Tests. Sometimes teachers need to know whether students are ready for certain learning tasks. Reading readiness tests are most commonly used in elementary school, but other types of test are also used. The reading readiness tests are used to determine whether the student has the necessary knowledge and skills to begin reading. They measure such skills as visual discrimination (identifying similarities and differences in works, letters, pictures), auditory discrimination (identifying similarities and differences in spoken words and sounds), and verbal comprehension. Other readiness tests measure basic concepts and skills that are important for school success. One example is the Boehm Test of Basic Concepts, which measures whether the student has learned concepts (e.g., biggest, nearest, several) needed to understand oral communication (Gronlund, 1985a).

Individual Achievement Tests. Individual achievement tests are used to identify students who may have learning disabilities. Individual achievement tests are more likely to determine the cause of a student's difficulty because the examiner can observer the student's attention, motivation, understanding of directions, and other factors that cannot be identified on a group achievement test. In the assessment of the learning problem discussed in Chapter 4, the Basic Achievement Skills Individual Screener (BASIS), an individual achievement test, was used by the special education teacher to understand the nature of Billy Field's problem. Other widely used individual achievement tests include the following:

- Peabody Individual Achievement Test (grades 1-12)
- Key Math Diagnostic Arithmetic Test (K-adult)
- Woodcock Reading Mastery Tests (K-grade 12)

SPECIAL ACHIEVEMENT TESTING PROGRAMS

A number of large-scale testing programs have important implications for judging both student and teacher competence. As you read about these testing programs, identify your position as to the usefulness of the programs in terms of improving the quality of instruction in schools.

National Assessment of Educational Progress

The National Assessment of Educational Progress is a nationwide testing program designed to report to the public and educational policymakers the educational achievement of children and young adults in the United States. This testing program involves approximately 1,500 schools representing the diversity of the school population. Student names are not identified on the tests, and no reports are available for school districts or for states. The purpose of the program is not to measure individuals or schools but to report the level of achievement attained by various groups and the nature of changes in achievement over time.

Student achievement is monitored in the following areas: reading and literature, writing, mathematics, science, social studies and citizenship, art, music, and career and occupational development. Each area is assessed every 4 to 8 years through representative sampling of students at ages, 9, 13, 17, and of young adults from 26 to 35 years of age. The results of the testing are reported separately by test item. National results are reported for each age group by religion, sex, race, size and type of community, and level of parent education.

Data from this testing program are useful in determining such questions as: What is the current level of academic achievement? What percentage of students can perform certain arithmetic computations? Does the assessment indicate that student performance is increasing or decreasing compared to previous assessments in the area(s)? How do rural and urban youth perform in relation t the rest of the nation? What are the differences among various groups over time? For example, have

females increased their mathematical ability since the previous assessment?

Minimum Competency Testing for Students

Public schools have been criticized because of declining test scores and because students are often passed each year to the next grade without regard to their level of academic achievement. Such a promotion policy results in many students graduating from high school unable to read and deficient in the basic skills necessary for survival in society.

To rectify this problem, many states have passes laws to mandate the establishment of **minimum competency testing** for elementary and secondary school students. The minimum standards are determined by local school districts, or they are established on a statewide basis. The standards are measured by tests given at various grade levels, and the tests must be passed by the time the student expects to graduate. Early testing allows those students who fail one or more parts of the tests to receive remedial instruction designed to help them pass the tests before completion of the twelfth grade. A student who continues to fail the tests is usually given a "certificate of attendance" rather than a diploma.

Each state has its own regulations and procedures to determine how the testing program should operate. Some states focus primarily on basic skills in reading, writing, and arithmetic. Other states incorporate a combination of basic academic skills with "survival skills" such as consumer knowledge, oral communication, and governmental processes. In addition, the setting of standards, the grade levels assessed, and the types of testing instrument used vary widely among the states. You should find out if your state has a minimum competency testing program and how it operates.

Included in the many issues concerning minimum competency testing programs that need to be resolved are the definition of competencies, the specification of minimal competency, and the testing of minimum competence (Hake & Madams, 1978). Not everyone agrees on the identification of "life skills," "essential skills", or "survival skills" that should be required in education. Are the "essential skills" for a baker, a lawyer, and a salesperson the same? Furthermore, you learned in Chapter 2 that school grades do not correlate highly with economic success (Jencks, 1972); perhaps skills taught in preparation for passing a test will be of little value later on. Nor does everyone agree on appropriate cutoff levels for minimum competency. Hake and Madams (1978) state: "In practice, the setting of minimum scores seems to be the result of compromise between judgments of what minimums seem plausible to expect and judgments about what proportions of failure seem politically tolerable" (p. 468. Larkins (1981) more critically states: "Improving the percentage of high school seniors who can pass a sixth-grade standardized reading test is not likely to halt the flight of parents who seek quality education in the form of a strong college preparatory program" (p. 136).

Testing for minimal competency raises many additional issues: Are the tests measuring what students have been taught? This is an especially important concern in states requiring one test for all students in the state. In Florida, a court order prohibited the use of its mandatory literacy test (which is what they call it) before it was t go into effect in 1979. The test was challenged by a suit brought by a group of parents whose children failed the test. In Debra P. v. Turlington, a federal judge ruled that there must be a sufficient amount of time before the test is announced and the date when diplomas are actually denied in order to allow students the opportunity to take remedial courses. An appeals court upheld the judge and issued an additional requirement that the state must show that the test is related to what is taught in the school curriculum (content validity). In 1983, the court was satisfied that all conditions were met by Florida and allowed students who did not pass the test to be denied diplomas (Brown, 1983).

Other issues that are still being resolved by the states are: What type of remedial instruction is most appropriate for students failing the tests? How can fairness to minority group members and handicapped students be ensured? How do we know what the testing program itself is contributing to student learning?

Not all educators believe that minimum competency tests are the panacea for improving school achievement. Larkins (1981) raises a number of concerns about minimum competency programs. First, the standards often are set very low. When improvement is noted, it is difficult to determine whether it has been caused by increased knowledge or instructors teaching to the test. Second, the testing programs do not address the problems with education. Larkins believes that the focus on rote learning and the overemphasis on ditto sheets to keep students busy at their desks are problems that are not being addressed. Third, there is danger that teachers will emphasize basic skills at the expense of the higher levels of learning because the higher levels are not emphasized on tests. In fact, Madams (1981) presents evidence that the decline in Scholastic Aptitude Test (SAT) scores reflects a decline in higher-level cognitive skills, not in basic skills. Fourth, the minimum competency tests will drive from school students who need to be there. If students are not promoted and are placed in grades where the disparity in ages between children becomes too great, the students who are not promoted will drop out of school rather than remain for continual remediation.

Although it is still too early to determine the merits of minimum competency programs, we have seen a movement by the states to increase the number of days in the school year as well as the number of credits required for graduation. This program represents an important development in education. It will be interesting to follow its progress to determine the extent to which issues are resolved and competency testing becomes widely accepted by legislators, parents, teachers, and students.

Minimum Competency Testing for Teachers

While the debate over minimum competency testing for students was occurring, attention moved to the ability of teachers

to provide the quality of instruction needed to improve student achievement. Teacher selection and training have come under sever criticism that has focused on the fact that many individuals entering the teaching profession have low or marginal academic ability as measured by such tests as the Graduate Record Examination (GRE). Standards for entering and graduating from accredited teacher education programs are too low, and state certification and school selection processes are inadequate. As a result, many teachers have entered the profession having neither gained nor demonstrated teaching competency and knowledge of basic academic skills (Hathaway, 1980).

To combat this problem, many states have passed laws requiring competency or literacy tests for prospective teachers in such areas as reading, writing, language, and mathematics as a requirement for certification. Some states also assess actual teaching performance as well. The argument often made by proponents of teacher competency tests is that if attorneys, real estate agents, and barbers have to take tests to obtain a license, why should teachers not be required to do the same? Most important, proponents believe that the testing will lead to the improvement of education as a profession.

Although the testing programs have focused on new teachers entering the profession, Arkansas and Texas went a step further and required experienced teachers to take tests for recertification. These teachers were given other opportunities to pass the test if they failed it the first time. As you might expect, testing new teachers has been more widely accepted in the profession than has testing experienced teachers. Teachers in Arkansas and Texas were critical of their new law requiring recertification, arguing that it was demeaning to require experienced teachers to take the tests.

A number of arguments are made against competency testing for teachers. In a strong indictment of teacher competency testing, Cole (1979) states:

> Minimum competency testing is a hollow means of judging the efficacy of teachers. It can only whittle away at the edges of the problem; it has no power to cure, because it treats symptoms rather than causes. The heart of the problem perceived by the public lies deep within the structure of the education system...competency testing is nothing more than search for victims, off on a false scent. The time for assurances of competence is at the beginning of the [teacher] educative process, not simply as a belated quality check at the end. (p. 23)

In addition to the argument that the test will not necessarily improve the quality of teaching, other criticisms of testing teachers generally fall into two major areas: The tests are not valid, and the standards are not high enough. The validity argument stems from the fact that the tests generally measure not classroom teaching ability but basic academic skills, which don't predict on-the-job success. In the few states that include some measure of actual teaching ability, there are also questions whether these

assessments can predict teacher effectiveness. If they cannot, the tests will do little to "weed out" inferior teachers. Finally, a good test performance on basic skills does not assure that a teacher is proficient in teaching advanced mathematics or English classes.

The proponents of competency testing for teachers reply to the foregoing arguments by saying that it is not the whole answer to improving the quality of education but that it can be part of the solution. The rest of the solution involves attracting more competent individuals to the profession, screening prospective teachers better, providing more effective teacher education, and instituting effective staff development programs in schools (Hathaway, 1980).

What is your opinion of the competency testing movement in education? Should both teachers and students be involved? Will these tests lead to improved education? Are there alternatives?

PREPARING STUDENTS FOR TESTS

One of the major controversies in the area of testing is the effect on test scores of coaching in test-taking skills, or test-wiseness training. **Test wiseness** is defined (Millman, Bishop, & Ebel, 1965) as "a set of cognitive skills that one may employ on a myriad of tests regardless of the nature of the tests or subject content" (p. 707). Test-wiseness training teaches students general test-taking skills regarding time management, strategies to avoid errors, and guessing strategies--information that is independent of any particular content preparation.

The term **coaching** as it relates to test preparation refers to training to increase scores on specific tests. Classes that prepare students for the Scholastic Aptitude Tests (SAT) are involved in coaching to improve tests scores. Although test-wiseness training and coaching have a different focus, there is some overlap of their goals. For example, certain general test-taking skills are taught in preparation for specific tests like the SAT. In addition, specific content areas can be introduced in test-wiseness training programs.

Although there is evidence that certain types of coaching can increase test scores (see Messick & Jungeblut, 1981; Messick, 1982), a number of questions remain. How much coaching? What kinds of coaching? How much improvement? Because programs vary greatly, it is important to identify the particular programs evaluated before reaching any conclusions about their effectiveness. Researchers have learned that the greatest change in test scores has been found with more intensive training programs in terms of the number of sessions and the degree to which the programs focus on broad cognitve skills (Bangert-Drowns, Kulik, & Kulik, 1983; Messick & Jungeblut, 1981). Also, certain types of test are more responsive to coaching than are others. For example, coaching is more likely to raise test scores on the mathematics section of the SAT than on the verbal section. Finally, is the improvement worth the effort, time, and money? Students do benefit slightly from coaching, particularly if they are not familiar with the testing procedures. However, there are few students with

mediocre ability who, as a result of their coaching experience alone, score high enough to get accepted into a program of study.

The results of investigations into training in test-taking skills, or test-wiseness, are similar: The training can produce a small but significant effect on academic achievement, and the longer the program, the greater the effect. Programs lasting from 5 to 7 weeks yielded greater gains (Samson, 1985; Sarnacki, 1979).

CRITICISMS OF STANDARDIZED TESTING

Not all educators are pleased with the extensive use of standardized testing in our nation's schools. Issues of validity and use are most frequently raised. Following is a brief discussion of the major criticisms that are independent of test validity (Holmen & Docter, 1974; Stenio, 1981), followed by responses by advocates of the proper use of standardized testing.

The Gatekeeper Function

Criticisms. One of the major criticisms of tests is that they are designed to measure differences among individuals to determine who receives and who is denied certain rewards and privileges. For example, students are placed in classes for the gifted, gain admission to college, or are admitted to advanced graduate study on the basis of their test scores. The important question is whether tests are the kind of gatekeepers we want in society. Do educators and employers rely too much on test, which can be unreliable predictors of future performance? Are there alternatives? Should we rely more on individuals' actual performance rather than on their test performance?

Response. Tests also open the gates for some students whose academic records and/or socioeconomic status would not permit them to attend college or enter a special program. For example, a high SAT score may be the deciding factor in admitting a student into a particular college or helping a student to receive financial aid. In other situations, tests may identify special ability or talents that encourage students to enter fields they had never before thought abut.

Harmful Effects on Cognitive Styles

Criticisms. Some educators believe that the widespread use of single-answer test item influences students' style of thinking. More specifically, the concern is that many students may believe that all issues or questions can be resolved by finding the one right answer.

Response. Well-constructed tests can measure higher-level cognitve skills and can enhance students' thinking.

Effect on Curricula and Change

Criticisms. Some educators argue that when teachers learn how and what their students will be tested on, they are less likely to cover important material because they do not believe it will be on the test. In addition, they may be less likely to try new methods of teaching or new resource materials if they believe that students' test results may be affected adversely.

Response. By knowing the general content of a standardized test given at the end of a course, a teacher is likely to cover more instructional objectives and help students to attain a higher level of mastery of content than they would if no standardized test were given.

Students' Self-Concept and Level of Aspiration

Criticisms. As you have learned from reading an earlier chapter, students make social comparisons to judge their own adequacy and self-worth. Students may come to believe early in their educational careers that they are less capable than their classmates and stop trying to achieve. Many parents have been misinformed by guidance counselors who have told them that their child or adolescent was not "college material"--information that may become a self-fulfilling prophecy.

Response. Although some students' self-concepts are negatively influenced by standardized testing, many students are positively influenced by learning that their academic efforts have paid off. With better training, teachers and counselors are less likely to provide inaccurate information about test results to parents and students.

Selection of Homogeneous Educational Groups

Criticisms. Many schools use test results to assign students to classes based on estimates of learning ability. It has been argued that many students have not learned the necessary test-taking skills, and as a result their ability is underestimated by the tests. The fact that students are assigned to different levels, or tracks, in school often influences the type of education they receive in terms of course content, quality of teachers, and expectations for achievement. Finally, students who score low on standardized tests are likely to be placed with other low-scoring students in homogeneous classes, as compared to heterogeneous classes where there is a mix of ability levels. There is some indication that homogeneous grouping does more harm than good for low achievers. Brookover and his colleagues (1982) have reported that more effective schools use fewer forms of stratification of students (e.g., tracking by type of educational program and ability grouping) than do less effective schools. Why do you think the amount of stratification in a school would be related to school achievement?

Could stratification be related too expectations for student performance? Do teachers treat all ability levels the same? What have you learned in previous chapters that can help you to understand this finding?

Response. Decisions about class or group placement of students should not be made on the basis of any single measurement. Course grades, test scores, work habits, and teacher comments are some of the major sources of information that should be used to make educational decisions. In some situations, standardized tests may indicate that a student has mastered more content than indicated by teacher grades, thus preventing inappropriate placement.

Invasion of Privacy

Criticisms. Some test critics believe that although the school has a right to measure achievement, it should not be giving intelligence, personality, and other nonacademic tests. They also say that test results should not be available to individuals other than the students' parents.

Response. Nonacademic tests are not usually given to students unless recommended by a counselor or school psychologist after consultation with a parent. Recent privacy laws attempt to reduce the possibility of anyone receiving test information concerning a student without permission.

Key Points

1. Evaluation is used for various purposes--placement (to determine where students should be placed in the proper sequence of instruction), formative (to determine student performance during instruction), diagnostic (to diagnose learning difficulties during instruction), and summative (to evaluate achievement at the end of instruction).

2. Norm-referenced measurement interprets student performance relative to a group.

3. Criterion-referenced measurement interprets student performance with respect to a specified behavioral criterion of proficiency.

4. Validity and reliability are two important criteria for evaluating the quality of tests.

5. Validity refers to the appropriateness of the interpretation of the test scores with regard to a particular use. A test may be valid for one purpose but not for another. Several approaches are taken to establish validity--content, criterion-related, and construct.

Validity is a matter of degree and should not be considered as an all-or-none characteristic.

6. Reliability refers to the consistency of test scores or other evaluation results from one measure to another. Because there are different types of consistency, different indicators of reliability are used (e.g., test-retest and equivalent forms).

7. Standardized tests are commercially prepared by measurement experts, and they measure behavior under uniform procedures.

8. Achievement tests measure a student's knowledge in a particular academic area at some point in time. Diagnostic and readiness tests are special types of achievement test.

9. Aptitude tests predict the student's probability of success in various education programs.

10. Intelligence tests are common aptitude tests used in school.

11. The Stanford-Binet and Wechsler are two individually administered intelligence tests.

12. Scholastic aptitude tests are group tests for measuring academic potential.

13. Teachers must not be hasty in interpreting intelligence scores of students who have not had an opportunity to learn tasks included in the test, are unmotivated, have poor reading skills, or are not well adjusted emotionally.

14. Information on standardized tests can be located in Bourses Tests in Print and Mental Measurements Yearbooks.

15. Minimum competency testing is a procedure in which students and teachers must demonstrate mastery of specific skills.

16. Training in test-taking skills can help students to perform more effectively in testing situations.

17. Many educators criticize standardized tests. These criticisms pertain to both the validity and use of the tests.

MEASUREMENT AND EVALUATION *

All teaching involves *evaluation.* At the heart of evaluation is judgment-making decisions based on values. In the process of evaluation, we compare information to criteria and then make judgments. Teachers must make all kinds of judgments. "Should we use a different text this year?" "Is the film appropriate for my students?" "Will Sarah do better if she repeats the first grade?" "Should Terry get a B- or a C+ on the project?"

Measurement is evaluation put in quantitative terms-the description of an event or characteristic in numbers. Measurement tells how much, how often, or how well by providing scores, ranks, or ratings. Instead of saying "Sarah doesn't seem understand addition," a teacher might say "Sarah answered only 2 of the 15 problems correctly on her addition work sheet." Measurement also allows a teacher to compare one student's performance on one particular task with a standard or with the performances of the other students.

Not all the evaluative decisions made by teachers involve measurement. Some decisions are based on information that is difficult to express numerically: student preferences, information from parents, previous experiences, even intuition. But measurement does play a large role in many classroom decisions, and properly done, it can provide unbiased data for evaluations.

The answers given on any type of test have no meaning by themselves; basic types of comparison are possible. A test score can be compared to the scores obtained by other people who have taken the same test, If you took a college entrance exam, the score you received told you (and the admissions offices of college) how your performance compared to performances of many other people who had previously taken the same test or one like it. The second type of comparison is to a fixed standard or minimum passing score. Most tests required for a driver's license are based on this kind of comparison.

Norm-Referenced Tests

In *norm-referenced testing,* the other people who have taken the test provide the norms for determining the meaning of a given individual's score. You can think of a norm as being the typical level of performance for a particular group. By comparing the individual's raw score (the actual number correct) to the norm, we can determine if the score is above, below, or around the average for that group. There are at least three types of norm groups

---

* Note. From Educational Psychology 3/e (pp. 488-521) by Anita E. Woolfolk, 1987, Englewood Cliffs, N.J.: Prentice Hall. Copyright 1987 by Allyn & Bacon. Reprinted by permission.

(comparison groups) in education. One frequently used norm group is the class or school itself. When a teacher compares the score of one student in a tenth-grade American history class with the scores of all the other students in the class, the class itself is the norm group. If the teacher happens to have three American history classes, all of about the same ability, then the norm group for evaluating individual performance might be all three classes.

Norm groups may also be drawn from wider areas. Sometimes, for example, school districts develop achievement tests. When students take this kind of test, their scores are compared to the scores of all the other students at their grade level throughout the district. A student whose score on the achievement test was in the top 25 percent at a particularly good school might be in the top 15 percent for the entire district. Finally, some tests have national norm groups. When students take the college entrance exam, their scores are compared with the scores of students all over the country.

Norm-referenced tests are constructed with certain objectives in mind, but the test items themselves tend to cover many different abilities rather than assess a limited number of specific objectives. Norm-referenced tests are especially useful in measuring overall achievement when students have come to understand complex material by different routes. Norm-referenced tests are also appropriate when only the top few candidates can be admitted to a program.

Hopkins and Antes (1979) have listed several limitations of norm-referenced measurements. Results of a norm-referenced test do not tell you whether students are ready to move on to more advanced material. Knowing that a student is in the top 3 percent of the class on a test of algebraic concepts will not tell you if he or she is ready to move on to trigonometry. Everyone in the class might have failed to achieve sufficient mastery of algebraic concepts.

Norm-referenced tests are also not particularly appropriate for measuring affective and psychomotor objectives. To measure psychomotor learning, a clear description of standards is necessary to judge individuals. Even the best gymnast in any school performs certain exercises better than others and needs specific guidance about how to improve. In the affective area, attitudes and values are personal; comparisons among individuals are not really appropriate. For example, what is an "average" performance on a measure of political values or opinions? Finally, norm-referenced tests tend to encourage competition and comparison of scores. Some students compete to be the best. Others, realizing that being the best is impossible, may compete to be the worst! Either goal has its casualties.

Criterion-Referenced Tests

When test scores are compared not to those of others but to a given criterion or standard of performance, the test is called *criterion-referenced.* In deciding who should be allowed to drive a car, it is important to determine just what standard of performance is appropriate for selecting safe drivers. It does not matter how

your test results compare to the results of others. If your performance on the test was in the top 10 percent but you consistently ran through red lights, you would not be a good candidate for receiving a license, even though your score was high.

Criterion-referenced tests measure the mastery of very specific objectives. The results of a criterion-referenced test should tell the teacher exactly what the students can do and what they cannot do, at least under certain conditions. For example, a criterion-referenced test would be useful in measuring the ability to add three-digit numbers. A test could be designed with 20 different problems. The standard for mastery could be set at 17 out of 20 correct. (The standard is often somewhat arbitrary but may be based on such things as the teacher's experience with other classes or the difficulty of the problems.) If two students receive scores of 7 and 11, it does not matter that one student did better than the other, since neither met the standard of 17. Both need more help with addition.

There are many such instances in the teaching of basic skills when comparison to a preset standard is more important than comparison to the performance of others. It is not very comforting to know, as a parent, that your child is better than most of the students in class in reading if all the students are unable to read material suited for their grade level. Sometimes standards for meeting the criterion must be set at 100 percent correct. You would not like to have your appendix removed by a surgeon who left surgical instruments inside the body only 10 percent of the time.

But criterion-reference tests are not appropriate for every situation. Not every subject can be broken down in to a set of specific objectives that exhausts all possible learning outcomes. And as we noted in Chapter 11, often the true objective is understanding, appreciating, or analyzing. Moreover, although standards are important in criterion-referenced test, they often tend to be arbitrary, as you have already seen. When deciding whether a student has mastered the addition of three-digit numbers comes down to the difference between 16 or 17 correct answers, t seems hard to justify one particular standard over another. Finally, at times it is valuable to know how the students in your class compare to other students at their grade level both locally and nationally. Remember, too, that admission to college is based in part on a test given t students all over the country. You will want your students who plan to go to college to have a good chance to do well on such a test. Table 14-1 offers a comparison of norm-referenced and criterion-referenced tests.

WHAT DO TEST SCORES MEAN?

On the average, more than 1 million standardized tests are given each school day in classes throughout this country (Lyman, 1978). Most of these are norm-referenced standardized tests.

Basic Concepts

Standardized tests are the official-looking pamphlets and piles of forms purchased by school systems and administered to students. More specifically, a standardized test is "a task or set of tasks given under standard conditions and designed to assess some aspect of a person's knowledge, skill, or personality....A test yields one or more objectively obtained quantitative scores, so that, as nearly as possible, each person is assessed in the same way" (Green, 1981, p. 1001). The tests are meant to be given under carefully controlled conditions, so that students all over the country undergo the same experience when they take the tests. Standard methods of developing items, administering the test, scoring it, and reporting the scores are all implied by the standardized test.

The test items and instructions have also been tried out to make sure they work and then rewritten and retested as necessary. The final version of the test is administered to a *norming sample*, a large sample of subjects as similar as possible to the students who will be taking the test in school systems throughout the country. This norming sample serves as a comparison group for all students who later take the test.

The test publishers provide one or more ways of comparing each student's raw score (number of correct answers) with the norming sample. Let's look at some of the measurements on which comparisons and interpretations are based.

Frequency Distributions.

A *frequency distribution* is simply a listing of the number of people who obtain each score or fall into each range of score on a test or other measuring device. For example, on a spelling test 19 students made these scores: 100, 95, 85, 85, 85, 80, 75, 75, 75, 70, 65, 60, 60, 55, 50, 50, 45, 40. As you can see, 1 student made a score of 100, 3 made 85, and so on. This kind of information is often expressed as a simple graph where one axis (the x or horizontal axis) indicates the possible scores and the other axis (the y or vertical axis) indicates the number of subjects who attained each score. A graph, in this case a histogram, or bar graph, of the spelling test scores is shown in Figure 14-1.

Measurements of Central Tendency and Standard Deviation.

You have probably had a great deal of experience with means. A *mean* is simply the arithmetical average of a group of scores. To calculate the mean, you add the scores and divide the total by the number of scores in the distribution. For example, the total of the 19 spelling scores is 1,340, so the mean is 1,340/19 or 70.53. The mean offers one way of measuring *central tendency*, the score that is typical or representative of the whole distribution. Two

other measures of central tendency are the median and the mode. The *median* is the middle score in the distribution, the point at which half the scores are larger and half are smaller. The median of the 19 scores is 75. Nine scores in the distribution are greater than or equal to 75, and nine are less. The *mode* is the score that occurs most often. The distribution in Figure 14-1 actually has two modes, 75 and 85. This makes it a bimodal distribution.

The measure of central tendency gives a score that is representative of the group of scores, but it does not tell you anything about how they are distributed. Two groups of scores may both have a mean of 50 but be alike in no other way. One group might contain the scores 50, 45, 55, 55, 45, 50, 50; the other group might contain the scores 100, 0, 50, 90, 10, 50, 50. In both cases the mean, median, and mode are all 50, but the distributions are quite different.

The *standard deviation* is a measure of how the scores spread out around the mean. The larger the standard deviation, the more spread out around the mean. The smaller the standard deviation, the more the scores are clustered around the mean. For example, in the distribution 50, 45, 55, 55, 45, 50, 50, the standard deviation is much smaller than in the distribution 100, 0, 50, 90,10, 50, 50. Another way of saying this is that distributions with very small standard deviations have less variability in the scores. The standard deviation is relatively easy to calculate if you remember you high school math. It does take time, however. The process is similar to taking an average, but square roots are used. To calculate the standard deviation, you follow these steps:

1. Calculate the mean (written as $\bar{X}$) of the scores.
2. Subtract the mean from each of the scores. This is written as $(X - \bar{X})$.
3. Square each difference (multiply each difference by itself). This is written $(X - \bar{X})^2$.
4. Add all the squared differences. This is written $\Sigma(X - \bar{X})^2$.
5. Divide this total by the number of scores. This is written $\frac{\Sigma(X-\bar{X})^2}{N}$.
6. Find the square root. This is written $\sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$ which is the formula for calculating the standard deviation.

The *Student Guide* that accompanies this text provides a more complete explanation of the standard deviation and a shortcut method for estimating it with classroom test data. Knowing the mean and standard deviation of a group of scores gives you a better picture of the meaning of an individual score. For example, suppose you received a score of 78 on a test. You would be very pleased with the score if the mean of the test were 70 and the standard deviation were 4. In this case, your score would be 2 standard deviations above the mean, a score well above the average. Consider the difference if the mean of the test had remained at 70 but the standard deviation had been 20. In the

second case, your score of 78 would be less than 1 standard deviation from the mean. You would be much closer to the middle of the group, with an above-average but not a high score. Knowing the standard deviation tells you much more than simply knowing the range of scores. One or two students may do very well or very poorly no matter how the majority scored on the tests.

The Normal Distribution.

Standard deviations are very useful in understanding test results. They are especially helpful if the results of the test form a normal distribution, like those in the two example tests in Figure 14-2. You may have met the normal distribution before. It is the bell-shaped curve, the most famous frequency distribution because it describes many naturally occurring physical and social phenomena. Many scores fall in the middle, giving the curve its puffed appearance. You find fewer and fewer scores as you look out toward the end points, or tails, of the distribution. The normal distribution has been thoroughly analyzed by statisticians. The mean of a normal distribution is also its midpoint. Half the scores are above the mean, and half are below it. In a normal distribution, the mean, median, and mode are all the same point.

Another convenient property of the normal distribution is that the percentage of scores falling within each area of the curve is known, as you can see in Figure 14-3. A person scoring within 1 standard deviation of the mean obviously has a lot of company. Many scores pile up here. In fact, 68 percent of all scores are located in the area plus and minus 1 standard deviation from the score. About 16 percent of the scores are higher than 1 standard deviation above the mean. Of this higher group, only 2.5 percent are better than 2 standard deviations above the mean. Similarly, only about 16 percent of the scores are less than 1 standard deviation below the mean, and of that group only about 2.5 percent are worse than 2 standard deviations below. At 2 standard deviations from the mean in either direction, the scorer has left the pack behind.

The SAT college entrance exam offers one example of a normal distribution. The mean of the SAT is 500 and the standard deviation is 100. If you know people who made scores of 700, you know they did very well. Only about 2.5 percent of the people who take the test do that well because only 2.5 percent of the scores are better than 2 standard deviations above the mean in a normal distribution.

Types of Scores

Now you have enough background for a discussion of the different kinds of scores you may encounter in reports of results from standardized tests.

Percentile Rank Scores.

The concept of ranking is the basis for one very useful kind of score reported on standardized tests: a percentile rank score. In percentile ranking, each student's raw score is compared with the raw scores obtained by the students in the norming sample. The percentile rank shows the percentage of students in the norming sample who scored at or below a particular raw score. If a student's score is the same or better than three-quarters of the students in the norming sample, the student would score in the 75th percentile, or have a percentile rank of 75. You can see that this does not mean that the student had a raw score of 75 correct answers or even that the student answered 75 percent of the questions correctly. Rather, the 75 refers to the percentage of people in the norming sample whole scores on the test were equal to or below this student's score. A percentile rank of 50 means that a student has scored as well or better than 50 percent of the norming sample and achieved an average score.

Figure 14-4 illustrates one problem in interpreting percentile scores. Differences in percentile ranks do not mean the same thing in terms of raw score points in the middle of the scale as they do at the fringes. The graph shows Joan's and Alice's percentile scores on the fictitious Test of Excellence in Language and Arithmetic. Both students are about average in arithmetic skills. One equaled or surpassed 50 percent of the norming sample; the other, 60 percent. But in the middle of the distribution, this difference in percentile ranks means a raw score difference of only a few points. Their raw scores actually were 75 and 77. In the language test, the difference in percentile ranks seems to be about the same as the difference in arithmetic since one ranked at the 90th percentile and the other at the 99th. But the difference in their raw scores on the language test is much greater. It takes a greater difference in raw score points to make a difference in percentile rank at the extreme ends of the scale. On the language test the differences in raw scores is about 10 points.

Grade-Equivalent Scores.

Grade-equivalent scores are generally obtained from separate norming samples for each grade level. The average of the scores of all the tenth graders in the norming sample defines the tenth-grade equivalent score. Suppose the raw-score average of the tenth-grade norming sample was 38. Any students who attains a raw score of 38 on the test will be assigned a grade-equivalent score of tenth grade. Grade-equivalent scores are generally listed in numbers, such as 8.3, 4.5, 7.6, 11.5, and so on. The whole number gives the grade and the decimals stand for tenths of a year, but they are usually interpreted as months.

Suppose a student with the grade-equivalent score of 10 is a seventh grader. Should this student be promoted immediately? No! Different forms of the test are used at different grade levels, so the seventh grader may not have had to answer items that would be given to tenth graders. The high score may represent superior mastery of material at the seventh-grade level, rather than a

capacity for doing advanced work. Even if an average tenth grader did as well as our seventh grader on this particular test, the tenth grader would certainly know much more than this test covered.

Because grade-equivalent scores are misleading and so often misinterpreted, especially by parents, most educators and psychologists strongly believe they should not be used at all. There are several other forms of reporting available that are more appropriate.

Standard scores.

As you may remember, one problem with percentile ranks is the difficulty in making comparisons among ranks. A discrepancy of a certain number of raw-score points has a different meaning at different places on the scale. With standard scores, on the other hand, a difference of 10 points is the same everywhere on the scale.

Standard scores are based on the standard deviation. A very common standard score is called the *z score*. A z score tells how many standard deviations above or below the average a raw score is. In the example described earlier in which you were fortunate to get a 78 on a test where the mean was 70 and the standard deviation was 4, your z score would be +2, or 2 standard deviations above the mean. If a person were to score 64 on this test, the score would be -1.5 standard deviation units below the mean, and the z score would be -1.5. A z score of 0 would be no standard deviations above the mean--in other words, right on the mean.

To calculate the z score for a given raw score, just subtract the mean from the raw score and divide the difference by the standard deviation. The formula is:

Since it is often inconvenient to use negative numbers, other standard scores have been devised to eliminate these difficulties. The T score has a mean of 50 and uses a standard deviation of 10. If you multiply the z score by 10 (which eliminates the decimal) and add 50 (which gets rid of the negative number), you get the equivalent T score as the answer. The person whose z score was -1.5 would have a T score of 35:

$$-1.5 \times 10 = -15$$
$$-15 + 50 = 35$$

The scoring of the College Entrance Examination Board test is based on a similar procedure. The mean of the scores is set at 500, and a standard deviation of 100 is used.

Before we leave this section on types of scores, we should mention one other widely used method. *Stanine scores* (the name comes from "standard nine") are standard scores. There are only nine possible scores on the stanine scale, the whole numbers 1 through 9. The mean is 5, and the standard deviation is 2. Each unit from 2 to 8 is equal to half a standard deviation. Stanine scores also provide a method of considering a student's rank, because each of the nine scores includes a specific range of percentile scores in the normal distribution. For example, a stanine score of 1 is assigned to the bottom 4 percent of scores in a

distribution. A stanine of 2 is assigned to the next 7 percent. Of course, some raw scores in this 7 percent range are better than others, but they all get a stanine score of 2.

Each stanine score represents a wide range of raw scores. This has the advantage of encouraging teachers and parents to view a student's score in more general terms instead of making fine distinctions based on a few points. Figure 14-5 compares the four types of standard scores we have considered, showing how each would fall on a normal distribution curve.

Interpreting Test Scores

One of the most common problems with the use of test is misinterpretation of scores. Often this takes the form of believing that the numbers are precise measurements of a student's ability. No test provides a perfect picture of a person's abilities; a test is only one small sample of behavior. You probably have had the experience of feeling that you really understood a subject only to have the test ask several questions you were not expecting or felt were simply unfair. Was the test an accurate measure of your ability? Two factors are important in developing good tests: reliability and validity. Both must be considered in interpreting test scores.

Reliability.

If you took a standardized test on Monday, then took the same test again one week later and received about the same score each time, you would have reason to believe the test was reliable. If 100 people took the test one day and then repeated it again the following week and the ranking of the individual score was about the same for both tests, you would be even more certain the test was reliable. (Of course, this assumes that no one looks up answers or studies before the second test.) A reliable test gives a consistent and stable "reading" of a person's ability from one occasion to the next, assuming the person's ability remains the same. A reliable thermometer works in a similar manner, giving you a reading of 100C each time you measure the temperature of boiling water. Measuring a test's *reliability* in this way gives an indication of test-retest reliability. If a group of people takes two equivalent versions of a test and the scores on both tests are comparable, this indicates alternate-form reliability.

Reliability can also refer to the internal consistency or the precision of a test. This type of reliability, known as split-half reliability, is calculated by comparing performance on half of the test questions with performance on the other half. If someone, for example, did quite well on all the odd-numbered items and not at all well on the even-numbered items, we could assume that the items were not very consistent or precise in measuring what they were intended to measure (Cronbach, 1970).

True Score.

All tests are imperfect estimators of the qualities or skills they are trying to measure. There is error involved in every testing situation. Sometimes the errors are in your favor, and you may score higher than your ability might warrant. This occurs when you happen to review a key section just before the test or are unusually well rested and alert the day of an unscheduled "pop" quiz. Sometimes the errors go against you. You don't feel well the day of the examination, haven just gotten bad news from home, or focused on the wrong material in your review. But if you could be tested over and over again without becoming tired and without memorizing the answers, the average of the test scores would bring you close to a true score. In other words, a student's true score can be thought of as the mean of all the scores the student would receive if the test were repeated many times.

But in reality students take a test only once. That means that the score each student receives is made up of the hypothetical true score plus some amount of error. How can error be reduced so that the actual score can be brought closer to a true score? As you might guess, this returns us to the question of reliability. The more reliable the test, the less error in the score actually obtained. On standardized tests, test developers take this into consideration and make estimations of how much the students' scores would probably vary if they were tested repeatedly. This estimation is called the *standard error of measurement.* It represents the standard deviation of the distribution of scores from our hypothetical repeated testings. Thus, a reliable test can also be defined as a test with a small standard error of measurement.

The most effective way to improve reliability is to add more items to the test. Generally speaking, longer tests are more reliable than shorter tests. In their interpretation of tests, teachers must also take the margin for error into consideration.

Confidence Interval.

Teachers should never base an opinion of a student's ability or achievement on the exact score the student obtains. Many test companies now report scores using a *confidence interval* or "standard error band" that encloses the student's actual score. This makes use of the standard error of measurement and allows a teacher to consider the range of scores within which a student's true score might be.

Let us assume, for example, that two students in your class take a standardized achievement test in Spanish. The standard error of measurement for this test is 5. One student receives a score of 79; the other, a score of 85. At first glance, these scores seem quite different. But when you consider the standard error bands around the scores instead of the scores alone, you see that the bands overlap. The first student's true score might be anywhere between 74 and 84 (that is, the actual score of 79 plus and minus the standard error of 5). The second student's true score might be anywhere between 80 and 90. If these two students took the test again, they might even switch rankings. It is crucial

to keep in mind the idea of standard error bands when selecting students for special programs. No child should be rejected simply because his or her obtained score misses the cutoff by one or two points. The student's true score might well be above the cutoff point.

Validity.

If a test is sufficiently reliable, the next question is whether it is valid. A test has *validity* if it measures what it is supposed to measure or predicts what it is supposed to predict. To be a valid test of Spanish grammar and vocabulary, the questions must measure just those things and not reading speed or lucky guessing. Tests of mathematics achievement ought to measure what students have learned in mathematics and not level of anxiety about math. A test is judged to be valid in relation to a specific purpose.

There are several ways to determine whether or not a test is valid for a specific purpose (Gronlund, 1985). If the purpose of a test is to measure the skills covered in a particular course or unit, the inclusion of questions on all the important topics and on no extraneous topics would provide content-related evidence of validity. Have you ever taken a test that dealt only with a few ideas from one lecture or a few pages of the textbook? That test would certainly show no evidence of content-related validity. Some tests are designed to predict outcomes. The SATs, for example, are intended to predict performance in college. If SAT scores correlate with academic performance in college as measured by, say, grade-point average in the first year, then we have criterion-related evidence of validity for the SAT. In other words, the test scores are fairly accurate predictors of how well the student would do in college. Most standardized tests are designed to measure some psychological characteristic or "construct" such as reasoning ability, reading comprehension, achievement motivation, intelligence, creativity, and so on. It is a bit more difficult to gather construct-related evidence of validity, yet this is a very important requirement. A test might be a good predictor of an outcome but still be an unsatisfactory measure of the construct under consideration. For example, both family income and intelligence test scores predict school achievement. Assuming that intelligence test scores and school achievement ought to be related, what makes the intelligence test a better measure of intelligence than family income? This has to do with our understanding of the construct of intelligence--our conceptual framework. Construct-related evidence of validity is gathered over many years. It is seen in a pattern of scores. For example, older children can answer more questions on intelligence tests than younger children. This fits with our construct of intelligence. If the average 5-year-old answered as many questions correctly on a test as the average 13-year-old, we would doubt that the test really measured intelligence. Construct-related evidence for validity can also be demonstrated when the results of a test correlate with the results of other well-established and valid measures of the same construct.

A number of factors may interfere with the validity of tests given in classroom situations. One problem has already been

mentioned--a poorly planned test with little or no relation to the important topics. Standardized tests must also be chosen so that the items on the test actually measure content covered in the classes. This match is absent more often than we might assume. And students must have the necessary skills to take the test. If students score low on a science test not because they lack knowledge about science but because they have difficulty reading the question, do not understand the directions, or do not have enough time to finish, the test is not a valid measure of science achievement.

A test must be reliable in order to be valid. For example, if an intelligence test yields different results each time it is given to the same child over a few months, then by definition it is not reliable. And it couldn't be a valid measure of intelligence because intelligence is assumed to be fairly stable, at least over a short period of time. However, reliability will not guarantee validity. If that intelligence test gave the same score every time for a particular child but didn't predict school achievement, speed of learning, or other characteristics associated with intelligence, then performance on the test would not be a true indicator of intelligence. The test would be reliable but invalid.

The Guidelines should help you increase the reliability and validity of the standardized tests you give.

TYPES OF STANDARDIZED TESTS

Several kinds of standardized tests are used in schools today. If you have seen cumulative folders, with testing records for individual students over several years, you know how many ways students are tested in school in this country. There are three broad categories of standardized test: achievement, diagnostic, and aptitude (including interest). As a teacher, you will probably encounter achievement and aptitude tests most frequently. An excellent source of information on all types of published tests is a series called the *Mental Measurements Yearbooks*. These yearbooks, once edited by Oscar K. Buros and now done by psychologists at the University of Nebraska, contain reviews of every major test, with information on the strengths and weaknesses of each, appropriate age levels, and how to order.

Achievement Tests: What Has the Student Learned?

The most common standardized test given to students are achievement tests. These are meant to measure how much a student has leaned in specific content areas such as reading comprehension, language usage, grammar, spelling, number operations, computation, science, social studies, mathematics, and logical reasoning.

As a teacher, you will undoubtedly give standardized tests. The results will be meaningless unless you administer them properly, following the procedures exactly as given in the

instructions. If you are well prepared, organized, and relaxed, it will help your students relax and perform better on the test.

Frequently Used Achievement Tests.

Achievement tests can be designed to be administered to a group or individually. Group tests are generally used for screening, to identify children who might need further testing. Results of group tests can also be used as a basis for grouping students according to achievement levels. Individual achievement tests are generally given to determine a child's academic level more precisely or to help diagnose learning problems.

Norm-referenced achievement tests that are commonly given to groups include the California Achievement Test, the Metropolitan Achievement Test, the Stanford Achievement Test, the Comprehensive Test of Basic Skills, the SRA Achievement Series, and the Iowa Test of Basic Skills. Individually administered norm-referenced tests include Part II of the Woodcock-Johnson Psycho-Educational Battery: Tests of Achievement; the Wide-Range Achievement Test; the Peabody Individual Achievement Test; and the Kaufman Assessment Battery for Children. These tests vary in their reliability and validity.

Using Information from a Norm-Referenced Achievement Test.

What kind of specific information do achievement tests results offer teachers? Test publishers usually provide individual profiles for each student, showing scores on each subtest. Figure 14-6 is an example of an individual profile for an eighth grader, Susie Pak, on the California Achievement Test. Note that the Individual Test Record reports the scores in many different ways. On the top of the form, after the identifying information about Susie's teacher, school, district, grade, and so on, is a list of the various tests--Vocabulary, Comprehension, Total Reading (Vocabulary and Comprehension combined), Language Mechanics, and so on. Beside each test are several different ways of reporting her score on that test:

GE: Susie's grade-equivalent score.
AAGE: Anticipated achievement grade-equivalent score, which is the average grade-equivalent score on this test for students around the country who are at Susie's grade level.
DIFF: An indication of whether or not the difference between Susie's actual and anticipated grade-equivalent scores is statistically significant (+ means her actual grade-equivalent score is significantly higher than the average, - means her score is significantly lower than the average).
SS: Susie's standard score.

LP: Susie's local percentile score; this tells us where Susie stands in relation to other students at her grade level in her district.

NP: Susie's national percentile score, telling us where Susie stands in relation to students at her grade level across the country.

RANGE: The range of national percentile scores in which Susie's true score is likely to fall. You may remember from our discussion of true scores that this range, or confidence interval band, is determined by adding and subtracting the standard error of the test from Susie's actual score. There is a 95 percent chance that Susie's true score is within this range.

Beside the scores is a graph showing Susie's national percentile and stanine scores, with the standard error bands indicated around the scores. Bands that show any overlap are probably not significantly different. But when there is no overlap between bands for two test scores, we can be reasonably certain that Susie's achievement in these areas is actually different.

Let's look at Susie' scores more carefully. In language mechanics she has a grade-equivalent score of 12.9, which is equal to a standard score of 763. This is at the 92nd percentile for Susie's district and at the 83rd percentile nationally. Her true national percentile score is probably in the range from 73 to 93 (that is, plus and minus 1 standard error of measurement from the actual score of 83). By looking at the graph, we can see that Susie's language mechanics score is equal to a stanine of 7. We can also see that her score bands on vocabulary and comprehension overlap a bit, so her achievement in these areas is probably similar, even though there seems to be a difference when you look at the NP scores alone. When we compare language mechanics and language expression, on the other hand, we see that the bands do not overlap. Susie probably is stronger in mechanics than in expression.

You may also have noticed that the difference between Susie's actual and anticipated grade-equivalent scores in language mechanics is significant. She scored significantly higher than the average student in her grade on this part of the test. But as discussed earlier, it is best not to interpret grade-equivalent scores literally. Susie is much better than the average eighth grader in language mechanics (in fact, she is as good or better than 92 percent of students at her grade level locally), but it's very unlikely that she could handle twelfth-grade English classes.

The profile in Figure 14-6 tells us a number of things. First, we can see that Susie is apparently strongest in language mechanics and math concepts and applications and weakest in language expression and science. But she is significantly below the average for her grade level only in science. By comparing the two columns under LP (local percentiles) and NP (national percentiles),

we can see that the eighth graders in Susie's district are achieving below the national level on every test except math computations. This is evident because Susie's performance places her generally in the 70th to 90th percentile range for her district but only in the 50th to 70th percentile range nationally. For example, Susie's performance in vocabulary is well above average for her district (86th percentile) but only average (48th percentile) for eighth graders nationally.

The scores we have just described are all norm-referenced. But results from standardized tests like the one Susie took can also be interpreted in a criterion-referenced way. The bottom portion of Susie's Individual Test Record in Figure 14-6 breaks down the larger categories of the top section and shows criterion-referenced scores that indicate mastery, partial knowledge, or nonmastery for specific skills like use of synonyms and antonyms, character analysis in reading comprehension, and abilities in geometry and physics. Teachers could use these results to get a relatively good idea of Susie's strengths and weaknesses with these specific skills and thus to determine her progress toward objectives in a given subject.

Diagnostic Tests: What Are the Student's Strengths and Weaknesses?

If teachers want to identify more general learning problems, they may need to refer to results from the various diagnostic tests that have been developed. Most diagnostic tests are given to students individually by a highly trained professional. The goal is usually to identify the specific problems a student is having. Achievement tests, both standardized and teacher-made, identify weaknesses in academic content areas like mathematics, computation, or reading; individually administered diagnostic tests identify weaknesses in learning processes. There are diagnostic tests to assess the ability to hear differences between sounds, remember spoken words or sentences, recall a sequence of symbols, separate figures from their background, express relationships, coordinate eye and hand movements, describe objects orally, blend sounds together to form words, recognize details in a picture, coordinate movements, and many other abilities needed to receive, process, and express information.

Frequently Used Diagnostic Tests.

Some diagnostic tests measure a student's ability in a variety of areas. These tests include the Detroit Test of Learning Aptitude and Part I of the Woodcock-Johnson Psycho-Educational Battery: Tests of Cognitive Ability. Others, however, assess a student's ability in a more specific area. Tests of motor skills include the Bender Gestalt Test and the Purdue Perceptual Motor Survey.

For assessing specific areas of perception, commonly used tests include the Wepman Auditory Discrimination Test, the Goldman-Fristoe-Woodcock Test of Auditory Discrimination, the

Frostig Developmental Test of Visual Perception, and the Motor-Free Visual Perception Test.

Elementary school teachers are more likely than secondary teacher to receive information from diagnostic tests. There are few such tests for older students. If you become a high school teacher, your students are more likely to be given aptitude tests.

Aptitude Tests: How Well Will the Student Do in the Future?

Both achievement and aptitude tests measure developed abilities. Achievement tests may measure abilities developed over a short period of time such as during a week-long unit on map reading, or over a longer period of time, such as a semester. Aptitude tests are meant to measure abilities developed over many years and to predict how well a student will do in learning unfamiliar material in the future. The greatest difference between the two types of tests is that they are used for different purposes--achievement tests to measure final performance (and perhaps give grades), and aptitude tests to predict how well people will do in particular programs like college or professional school (Anastasi, 1981).

Scholastic Aptitude.

The purpose of a scholastic aptitude test, like the SAT or ACT, is to predict how well you would do in college. Colleges use such scores to help decide on acceptances and rejections. The SAT may have seemed like an achievement test to you, measuring what you had already learned in high school. Although the test is designed to avoid drawing too heavily on specific high school curricula, the questions are very similar to achievement test questions.

Standardized aptitude tests such as the SAT (and the SCAT for younger students) seem to be fairly reliable in predicting future achievement. Since standardized tests are less open to teacher bias, they may be even fairer predictor of future achievement than high school grades are. Indeed, some psychologists believe grade inflation in high schools has made tests like the SAT even more important.

IQ and Scholastic Aptitude.

In Chapter 4 we discussed one of the most influential aptitude tests of all, the IQ test. The IQ test as we know it could well be called a test of scholastic aptitude. Figure 14-7 shows how IQ scores are distributed based on the results of the major individual tests. Now that you understand the concept of standard deviation, you will be able to appreciate several statistical characteristics of the tests.

For example, the IQ score is really a standard score with a mean of 100 and a standard deviation of 15 (for the Wechsler Scales, the Cognitive Abilities section of the Woodcock-Johnson Psycho-Educational Battery, and the Global Scale of the Kaufman Assessment Battery for Children) or 16 (for the Stanford-Binet and

the McCarthy Scales for Children). Thus, about 68 percent of the general population would score between +1 and -1 standard deviations from the mean, or between about 85 and 115. Only about 2.5 percent of the general population would have a score higher than 2 standard deviations above the mean--that is, above 130 on the Wechsler Scales.

A difference of a few points between two student's IQ scores should not be viewed as important. Scores between 90 and 109 are within the average range. In fact, scores between 80 and 119 are considered to range from low average to high average. To see the problems that may arise, consider the following conversation:

> Parent: We came to speak with you today because we are shocked at our son's IQ score. We can't believe he has only a 99 IQ when his sister scored much higher on the same test. We know they are about the same. In fact, Sammy has better marks than Lauren did in the fifth grade.
> Teacher: What was Lauren's score?
> Parent: Well, she did much better. She scored a 103.

Clearly brother and sister have both scored within the average range. The standard error of measurement on the WISC-R varies slightly from one age to the next, but the average standard error is 3.19. So the bands around Sammy's and Lauren's IQ scores--about 96 to 102 and 100 to 106--are overlapping. Either child could have scored 100, 101, or 102. The scores are so close that, on a second testing, Sammy might score slightly higher than Lauren.

Vocational Aptitude and Interest.

In schools, the guidance counselor is generally the person most concerned with students' career decisions. It is the responsibility of people in the guidance office to know what aptitude test scores really mean and how to help each student make an appropriate decision. Two kinds of tests, vocational aptitude and vocational interest, may provide useful information for educational planning. But as with any tests, interpretation must be cautious.

If you teach in a junior high or high school, your school may administer vocational aptitude test to the students. One test designed to measure aptitudes relevant to career decisions is the Differential Aptitude Test (DAT). Students in grades 8 through 12 may take the test. Questions cover seven areas: (1) verbal reasoning; (2) numerical ability; (3) abstract reasoning; (4) clerical speed and accuracy; (5) mechanical reasoning; (6) space relations; and (7) spelling and language.

The test results on the DAT are converted into percentiles, and a percentile band is reported for each subtest. After the tests have been scored, the guidance counselors in a school should be able to help students relate their DAT profile scores to career-

planning decisions. In general, people in different occupational groups do tend to have different patterns of scores on the DAT, which gives the test some validity.

In many high schools, vocational interest tests are also given. Three examples are the Kuder Preference Record, the Strong-Campbell Interest Blank, and Part III of the Woodcock-Johnson Psycho-Educational Battery: Tests of Interest Level. In these tests students may be asked to indicate which of several activities (such as collecting books, collecting shells, or collecting postcards) they would like most and which they would like least. The pattern of the students' answers is then compared to the answer patterns of adults working in different occupations. It must be remembered, however, that the results on such a test indicate interest, not aptitude or talent.

Occupational interest tests cannot tell you exactly what students will be or should be when they grow up. Results of such tests should be interpreted in the light of all the other information available about a student, as well as with some healthy skepticism. As a teacher, information you gain about a student's interests from test results can be used most appropriately to help motivate the student. No career option should be permanently closed to an adolescent on the basis of an occupational interest test.

BASIC CONCEPTS IN MEASUREMENT AND EVALUATION *

Overview

In this chapter we examine some fundamental ideas about measurement, testing, reliability, validity, and evaluation. Teachers do more than interact with students. They have to plan. And some of that planning involves the evaluation of student achievement after instruction has taken place.

If tests are going to be used in evaluating achievement, their form and content should, for the most part, be known before instruction begins so that the test items can guide and focus instruction. You may have heard that it is wrong to teach to the test. We don't agree. Test items should directly reflect the objectives of the instructional sequence. If they do, then teaching to the test is justified. But this does not mean that a test just measures memory, the student's ability to recall exactly what was said in class or what was read in the textbook. We can measure achievement of the objectives by using different words and different applications. Furthermore, tests must yield results that are consistent or stable; otherwise your judgments about a student's performance will not be accurate. You also need to know the ways in which student performance can be judged. Different criteria can be used for deciding whether students are succeeding or need more help. You need to know which criteria you are using and why you are using them when you judge student performance.

Testing is often the heart of an evaluation program. So we start with a critical look at the nature of testing and the characteristics by which they should be judged.

MEASUREMENT

TEST: A DEFINITION

A test is s systematic procedure for *measuring a sample* of a person's *behavior* in order to *evaluate* that *behavior* against *standards* and *norms*. Let's take a closer look at some of the concepts that make up this definition.

Systematic Procedures.

We're all observers, constantly watching the world around us. But most of our observations are unsystematic--and what they tell us

* Note. From Educational Psychology 4/e (pp. 568-611) by N.L. Gage and David C. Berliner, 1988, Boston, Mass.: Houghton Mifflin Company.

may well turn out to be untrue or only partially true. Suppose you're watching some young children playing on swings. Yes, you can see what they are doing, but you can't judge their psychomotor skills. To do that, you would have to hold constant the kind of swing, the chain or rope holding the swing, how they children get started (with a push or by themselves). Your observations would have to follow some standardized procedures, rules and schedules--a system--so that all the children have an equal chance to show their psychomotor ability.

In the same way, you can't judge problem-solving ability by just watching. Suppose your brother and his friend are both trying to figure out a way to get a raise in their allowances. If your brother manages to get a raise and his friend doesn't, does this mean his friend has less problem-solving ability? Without systematic procedures, you can't tell. Maybe he does, or maybe your parents are wealthier than the friend's parents.

The message we are trying to get across is that causal, unsystematic observation is not enough for teachers. For feedback to students and their parents, and for examining the effectiveness of their own teaching, irregular, unsystematic observation is much too fallible. Too much is at stake! When we want estimates of student's achievement, as we surely do a great many times each school year, we have to use systematic procedures to obtain those estimates. A test is a method
for obtaining trustworthy estimates of achievement. A test, when designed correctly, provides the same stimuli for all students. Tests are not substitutes for observations. They are themselves observations of behavior that are more efficient, more refined, and less biased than other ways of observing. Tests are also easier to summarize and interpret than most other kinds of observation.

Measuring.

One reason for the popularity of tests is that they give us a quantitative estimate of ability or achievement; they tell us how much. In education the attributes that interest us--the abilities and achievements of students--are intelligence, creativity, spelling ability, science knowledge, interest in art, and the like. When we quantify a student's social studies achievement, academic aptitude, or appreciation of poetry, we are measuring.

Some tests are *objective*, in that it is easy to get different judges to agree about the score of measure yielded by the test. Objective tests can be scored by clerks or machines. Other tests require expert judgment, and it is harder to get the experts to agree closely or exactly because the criteria each expert uses either are not described or cannot be described; these tests are called *subjective* tests.

Sample.

We must remember that a test samples behavior. We do not test all of a student's mathematic knowledge or ability to understand French. We use only a small number of many possible tasks or problems to determine whether the student knows how to add or knows the meaning of certain French words. From that

sample, we determine whether the student can use information or apply ideas. So that sample should be as unbiased as possible, should cover proper areas of the curriculum, and should provide many ways for the student to demonstrate competence. When certain students perform poorly on a test, the fault may lie with our sampling procedure, not their ability. Every one of us has at some time or other felt ready to take a test only to find out we knew the answers to a lot of questions the teacher never asked. Some methods of seeing to it that a test is a fair sampling of behavior or achievement are discussed in Chapter 24.

Behavior.

Tests are designed to elicit behavior from students. We cannot deal with what students are thinking when they solve problems unless we can observe their thinking through their thinking aloud or their writing or their solutions to problems. We cannot study students' creativity unless we can see its processes and its products. What we examine are what people say or do and their solutions to problems or their productions in music, writing, or art;
that is, we examine creative behavior. That we must rely on what is observable is an important point. The little girl nodding her head in the second row may understand the point you just made, or she may just be socially acquiescent. When you want to measure thoughtfulness, physical prowess, or understanding, you must convert these dimensions into observable behaviors.

Evaluation.

Evaluation is the process by which we attach value to something. Measuring a sample of a student's observable behavior tells us how much of a given attribute that student has. Evaluating the measurement is an altogether separate issue. For example, suppose a student scores 40 points on a reading test. What does this measurement tell us about the student's reading comprehension? Is it good enough or weak, commendable or regrettable? In the process of evaluating, we determine whether a student has mastered the material. Or an evaluation can tell us whether a student should be recommended for advanced training or whether a class is doing poorly or whether a school is achieving at an excellent level. Measurement gives us numbers. Human judgment, concern, and interpretation turn those numbers into evaluations.

Standards and Norms.

Standards and norms are both means of comparison. When we talk about *standards* we mean that people have made a judgment about what is acceptable and reasonable performance for an individual or a group of individuals on a specific test. A student's numerical score is evaluated differently depending on the standards we use to interpret the score. We might want to know whether a score of 40 points on a reading comprehension test is above or below the *established criterion for acceptable competence* in reading.

Such a standard is different from a *norm*, which is used to interpret a student's score in relation to the scores of other

students. Norms require comparisons among students. Some norms, such as those based on students in the same class, are called *immediate peer norms.* Other norms, such as those based on students in a given state or the whole country, are called *distant peer norms.* Both of these kinds of norms are used in norm-referenced testing. Standards, on the other hand, are used in criterion-referenced testing.

NORM-REFERENCED TESTING

Norm-referenced tests are those that use the test performance of other people on the same measuring instrument as a basis for interpreting an individual's relative test performance. A norm-referenced measure allows us to compare one individual with other individuals (see Glaser, 1963).

Using the Immediate Peer Group.

Suppose Lisa received the highest score, an A, on her test in auto mechanics class (her peer group). Gloria, with a B, was seventh in the class. In this norm-referenced situation, using the scores of their immediate peer group as our norms, we can say that Lisa knows more about auto mechanics than Gloria knows, at least as measured by this test. But unfortunately the scores by themselves tell us nothing about what the two girls really know.

Suppose the test they took was used as a basis for admission to an advanced class in auto mechanics. If only six openings were available, Lisa would get into the class and Gloria would not. It may be that even Lisa does not know enough to be able to benefit from the advanced class, but we could not determine this fact from the norm-referenced test because it does not necessarily tell us what Lisa knows about how cars work. It tells us only where she ranks in comparison with others who took the same test. Norm-referenced tests give us a way to pick the more talented from the less talented students in a particular subject matter area. When the norm group is a student's own class or fellow students in the same grade at school, the immediate peer group provides the norms by which we judge performance.

Using a Distant Peer Group.

Let's say that Lisa and Perry both take a national test on knowledge of auto mechanics. Perry scores at the 74th percentile in comparison with thousands of others who have taken the same test. That is, his knowledge of auto mechanics is equal to or better than 74 percent of the students who took this test. We could also say that only 26 percent of the other students know more about auto mechanics than Perry does. The other students, a representative norm group against which Perry's score can be judged, are like a distant, somewhat invisible peer group. To the degree that the distant peer group is meaningful--that is, appropriate as a basis for comparison--we can interpret Perry's score in light of the group's score.

Now suppose Lisa receives a score that places her at the

53rd percentile rank on the national norms. Although Lisa's knowledge when compared with that of her immediate peer group was at the 99th percentile rank, she is clearly less knowledgeable when compared with the distant peer group. If we learn, however, that the distant peer group, the norm group for the test, contains mostly males, we can evaluate Lisa's score differently. In this culture, males tend to know more about the workings of cars than do females. So the comparison may not be a fair one. This is the same problem that many Blacks, Chicanos, and Native Americans face with norm-reference testing when the norms are based on distant, but supposedly representative, peer groups. Minority-group students who show excellent performance relative to their actual peers often seem to perform poorly on nationally standardized tests. Why? Because the norm group used to judge their scores is not really representative; it does not include enough members of minority groups. For certain purposes, then, the norm group is biased. A biased norm group can unfairly penalize a minority-group student who has a different cultural background. In the case of Lisa, if the test is being used as a basis for admission to an auto mechanics course, perhaps her performance should be compared only with that of other women. That is, we might want to use different norms to evaluate performance on a test if we are concerned about equal opportunity, as we are in an admissions test. Here social class, gender, race, and ethnicity can be relevant considerations. But if the test Lisa took was an evaluation of course work and we want highly competent rather than mediocre auto mechanics, then social class, gender, race, and ethnicity are probably irrelevant.

CRITERION-REFERENCED TESTING

If we really want to determine what Lisa knows about cars, we would have to abandon the norm-referenced approach and turn to a criterion-referenced approach. No matter what we learn about a person's standing relative to his or her peers, in the norm-referenced approach we can never learn whether the person knows a certain thing, such as how to diagnose problems with automobile transmissions. Except when the purpose of testing is to select a fraction of students for scarce positions (for example, at a highly competitive college or for a small honors class), a student's achievement should be assessed through criterion-referenced, approaches. For most school purposes, criterion-referenced testing markedly improves assessment and evaluation.

Criterion-referenced tests measure an individual's ability with respect to some criterion. The criterion, or standard, is determined in advance by knowledgeable people in the field. Automobile driving tests are not norm referenced (who wants the best of a group of bad drivers?); they are criterion referenced, based on standards of performance. Comparison among students is not a factor here. The criterion-referenced measure is used when we want to know what students can do, rather than how they compare with immediate or distant peer groups (Popham & Husek, 1969). The criterion-referenced test is deliberately constructed to give information that is directly interpretable in terms of an absolute

criterion of performance (Glaser & Nitko, 1971).

The absolute criterion is usually based on a teacher's experience with students, on the particular curriculum area, on records of past performance, and also on the teacher's intuition and values. Suppose a teacher decides that students must be able to solve eight out of ten problems of the following kind:

Which word is the name for something worn on the head? (HAT/FAT/CAT)

In this case the teacher would read the question, and the students would pick answers that demonstrate reading skills. If a student can do eight, nine, or ten problems correctly, the teacher concludes that reading skills of the type tested have been mastered. Failure to reach the teacher-set criterion is defined as seven or fewer items correct.

In this kind of testing, we learn that a student can or cannot read. Comparing Henry's test score with Ann's test score does not meet a parent's or a teacher's need for information. It is too bad that parents do not always recognize this. They often want information about how their child compares to others rather than information about the actual knowledge and skills their child has learned. Teachers need to educate parents that the criterion-referenced approach is often more in tune with classroom needs than is the norm-referenced approach.

Usually criterion-referenced tests are graded with a pass-fail system. Our concern here is whether each student has mastered the material, not how much each student knows in relation to what other students know. One student may do more work in an area than another. One student may be quicker to finish work than another. But when it comes to judging proficiency in certain curricular areas, criterion-referenced tests allow us to determine how well students are meeting the criterion. Perhaps all the students will pass a test; perhaps none of them will. If there are several standards for different letter grades, perhaps all students will receive a C, or none will.

It is very important that the criteria of performance be set before students take a test. Then the criteria cannot be influenced by how well the students do on the test. And then it is impossible for the teacher to adjust standards so that some predetermined percentages of the students receive grades of A, B, and so on. The choice of either norm-referenced or criterion-referenced testing, then, has important consequences for your classroom and school. Table 22.1 lists some pros and cons of each type of testing (Clift & Imrie, 1981; see also L. Shepard, 1979).

Reliability

There are two major criteria for judging the tests we use: reliability and validity. A good test is both reliable and valid. Here we examine test reliability; in the next section, we look at test validity.

Suppose we are interested in the academic performance of a student named Carmine. She has just finished a test of knowledge about botany. The test had 40 items, each worth 1 point. Carmine

received a score of 24. To interpret Carmine's test performance in terms of norms we would find it useful to know that the mean or average score of the class was 29. To interpret her test performance in terms of criteria, we would have to know that the teacher had set 80 percent or more items correct as the criterion of competent. So 32 (.80 x 40) was the cut-off score to determine who has or has not mastered botany. Given this information and either a norm-referenced or criterion-referenced perspective, we could say that Carmine's performance is somewhat below average, or that she has not reached the mastery criterion.

But the issue is not this simple. With any test, we must ask whether we would make the same decisions about Carmine's performance if she took the same or a similar test again within a short period of time. We must consider how precise, consistent, or stable the test performance of a student is. To do so, we need to know his performance over different testings and time spans. The precision, consistency, or stability of a score are different aspects of what we call *test reliability*, one of the fundamental characteristics of tests. If we believe that the decisions we make about people on the basis of a test will be the same from time to time, we are assuming that the test is reliable.

TEST-RETEST RELIABILITY

How do we know whether a test gives us reliable information about a student? One way is to retest Carmine (with the same test or a parallel test, at a later time) and see whether her score is about 24 on the retest. If it is, our evaluation that Carmine is slightly below average in knowledge of botany or has not mastered botany seems to hold. The test appears to be a reliable instrument. Moving from one student to a class of thirty, we see that reliability refers to the degree to which the scores received by students on one test occasion are about the same as the scores they would receive if tested again on a different occasion.

When we are interested in norm-referenced interpretations of scores, the reliability question becomes a question of the degree to which the rank ordering of individuals will be the same from one time to the next. Insofar as students' ranks according to their scores are about the same from one test occasion to another, the test may be considered reliable in the sense that it is highly stable.

One estimate of reliability for norm-referenced tests that we might use is the correlation between the ranks obtained on the first testing occasion and the ranks obtained on the second testing occasion. Remember that a correlation tells us the degree of relationship between two things. Suppose we had five students take a test and then retake it, with the results of the second testing described below in columns A or B.

If we obtained the results in column A, we would not be too surprised. There have been some changes in ranks: For example, the first- and second-ranked students switched ranks, and so did the third- and fourth-ranked students. But the least able student on the first test remained fifth ranked on the second test. The test has some stability in terms of how it orders people from lowest to highest in ability, although it clearly is *not* perfect. (You can

compute the rank-difference coefficient of correlation with the procedure described in Appendix A.) But if the results were like those in column B, there are too many changes in rank to feel the test is stable or dependable. In this case the test probably is not reliable. (Again, you can compute the coefficient. Is it further from + 1.00 than the first coefficient?)

The coefficient of correlation between the ranks obtained by students in a class can be used to estimate stability, dependability, or reliability, giving us a numerical value to use for interpreting a test's reliability. If two rank orderings of test scores were exactly the same, the correlation would equal 1.00, which would mean the test was perfectly reliable. If the correlation equals .90 or .84, the reliability is high because the rank orderings are pretty much the same on the two testing occasions; that is, the information we get from a score, which we then use to make decisions, stays relatively constant from one testing occasion to another.

Generally, we need reliability coefficients above .80 to make important decisions about students from norm-referenced tests. Lower reliabilities generally are not acceptable. But if a student scores extremely high or low on a test, some decisions become safe even if the reliability is lower than .80. That is, even with low test reliability, say about .60, we would not expect the top- and bottom-scores to switch places completely. Perhaps the low-scorers move up a bit and the high-scorers move down a bit, but those at the extremes of a distribution move around a lot only if reliability is down near .00.

When we deal with criterion-referenced tests, reliability has a different meaning. The precision of the score is less important to us than the dependability of our *decision* about whether Carmine has or has not mastered the botany unit (Mehrens & Lehmann, 1980). In a way, Carmine's exact score is not very important--all scores under 32 are grouped together as failures, and all scores at 32 or above are grouped together as passes. The interpretive decision, not the actual score, is what matters.

The test-retest method can also be used to estimate the reliability of our decision about Carmine. If she was retested and we found once again that she has not mastered the unit, the test shows some reliability. For a group of students, we would analyze the decisions made about them each time they were tested. We might ask how we classified Carmine, Henry, Phyllis, Herman, and the other member of the class on each test. Inspecting these data tells us whether or not the decisions are being confirmed.

The numerical indices of test-retest reliability for criterion-referenced tests provide either correlational information, which is what we've been discussing, or probability estimates (Berk, 1980; Sweezy, 1981). If one criterion-referenced test has a probability test has a probability of .82 for correctly classifying "masters" and "nonmasters," it is more reliable than a test that has a probability of only .55. Both correlational and probability estimates of reliability are numerical values between 0 and 1.00 and are interpreted in a similar way--the closer to 1.00, the higher the estimate of reliability.

INTERNAL-CONSISTENCY RELIABILITY

To determine test-retest reliability, we have to give students the same test (or a parallel form of the same test) twice. But most teachers do not have the time to retest students. They have to be able to determine reliability from a single administration. Internal-consistency reliability does not reflect stability over time; it indicates how precisely a single test measures whatever it is supposed to measure.

How does it work? What we do is estimate the correlation between the test we are giving and some hypothetical test that would be given at the same time. The statistical procedures for computing internal-consistency reliability are too complicated to talk about here. But an introductory testing or measurement textbook can give you the information you need to estimate internal-consistency reliability for your own norm-referenced tests. Still under development are procedures for determining from a single administration the reliability of criterion-referenced test--tests that tell us whether a student has mastered a subject matter area. In their current form, these procedures are still too difficult for classroom teachers to use easily (see Hambleton, Swaminathan, Algina, & Coulson, 1978; Subkoviak, 1980).

THE STANDARD ERROR OF MEASUREMENT

More important than the ability to compute reliability estimates is the realization that scores from tests have less than perfect stability and internal consistency. Reliability estimates for tests in education are always less than 1.00 and often less than .80. Therefore the scores for individuals will fluctuate a little (or a lot) from one occasion to the next. Be sure to remember, then, if you ever have to make decisions based on a single score for a student (say, assigning all those who have scored above 45 on a certain test to the gifted program, or giving Ds to all those who have scored below 18 on a certain test), that the observed scores you are using to characterize each student are not all that precise. The lack of precision in scores--a lack that results from unreliability--is reflected in a statistic we call the *standard error of measurement.*

Any score on any test is made up of "error" as well as the "true" level of the attribute we are measuring. When we measure IQ, we measure an aptitude for performing certain intellectual tasks. But that measurement is affected on any given day by the person's health, emotional state, motivation, rapport with the examiner, recent practice in the area being tested, attention, coordination, memory, and fatigue (see Cronbach, 1970; Stanley, 1971). Of particular importance here is the individual's luck in guessing on a given test occasion.

An observed score, say, Carmin's 24 points on the botany test, is made up of true score and error score. The estimated amount of the error in a person's score is what we want to know. After we estimate the error, we can estimate a *confidence band* around the observed score and be pretty sure that a person's true score is within that confidence band. The standard error

measurement (further described, with its formula, in almost all introductory testing and measurement texts) provides us with the information we need to develop a confidence band for observed scores. By tradition we usually talk of 68, 95, or 99 percent confidence in estimating true scores.

Suppose we learn from the use of the statistical formula that the standard error of measurement was 4 points for the botany test that Carmine took. Then we would know that Carmine's true score was probably somewhere in the range from 4 points below to 4 points above the score she received. That is, her true score would probably be between 20 and 28. This is the confidence band, and from our statistics we would have confidence that we were right 68 percent of the time.

Because of the errors in our measurement system, we must learn to think of Carmine's score, and those of our other students, as falling within ranges, not at precise points, on our tests. Knowing there is a standard error of measurement also keeps us on guard when we compare two or more students whose observed scores seem different. Suppose Henry's observed score on the botany test was 31. At first glance it would appear that Henry is slightly over and Carmine under the average for the class, and that Henry scored 7 points higher than Carmine. But when we use the standard error of measurement to determine confidence bands, we find a situation like the one shown in Figure 22.1. Thinking in terms of confidence bands, not precise scores, leads us to different interpretations. First, we notice that the two confidence bands overlap. So we probably do not want to conclude automatically that Henry's true score is higher than Carmine's. If Henry's true score is at the low end of the confidence band determined for his observed score and Carmin's true score is at the high end of the confidence band around her observed score, Carmine would actually be performing better on the botany test than Henry!

Because scores are not precise points, but rather indicators of bands, many people who use criterion-referenced tests add a third category to "mastery" and "nonmastery"--a category called "no decision." To reflect the acknowledged lack of precision in the scores on the botany examination, it might pay to designate 30 to 33 items correct as a band within which we need more information before classifying someone showing a "mastery" or a "nonmastery." The size of the band we choose reflects our concern about the magnitude of the standard error of measurement.

All this discussion of unreliability and error of measurement is designed to keep you cautious. In any testing you are involved with, in any tests you interpret, in any selection or grading you do, remember that each score has associated with it a confidence band that defines the range of error for that particular score. If reliability is less than 1.00, the obtained score contains some error. Do not think of numerical scores as God given. They are fallible, as are the people who must interpret them.

IMPROVING RELIABILITY

One way to improve reliability, and thus reduce the standard error measurement, is to increase the length of a test. This is true for norm-referenced or criterion-referenced tests. You should remember that all tests sample behavior. Of the hundreds of test questions that could be asked in an area of arithmetic or about a novel, only a subset of items are used to make up a test. If a student has a momentary lapse in attention or makes a careless error on one question, he or she is sunk if the test is short. Or if a student doesn't remember one character or incident in a novel, and that is what all the questions on a test are about, again the student is in trouble. More items that sample widely from a domain of knowledge really work to a student's advantage when taking a test. Although students who take tests and teachers who construct tests both complain about the tedium of lengthy examinations, the relation between reliability of .20, adding 5 items of the same type would increase the estimated reliability coefficient to .33. If you add 15 items of the same type, the reliability estimate increases to .50. When the reliability of a test is low, adding items increases the reliability coefficient. When reliability is already high, adding items does not have much effect.

Validity

Test validity is the degree to which testing procedures and interpretations help us measure what we want to measure. It is the single most important issue to consider when evaluating a test (Committee to Develop Standards for Educational and Psychological Testing, 1985). This is because so many tests really do not function as they should. A highly reliable test of ability may not identify the most creative students in art, only those who have learned art terms. Or the test may identify students for an art program who are really no better at art than the ones who were not picked. Reliability is important, but ultimately what we most want in a test is validity. We want the test to measure what we intend to measure--measure achievement of a certain kind, the promise of various candidates for scarce positions, the latent talent of students, and so forth. There are many different kinds of validity, but we concentrate here on just three: content validity, construct validity, and criterion validity (see Cronbach, 1971; Messick, 1980).

CONTENT VALIDITY

We must be sure when we construct, say, a geology test that we are really measuring geology knowledge and skills. We have to be sure that the questions pertain to what we've taught--either the entire curriculum or just the material in one class. We have to know that the questions are representative of the material and that there are enough of them to sample adequately the different kinds of knowledge and skill in that domain.

If a social studies test asks questions that could be answered on the basis of general intelligence, test wiseness, or regular

newspaper reading, the course content in social studies is not being tested adequately. *The items on an achievement test should be tied to an instructional domain that students have had an opportunity to learn.* If independent experts agree that a test in eight-grade social studies is measuring the common curriculum in that subject area, the test has content validity. As the social studies subject matter changes or as new subtopics are stressed, the content sample for the eight-grade social studies test must also change if it is to remain valid.

Content validation is a logical procedure; it is based on good sense. (In Chapter 24 we describe how to define a domain and sample from it.) If we don't rely on sensible sampling procedures, we probably cannot interpret our tests in the way we intended to. That is, a test is not a valid test of geology or social studies or whatever unless its content is appropriately matched with the defined domain of interest in geology or social studies or whatever. Content validity is a special problem for norm-referenced tests; it is less of a problem with criterion-referenced tests, in which instructional objectives are tied directly to test items.

CONSTRUCT VALIDITY

Construct validity is perhaps the most difficult kind of validity to understand. It deals with the question of whether a test measures the attribute or characteristic it claims to measure. We call certain abstract characteristics or attributes of people *constructs.* Intelligence is a construct. So are creativity and anxiety. We cannot measure these attributes, characteristics, traits, or constructs directly, the way we do arithmetic achievement or spelling ability. So we invent the idea of intelligence or creativity or anxiety to talk about a complex set of behaviors that, all together, seem to indicate that a person will act intelligently, creatively, or anxiously. How do we tell if a test we are using to measure, say, scientific aptitude is really measuring the construct of scientific aptitude? Maybe it is measuring only knowledge of science, general intelligence, reading comprehension, or all of these. This is a tricky question. Too many tests pass as tests of creativity, art aptitude, or mechanical knowledge when they do <u>not</u> measure the attribute they claim to measure.

How do we check for construct validity? One way is to use correlations. If we have a test of art aptitude and a rating of how well students did in art, the two measures ought to be related. The same would be true for a test of mechanical aptitude and performance in auto mechanics, or any other aptitude test and a criterion measure of the corresponding kind of performance. Even a moderate correlation would be reassuring. We can also correlate one test of, say, IQ with another. For example, a new short IQ test had better correlate substantially with the Stanford-Binet and Wechsler IQ test, both well-accepted tests of the construct of IQ. If the new measure of IQ does not have much in common with these tests, then what it measures is not what we usually mean by intelligence, regardless of what the test is intended to measure.

Another way to check a test for construct validity is to test hypotheses about how high-scorers and low-scorers should act. When we test leadership, self-concept, attitude toward mathematics, or other constructs, high-scorers and low-scorers on these tests should act differently. If they do in fact behave the way we expect them to, the tests have a claim to construct validity.

Claims about what a particular drug will do are scrupulously examined by federal and state agencies, but there is no such regulation of test development or the claims associated with tests. Education is filled with people who make up tests and advocate their use without looking carefully at whether they measure what they are supposed to measure. Teachers, then, have to be careful about the tests they use.

CRITERION VALIDITY

If you are using a test for selecting students for admission to a school, curriculum, or course, you must make sure that it is valid for the purpose. Suppose you have a program for academically gifted students and are using a creativity test for selection. Does the test actually select students who are more likely than others to profit from the program? To estimate the criterion validity, we need to test a group of students and let all of them, whatever their scores, into the program. We would then correlate scores on the selection test with scores on some criterion measure that reflects success in the instructional program. This allows us to see the degree to which the high-scorers on the selection test profit more from the program than the low-scorers on the selection test. If they do profit more from instruction, and if in the future we want to choose students who will get the most from the special program, we can use the test for selection. The test is valid for predicting who will do well on the criterion. (What we are calling criterion validity is sometimes called ***concurrent*** or *predictive* validity.)

Once again, a warning: Criterion validity coefficients provide us with a useful basis for selecting and counseling students in various curriculum areas, but only when a particular student is like the students who were in the sample on which the validity coefficient was determined. Teachers who use tests as a way of getting information to help them make decisions should find out whether a test has a substantial criterion-validity coefficient. Then they must check whether the group used to determine that coefficient is like the group to which the information is now being applied. Selection for special programs can be helped by valid tests, but selection should also take into account the unique circumstances of a particular student, the student's motivation, and the meaning of wrong decisions.

RECAP

Good tests are reliable. This means we can trust the information they give us about a student, believing it to be precise,

consistent, and stable. The reliability coefficient and the standard error of measurement tell us how much faith we can have in these aspects of a particular test score. We need to think about test scores as having confidence bands around them. Errors of measurement and true scores are always mixed together.

Most important, good tests are valid: They measure what we intend them to measure. Achievement tests should have content validity--a logical match between a test and the domain it is intended to sample. Other tests should have construct validity--they should measure the constructs (attributes, traits, tendencies, among others) they claim to measure. What a test is called and what a test actually measures may be very different. Finally, the criterion validity of tests tells the degree to which tests used in selection and counseling predict some criterion (or accepted measure) of performance in courses or jobs. Valid tests measure what they are supposed to measure.

Evaluation

Evaluation has become a specialized activity in education. People now are trained to be evaluators. Agencies that fund education projects are specifying that the projects be evaluated; teachers and administrators have being asked more and more for evidence that what they do is working. Evaluation is part of educational accountability (holding educators accountable for the success of their efforts), product development, and curriculum development.

Evaluation of this type is concerned with the collection and use of information for making decisions about programs, curricula, teaching methods, and other school activities (see Cronbach, 1963). Evaluators address these kinds of questions:

- How have the new attendance boundaries in the district affected segregation?
- How can I change the reading program to help students learn more?
- How do students like the food since we hired the new cafeteria manager?
- Should this teacher get a merit raise?
- Is the new curriculum on patriotism doing what it should?

Evaluators are educators, helping members of a policy-shaping community to recognize their own interests, weigh consequences of alternative approaches, and discover new ways to perform their tasks (Cronbach, 1980). Evaluators are detectives: They look for evidence to shed light on some problem. They must collect evidence to help others make decisions. The evidence may include measure of student achievement by means of teacher-made and standardized tests, observations of teacher and student behavior, attitude measurements, surveys of parental opinion, financial costs, and a variety of other things.

Whatever is being evaluated--a reading unit or an entire science curriculum--two kinds of evaluation go on: formative evaluation, which is used to change a program so that it operates as it was intended to operate, and summative evaluation, which is

used to judge a program on the basis of how well and at what cost it brings about wanted outcomes (Scriven, 1967).

FORMATIVE EVALUATION

Formative evaluation is what we use as a basis for revising materials or programs. It is the responsibility, not only of the developers of the materials or programs, but of teachers and administrators too. Is the new film on mitosis effective? Are the inductive methods you are using to teach the concept of "peninsula" working? Are the instructional games for use with Native American youngsters in the Southwest worthwhile? Whenever you try out new materials, teaching methods, or programs, you must implement a system for monitoring the innovation. You can work alone or with other teachers. Your evaluation is the basis for improving the program or abandoning it.

Of course an innovation should be used in a way that gives it a fair chance to work in your unique classroom circumstances. If it is a new product, say, a teacher-training module or a logic game for students, it might have to be modified many times by yourself or the developer before you are satisfied with it. Formative evaluation is concerned with making activities or materials work the way they are supposed to.

SUMMATIVE EVALUATION

After a product, program, or activity has been refined, modified, and used, a summative evaluation may be called for. A good summative evaluation examines competing methods or programs too, to see whether they can produce the same or better results fore less money or in less time. The usual summative evaluation is like a horse race: Curriculum A is pitted against curriculum B; textbook C is pitted against textbook D.

The horse-race approach is narrow, but it seems to be valued. From experience with these kinds of comparisons, educational psychologists now know enough to expect that in any comparison of educational programs or products, each will have better effects in those curriculum areas that are deliberately stressed by the materials, and each will do about as well as the other in those areas of the curriculum that are not especially emphasized (see Walker & Schaffarzick, 1974).

The evaluation of materials or activities must also take into account factors other than educational effects, namely, cost, time required, ease of use, needs met by the program, attitudes of students, preferences of teachers, judgments of specialists, and so on. Summative evaluation is complex because many sources are used to obtain information for making decisions about keeping or abandoning the materials or activities.

Teachers are rarely involved in summative evaluation. This is a job for unbiased evaluators from outside the school system. These evaluators must be skilled in measuring achievement, attitudes, opinions, and interests and in examining the economics,

management, and politics of education. Cronbach and his associates (1980) called these evaluators "public scientists"--people who use systematic inquiry techniques to affect social systems.

Teachers have a right and a responsibility to demand summative evaluation reports. When consumers want information about washing machines, automobiles, or cameras, they can get that information from local consumers groups, agencies of the federal government, or Consumer Reports, a magazine that offers summative evaluations of many products each year.

What do we do? Where do we get the information we need about the effectiveness of textbooks, curricula, audiovisual aids, workbooks, educational games, computer programs, and other educational projects? We have to begin to ask product developers to answer questions like these:

- Which of my students will profit from these materials--the brightest only, middle-class students only, or others as well?
- What skills and knowledge will my students have after using the materials? Can you show me what a final test would look like?
- What is the cost per student? Are the materials reusable? Will the books last? Can I make copies?
- How much of my time will be needed? Of students' time? What are the prerequisites for students and teachers who use these materials?
- How does this product compare with the one I have been using? How do you know?
- Can you give me the name of someone in the area who is using these materials?

Teachers have a responsibility to help improve educational quality as much and as quickly as possible. One way to meet that responsibility is to demand summative evaluations of new teaching programs and materials--in short to be good consumers.

SUMMARY

A primary function of teachers is the measurement and evaluation of students and programs. Often we use tests in the evaluation process. A good test relies on systematic procedures for observing student performance and quantifies that performance. A test samples only a small amount of what students have learned. But that small amount can provide a meaningful indicator of what has been learned. We use the measurement of observable behavior--the student's score--to evaluate learning.

Norms allow us to interpret a student's score in relation to the scores of other students--in the immediate peer group or a distant peer group. Norm-referenced testing tells us how a student's knowledge compares with that of other students; it does not tell us what the student actually knows. For this kind of interpretation, we have to use criterion-referenced testing--tests

that allow us to compare student behavior with some preset standard of performance.

A good test must be both reliable and valid. Reliability has to do with stability (reliability over time) and internal consistency (precision). No test is ever perfectly reliable. To take account of the standard error of measurement, we have to think of a student's score, not as a precise point, but as a range--a confidence band around the observed score.

Validity--that a test measures what we want it to measure--is the most important element in the evaluation process. An achievement test has content validity when it tests material that students have had an opportunity to learn. Construct validity is more difficult to determine because it deals with the degree to which a test measures an abstraction. Here the question is whether a test measures the construct (ability, trait, or tendency) it is intended to measure. Finally, criterion validity tells us how well a test predicts success in an instructional program or a job.

We've been talking about testing as a means of measuring and evaluating learning. But teachers are also responsible for evaluating the programs and materials they are using or plan to use. That is, we use evaluation to make decisions not only about students' learning but also about the programs and materials we use to teach them.

Formative evaluation gives teachers and students the information they need to improve programs or curricular materials. Summative evaluation, which is usually conducted by outside experts, tells us whether programs and products are producing what they were intended to produce at reasonable costs in money, effort, and time. Teachers should demand more information about evaluation from curriculum developers and those who sell educational materials.

Popham (1981) dramatically summed up the need for educators to understand tests, measurement, and evaluation:

> We live in an era when everybody adores evidence. We want evidence that patent sleeping pills do indeed send us off to the land of Nod. We want evidence that food additives are not carcinogenic...we want evidence that the nation's schools are effective.
>
> There is no doubt that we are living smack in the middle of an evidence-oriented era. That evidence orientation having intruded most dramatically on the educational enterprise, should force educators to consider a fundamental truth. In an evidence-oriented enterprise, those who control the evidence-gathering mechanisms control the entire enterprise. Since, in education, tests constitute our chief evidence-gathering mechanisms, it is apparent that all educators should become knowledgeable regarding the fundamentals of educational measurement. (pp. 5-6)

CHAPTER 23
STANDARDIZED TESTS AND THE TEACHER

Overview

All through your schooling you have been dealing with standardized testing. But you have always been on the student's side of the process. In this chapter we talk about the other side--the teacher's and, more generally, the educator's. Both standardized tests and teacher-made tests rest on the basic approaches and theory we looked at in Chapter 22, but they offer different advantages and serve different purposes. Let's look at these before we go on to the ways in which we select, administer, and interpret standardized tests.

ADVANTAGES AND SPECIAL USES OF STANDARDIZED TESTS

A standardized test is one that has been given to a large representative sample of some population so that scores on the test can be compared with those of the people in that sample. That is, norm-referenced standardized tests provide norms that make possible the comparison of any student's score with those of many other students. Whenever you want to evaluate a student against criteria that go beyond a single teacher's classroom and that teacher's conception of what should be taught, you need a standardized test. A comparison of standardized and teacher-made achievement tests (discussed in the next chapter) is shown in Table 23.1.

In addition to having norms--that is, bases for interpreting scores in terms of other students' scores--standardized tests are usually more carefully constructed than teacher-made tests because they are constructed by experts, using the technical, statistical, and research knowledge of the testing field. Also standardized tests usually come with more or less detailed background information about the test--its rationale, its purposes, and the ways in which its content and items were chosen. This information includes evidence about the test's reliability and validity (see Chapter 22). And standardized tests usually have detailed instructions abut how the test should be administered--the exact directions to be given the students, the time limits (if any), and the ways in which the teacher should handle any special problems that may arise.

Criterion-referenced standardized tests are becoming more popular. These tests yield direct information about how well a student has performed in relation to some specific criterion, rather than in relation to the performance of other students. Some tests

of reading disabilities, the Red Cross lifesavers' tests, the requirements for merit badges in the Boy Scouts--all are criterion-referenced standardized tests that are being used today in educational programs. And criterion-referenced tests are rapidly being developed to measure achievement in many other areas of the curriculum. Example of the information a teacher receives from a well-known norm-referenced standardized test and a criterion-referenced test are given in Figures 23.1 and 23.2.

TYPES OF STANDARDIZED TESTS USED IN SCHOOLS

The two main kinds of standardized tests used in schools are aptitude tests and achievement tests.

APTITUDE TESTS

When tests cover broad areas of intellectual functioning, they are called intelligence tests, scholastic aptitude tests, academic aptitude tests, or general ability tests. When they are aimed at more specific kinds of intellectual ability, they are called special ability or aptitude tests. The most frequently used tests of this kind are verbal, mathematical, spatial, mechanical, and clerical aptitude tests.

In general, aptitude tests are norm-referenced and provide information for student guidance and counseling. In colleges, business, and industry they are used for selection and placement. Students have been put into special classes, or treated differently within regular classes, on the basis of their scores on aptitude tests. As we saw in Chapter 4, controversies rage over whether aptitude tests (in the form of intelligence tests) in the schools do more harm than good. Certainly when Spanish-speaking children are classified and treated as mentally retarded because they do poorly on scholastic aptitude tests printed in English, the conclusion that these tests are harmful is easy to defend. Some writers think that the tests do harm by influencing teachers' expectations of their students. They argue that teachers act on their expectations, treating some of their students inappropriately--calling on them less often, "staying with" them in classroom recitations less often, giving them enriched assignments less often, and so on (see Good, 1983). It is argued that any measurement of any student characteristic that is thought of as an aptitude measure--a prediction future learning or performance--is in effect a prophecy that teachers may then unconsciously fulfill, to the detriment of the students who have done poorly on this kind of test. Others argue that teacher knowledge of a student's general or special ability can help the teacher adjust teaching, explanations, assignments, and overall treatment so as to challenge the more able and avoid frustrating and discouraging the less able. So far research has only revealed the possibilities; it has not shown how to make sure students benefit. It is also clear that teachers can form fairly accurate impressions of student aptitude just from

student performance in class, without using aptitude tests. These impressions, however, can have the same dangers or advantages as those based on aptitude tests.

ACHIEVEMENT TESTS

Standardized achievement tests are used to measure students' achievement of the objectives of instruction in a given course or other curriculum unit. We have standardized achievements tests in, say, third-grade reading, fifth-grade arithmetic, seventh-grade social studies, ninth-grade algebra, and eleventh-grade chemistry. The tests have been constructed by specialists in the curriculum areas and have been administered to large samples of students in the appropriate grade levels and courses. They tell us how well any particular student, class, school, school district, county, or state has done in comparison with the norm group.

Thus you often see newspaper articles reporting that the students of a certain city have fallen above (or below) the norm for the state as a whole in reading achievement. Or an article may state that, within a given city, certain schools do better and others do worse than the citywide average in achievement in sixth-grade arithmetic. Comparisons of this kind are possible only with standardized, as against teacher-made, tests of achievement, and, as we said in Chapter 22, only if the group being compared is like the norm group.

THE DIFFERENCES BETWEEN APTITUDE AND ACHIEVEMENT TESTS

How do aptitude tests differ from achievement tests? They differ primarily in function: Aptitude tests are used to *predict* achievement, or the outcomes of *future* learning experiences; achievement tests are used to measure and *evaluate* achievement, or the outcomes of *past* learning experiences. But achievement tests often predict future achievement as well as, if not better than, aptitude tests do. So the distinction between the two kinds of tests in terms of function does not always hold up well.

We can also make a distinction in terms of content. In an achievement test, the content should, and usually does, deal with what is taught directly and intentionally in schools. Here we ask questions about the content of what has been read in textbooks, discussed in class, practiced in homework and at the chalkboard, and explained by the teacher. Achievement test content should always be high in what we might call "taughtness"; aptitude test content need not be. We would not usually ask questions about the geography of a state or French vocabulary on an aptitude test, and we would not usually include questions that measure spatial ability (which is not taught in most schools) on an achievement test. But this is sometimes done, and, when it is, confusion reigns.

Experts also have trouble with the distinction between aptitude and achievement (Anastasi, 1980; Ebel, 1980, Green,

1974). For your purposes, it enough to remember that aptitude tests should have validity for selection and prediction, and that achievement tests should be high in "taughtness" and content validity. Remember too that aptitude tests, which predict future performance, cannot be criterion referenced. We do not expect a future astronomer or jet engine mechanic to meet the criteria for acceptable performance in those fields now. We only want to find the best candidates for training programs in astronomy or engine repair. Achievement tests, however, can be either norm or criterion referenced.

NONCOGNITIVE TESTS

In addition to ability and achievement, teachers and schools sometimes use standardized tests to measure other human characteristics. These kinds of tests are often called *noncognitive*, but you should realize that it is virtually impossible to have a truly "noncognitive" test (see Weiss, 1980). Messick (1979) has proposed a classification of the noncognitive variables that are of interest to teachers. He identified twelve areas, among them attitudes, interest, motives, temperament, social sensitivity, cognitive styles, and values. Because the use of the instruments to assess these characteristics lies beyond the scope of this book, we do not discuss them in detail here. You should plan to work closely with school psychologists and counseling psychologists when measure of temperament, cognitive style, attitudes, and values are needed.

SELECTING STANDARDIZED TESTS

Hundreds of standardized tests have been published. How do we decide which to use? Actually there are two kinds of problems here: deciding which criteria to use in making a choice, and getting the necessary information about standardized tests to use in applying the criteria. These problems usually must be solved either by individual teachers or, more often, by committees of teachers in a given school district who have been appointed to recommend standardized tests for use throughout the district. Sometimes the teachers are helped by a school psychologist or curriculum director who is relatively well versed in the technicalities of test construction. In any case, teacher themselves can do a better job if they know what to look for and where to find it in selecting standardized tests.

THE QUESTIONS TO ASK?

This is not simple matter. The criteria for selecting standardized tests have been the subject of scores of textbooks and monographs over the last several decades. Our list, based on Popham (1981) and the Center for the Study of Evaluation (for

example, CSE, 1976) is presented in checklist form in Figure 23.3. Briefly, you might ask the following twelve questions about standardized tests.

1. Is the behavior that is measured adequately described? The developers of commercial, statewide or districtwide standardized tests always note in the test manual some description of what they think they are measuring. It could be a brief statement of an objective or a detailed statement of the behavior-content matrix used to develop test items. You are likely to find that the manuals for norm-referenced standardized tests provide more general statements about the attributes they are measuring and the procedures for choosing items than do the manuals that accompany criterion-referenced tests. Your job is to decide whether the description gives you confidence that you know what the test claims to measure.

2. How many items per measured behavior are there? It is customary in criterion-referenced tests to have only a small number of items per objective. Perhaps three items of two-column addition may be enough to satisfy some people that a student's ability in two-column addition has been adequately measured. But would three items be enough to satisfy you? What if we had a three-item test of vocabulary? Almost everyone would agree that you cannot measure vocabulary with a three-item test. So part of the decision about whether or not we have a sufficient number of items depends on the subject matter we are teaching. Norm-referenced tests face a similar problem.
If you have a sixty-item elementary reading and language arts test, there may be one item on prefixes, one item on syllabication, four items on comprehension, two items on grammar, and so on. To cover the broad range of curricula in use in some countries, a nationally normed standardized test may have only an item or two for each area in your language arts program. Is this enough? Obviously there are only subjective answers to this kind of question. But you are going to have to make informed subjective judgments if you are going to be an intelligent user of standardized tests.

3. Is the test valid for your use? In Chapter 22 we noted that criterion-referenced test developers generally do a good job on *content validity*. But this can be a problem for norm-referenced test developers. A careful comparison of three leading textbooks in fourth-grade mathematics with five frequently used norm-referenced standardized tests had startling results. As you can see in Figure 23.4, in the best match between what a textbook taught and what a test measured, only 71 percent of the topics tested actually were taught! In the worst case, only 47 percent of the topics tested were covered in the curriculum. These kinds of mismatches between curriculum and tests grossly underestimate what students learn in school. A teacher or committee can judge content validity by studying the test's questions and asking for each question, Is this something I (we) teach? Is the topic dealt wit

the way I teach it? The answers to such questions lead to decisions about content validity.

Construct validity can be evaluated quantitatively by examining the correlations of a test's scores with other indicators of the construct it supposedly measures. A group IQ test ought to show a moderate to high correlation with an individualized test of intelligence such as the Stanford-Binet. And an achievement test ought to correlate with teachers' ratings of student achievement in that curricular area. If a test is going to be used to predict something important, say, performance in a special school, it must have high enough *criterion validity* to make you comfortable with your decisions. How high is high? There is no simple answer. You must think through how you are going to use the information from the test and what the costs are of a wrong decision.

4. How reliable is the test? To what degree do two or more testings with the same or parallel instruments give the same information for decisions? How large is the standard error of measurement? You must ask whether there is too much error associated with scores or with decision making for the test to be useful for the purposes you have in mind.

5. Does the test provide sufficient feedback to the teacher? Tests that give us clear, simple information that can be used by relatively untrained people for making decisions rate high on "teaching feedback." For norm-referenced tests this has a lot to do with the norm group because feedback to a teacher about a student's or a class's performance is dependent, in part, on how easy it is to interpret the performance given the norms. Thus the value of feedback to a teacher may depend on the breadth of the age, ability, and educational levels covered by the test's norms. It may also depend on how representative the norm group is--the geographic areas, ages, cultural groups, and types of schools and school districts it was drawn from the recency of its testing. Remember that the interpretation of a single individual's score on a norm-referenced test depends on how closely that individual resembles the norm group. Another consideration is how easily the raw scores yielded by the test can be converted into the kinds of scores (percentile ranks, standard scores, or whatever) used in the table of norms. For criterion-referenced tests we should expect that when an objective like "The student can comprehend literally stated written materials" is tested, we would be told through the wonders of computer analysis and printouts who passed, who failed, and whom we need more information about. This feedback is crucial for the teacher who wants to improve instruction.

6. Does the test provide student feedback? How well can students and their parents understand test results? Most test publishers today supply every student and parent with a printout like the one in Figure 23.5. This is the kind of student and parent feedback you should look for in a standardized test.

7. Does the test show examinee appropriateness? Examinee appropriateness is the suitability of the test for the people who are

going to be tested. This criterion refers to the level of comprehension required by the test, its physical format, and the way in which students make their responses. The age or grade range of the test should not be too large; otherwise the students at the lower part of the range will find the test too hard to understand, and those at the upper levels will find it childish. Appropriateness can mean the right level of difficulty for people of a given age, grade, sex, or cultural background. The directions for taking the test or making the observations should be clear and appropriate. The visual layout of the question, the typography, and the use of white space and color should make for clarity. The same kinds of factors apply to tests designed to be presented by sound rather than in print. Here you must think about the test's timing, pacing, and ways of recording answers (speaking, writing, marking, or performing), and the necessary equipment.

8. Is the test free of obvious bias? Despite many noble attempts it is hard to create a test that is fair to every group. Hidden biases permeate standardized and teacher-made tests. But some obvious sources of bias can be eliminated by checking that the publisher has had the test reviewed by members of various cultural and ethnic groups and by people from different regions of the country. You should also examine the test items to see whether they are (a) relevant to the life experiences of the examinee, (b) direct rather than wordy, and (c) moderately stimulating. For example, in an attempt to be relevant, one test writer made a reference to acne, which is such an extremely sensitive subject for many American teenagers that is may have distracted some of them while taking the test (Popham, 1981).

9. Is the test easy to administer? Tests that can be given to large groups are much easier to use than those that can be given only to small groups or to individuals. Tests that require less time but still give reliable information are also easier to use. And you should consider how clearly the purposes and limitations of the test and the directions for administering it are stated. How much training is required to give the test? How long does it take to get ready to give the test? Can students take the test on their own, or does it have to be administered by a trained psychometrist?

10. Does the test show ethical propriety? The ethics of testing can apply to the way in which the test is administered, the content of the test, or the kinds of recommendations that can be based on the tests. Does the test involve more than a normal amount of stress (for example, very short time limits that prevent anyone from finishing it)? If the content could be offensive, insulting, or embarrassing, the test is questionable on ethical grounds. Or if it leads to recommendations, including feedback to the student, that are possibly offensive or insulting, again it rates low in ethical propriety.

11. Does the test have retest potential? Are equivalent forms of the test available so that students can be retested with them if necessary? Is there adequate evidence that the forms are truly

equivalent? This is particularly important for criterion-referenced tests. When this issue arises for norm-referenced test, it is appropriate to ask if each of the alternate forms have norms.

12. Is the cost of the test acceptable? Educational testing is expensive. Do you really want to spend; the students' time and the district's money for the information the test scores give you? If the district intends to use tests, can you get the information it needs by some other method, and use the money for another educational purpose? The cost of a test should be commensurate with the benefits derived from it.

SOURCES OF INFORMATION

Where can you get the information you need to evaluate a set of standardized tests? The sources take three forms.

The test, its manual, scoring keys, and the like.

You can write to the test's publisher and ask for a catalog. If the publisher has the kind of test you are looking for (for example, a test of third-grade reading achievement) you can write or call for a sample, or specimen, set of the test materials by following the directions given in the catalog. Most reputable test publishers require that purchasers furnish some evidence of their qualifications for using the particular kind of test they want to purchase.

Once you have the test and its manual, scoring key, tables of norms, and other information, you can apply the criteria noted above to evaluate the test. The choice requires some judgment and insight, as well as experience in teaching the subject, knowledge of the curriculum, or psychological training relevant to the test. In choosing tests, however, there really is no satisfactory substitute for your own investigation and common sense. A good strategy to use is to take the test yourself and, as you do so, make judgments item by item as to content validity and other criteria. Then score your test and interpret your score, using the norms and other materials provided. If the test requires timing and reading directions aloud to students, have another teacher administer the test for you. This, along with a careful reading of the test manual on how the test was developed and what evidence is available regarding its validity, should help you make an intelligent decision.

The literature concerning the test.

For many tests there is considerable research literature. It can often be located through the bibliography in the test manual. In addition, the many volumes of the Mental Measurements Yearbook, described below, contain exhaustive bibliographies on many tests. A four-volume reference work, Test Critiques, is also a good source of information (Keyser & Sweetland, 1984-1985). Two useful documents that describe and analyze tests for children from birth through middle elementary school are Goodwin and Driscoll (1980) and Johnson (1979).

Journals for teachers in the various subject matter fields--such as the Journal of Research in Science Teaching, The Mathematics Teacher, and the Elementary School Journal--carry articles reporting research on tests. Various bibliographic aids--ERIC's CIJE (the Current Index to Journals in Education of the Educational Resources Information Center), the Education Index, and Psychological Abstracts--also can lead you to testing literature. One journal, Educational and Psychological Measurement, has a section called "Validity Studies of Academic Achievement," in which evidence on the validity of tests is reported.

The Mental Measurements Yearbooks.

This series of volumes, originally under the editorship of Oscar K. Buros, has been appearing about every five years since the 1930s. Each volume contains experts' reviews of many standardized tests of educational achievement, general and special aptitude, temperament, and personality. The ninth yearbook, the last to appear, came out in 1985. It contains reviews of over fourteen hundred tests (see J.V. Mitchell, 1985). The reviews tend to be searching and critical. For example, Buros (1972) warned that "at least half of the tests currently on the market should never have been published." Broad reading in these volumes in the area of the test you are considering will tell you about the strengths and weaknesses of tests in the field. As noted above, the bibliographies of research on many of the tests are exhaustive, running to more than a thousand items for some of the older and more widely used tests. Carefully edited, Buro's yearbooks have become a major resource of the testing function in the United States.

Administering Standardized Tests

The term *standardized* refers in part to the way the tests are given. Unless directions are followed carefully, the results are meaningless. You should prepare the students as necessary, depending on their age and kind of test. Motivate them, but don't make them anxious about the test. And watch for signs of parent-induced anxiety. Older students who know more about what the tests mean will motivate themselves, to the extent that they want to do well in school. But they too can feel debilitating anxiety. In many cultures, where tests often assume too much importance, testing sessions can be traumatic for some students. Work with the school counselor to alleviate test anxiety when it appears.

You must see to it that students know how to take the test. Some youngsters may not be familiar with the format of a test. Others may not be used to working alone. Still others may not understand that they shouldn't move ahead in a test when each section is timed. Many children, particularly poor children and youngsters from different cultures, may need special coaching in the skills of test taking. New students in a district may also need coaching, especially if other students in the group have taken a similar test before.

It's a good idea to teach students item formats that are strange to them ("Baker is to bread as _____ is to poem," "There are

men (who, whom) many say cannot be trusted.") It was only a few years ago that most people believed that scores on the Graduate Record Examination, the Scholastic Aptitude Test, and other standardized tests could not be improved by preparation. It is now clear that an allocation of study time for developing familiarity with item types and test formats does improve student performance on standardized tests. The best estimate we have is a correlation of over .70 between the time spent coaching students and improvement on the Scholastic Aptitude Test, with a great deal of that improvement coming from just small amounts of coaching time (Messick, 1982). One commercial course for preparing students for the Scholastic Aptitude Test takes about eighteen hours and costs $500. Students who are motivated and can afford to take the course have reported as much as a 200-point gain in SAT scores (Owen, 1985). If the reports are true, there is a moral question for our society to consider: Should this kind of advantage be available primarily for the wealthy? That wealth is the issue here is evident from research that shows coaching of minority-group and low-income students can lead to dramatic test score improvements (Anastasi, 1981; Messick, 1982).

If a test has time limits for the test as a whole or for various parts, observe them exactly. If a test has unusual kinds of answer sheets, you should become familiar with them in advance so that you can warn students about them and help students handle them (but only in keeping with the directions for administration!). If students need special supplies (scratch paper, calculators, special pencils), make sure they are on hand ahead of time.

Giving a standardized test is not particulary difficult for teachers who are prepared. But it can be a nightmare for teachers (and their students) who discover too late that they are not prepared. Confusion, unclear directions, unobserved time limits, the inability to answer questions about what is (and is not) permissible can make students do much more poorly (or meaninglessly better) than they should.

So, prepare yourself. Read a copy of the test, the manual, and the directions for administering the test at least a few days in advance. You might even try a trial run with another teacher acting as a student to help you do justice to your students and to the standardized test itself.

Interpreting Standardized Tests

In norm-referenced tests, the raw score is the number of right answers--a relatively meaningless measurement. Does a score of 36 mean excellent or poor performance? What about a score of 58? You can't tell.

You learn a little more if you know how many questions were on the test or the total possible score. If a test had 60 questions, a score of 36 seems just fair and a score of 58 seems excellent. But you may be wrong. What if the test was very easy for the subject matter, grade level, and type of student being tested? In that case, even the score of 58 may not indicate very good performance. And

what if the test was very difficult? Perhaps a score of 36 indicates quite good achievement.

How can we tell whether a test was very difficult or very easy or just right? Yes, we can look at the questions, but this isn't always effective in every subject matter area. What one teacher thinks is difficult, another teacher may feel is easy. Or a question that looks hard may turn out to be easy, and vice versa.

All of what we've just said does not hold if we are talking about a criterion-referenced test. By definition, we do know what a single raw score means here. But most present-day standardized tests are norm-referenced tests. To use them, we need to understand how norms are developed and how they function.

We can put the types of norms into a hierarchy of increasing sophistication: 1) ranking; 2) percentile ranks; 3) frequency distributions, medians, and means: 4) standard scores; 5) age- or grade-level norms.

RANKING

The first step you can take to interpret a collection of miscellaneous raw scores is to put them in rank order. Suppose you had the following raw scores for thirty fourth-graders on a fifty-item spelling test used by your district: 45, 22, 49, 17, 45, 31, 27, 40, 18, 19, 40, 25, 46, 22, 41, 30, 30, 37, 22, 23, 19, 16, 39, 28, 16, 28, 31, 36, 36, 27. We know it is hard to interpret a single score by itself. But if you rank-order the scores, you will see that one student with a score of 49 had the highest score. Two other students with scores of 16 ranked lowest. You have begun to make sense of the scores. Ranking allows you to compare scores with one another and, within this class, to interpret the scores of individual students in relation to the scores of their classmates.

Percentile ranks

Now suppose you want to compare the rank of a student in a class of 30 with the rank of another student in a class of 25. You would convert the ranks of both students to what they would be in a class of 100. The student who outranked 20 classmates in the class of 30 has a *percentile rank* of 67 (30 x 100/20). The student who ranked higher than 20 classmates in the class of 25 has a percentile rank of 80 (25 x 100/20). The percentile rank of a given raw score tells what percentage of students is excelled by that raw score.

FREQUENCY DISTRIBUTIONS, MEDIANS, AND MEANS

A frequency distribution is another way to see how students scored in comparison to other students. For the 30 scores presented above a frequency distribution of grouped data would look like the one shown in Table 23.2. In the far left column the scores have been grouped by threes starting with the highest score

earned, 49, and continuing through the range including the lowest score earned, 16. In the next column a tally has been made of the number of scores in each range, followed by the number representing the total of each tally. Looking at the table you can see that there was only 1 score in the range between 47 and 49, there were 3 scores between 44 and 46, and so on. The column on the far right is of cumulative frequencies. Starting at the bottom, with the range containing the lowest scores, there are 2 scores. In the next lowest range, 17 to 19, there are 4 scores. Added to the 2 lower scores, the cumulative frequency of all scores below 19 is 6. After the column has been tallied, one group of scores at a time, the highest cumulative frequency should be the total number of scores, in this case 30.

What if we want to know the score below which half, or 50 percent, of the students fell? We would look up the cumulative frequency column until we found half the total number of scores--in this case 15. Here the 15th of the 30 scores fell into the 26 to 28-point range. Therefore 29 is the raw score below which 50 percent of the scores are located, or 29 is the 50th percentile. This score, which is higher than half and lower than half of the scores, is called the *median*. The median is a statistic that we call a *measure of central tendency*. Another measure of central tendency is the arithmetic *mean*, commonly called the *average*. We obtain the mean by adding all the scores together and dividing the total by the number of scores. In our example the sum of the 30 scores is 905. Dividing by 30, the number of scores, we find that the mean, or average, is 30.2. When scores tend to be normally distributed, the median and mean tend to be close in value, which makes sense because both statistics are estimates of where the middle is in a distribution of scores.

STANDARD SCORES

The mean is useful in many ways. For one thing it gives us a baseline from which we can measure the performance of each student. Knowing not only whether a student is above or below the mean but also how far above or below tells us a great deal about where the student stands in the group. But first we need a unit of distance that is independent of the number of questions in the test, so that it can be used to compare (a) standing in a group of one size on a test with one mean with (b) standing in a group of another size on a test with a different mean. This kind of comparison would have meaning regardless of the number of questions on the test or the size of the group. The percentile rank makes this kind of comparison possible, but it has the disadvantage of distorting units so that a small change in raw score can make a big change in percentile rank at any region in the frequency distribution where scores tend to pile up. Because scores usually pile up in the middle of a frequency distribution, as they do in a normal distribution, slight changes in raw raw score make big changes in percentile rank in this region. It would be better to have a converted score that did not impose this kind of distortion.

The solution to the problem takes the form of the *standard score.* This is a score that converts raw scores into scores that indicate a certain deviation from the mean measured in *standard deviation* units. The standard deviation is useful as a measure of the variability, or spread, of the frequency distribution. (Its computation is described in Appendix B.) It is a statistic that helps us describe, and therefore understand, the variability (or spread) in a distribution, the amounts by which the scores differ from one another. It also helps us determine the standard score so that we can compare scores in one distribution with those in another--our goal here.

The z Score

The standard score, often called z, is derived as follows:

$$z \text{ score} = \frac{X - M}{SD}$$

where X is any raw score, M is the arithmetic mean, and SD is the standard deviation. That is, a standard score, or z score, is defined as the deviation of a score from the mean in standard deviation units. The z score, which underlies the various types of standard scores in Figure 4.1, generally ranges from -3.00 to +3.00, and has a mean of 0.00.

How does it work? Suppose we want to compare a student's relative performance on each of four standardized tests. In Table 23.3 we've taken the four raw scores and converted them to z scores. It is clear that this student is well above average in algebra and French, with z scores about half a standard deviation (.50) above the mean in both subjects. And the student is below average in physics and exactly average in reading. By converting to z scores, we can compare any test score with any other test score, whatever their distribution or means. Standard scores also allow us to determine an overall average for individual students: We just add the z scores and average them.

T Scores and Stanines

To avoid the minus signs and decimals of z scores, these scores are often converted to T scores by multiplying the standard score by 10 and adding 50:

$$T \text{ score} = 10z + 50.$$

A z score of -.12 on a test becomes a T score of 49 (10 x -.1.2 + 50). And a z score of +.80 on a test becomes a T score of 58(10 x .80 + 50). The mean of the T distribution is 50, and the standard deviation is 10. Now we can use what we know about the normal distribution to interpret T scores, which generally run between 20 and 80. For example, we would get a T score of 60 for a raw score equal to 1 standard deviation above the mean (which exceeds 84 percent of the scores in a normal distribution, as shown in Figure 4.1). A T score of 70 tells us that the raw score falls 2 standard deviations above the mean (and thus exceeds 98 percent of the scores in a normal distribution). T scores of 40, 30, and 20 indicate raw scores at 1, 2, and 3 standard deviations below the

mean. In Table 23.4 we've converted to T scores the z scores listed in Table 23.3. The T scores, like the z scores, can be added together and averaged.

To interpret any standardized score, you must know its mean and standard deviation. The College Boards and Graduate Record Examinations, for example, use standard scores with a mean of 500 and a standard deviation of 100. A score of 600, then, is 1 standard deviation above the mean.

During World War II, the U.S. Aviation Psychology Program developed a converted score that has a range of 9 points. This "standard-nine" score, the *stanine*, has a mean of 5 and a standard deviation of almost 2. Each stanine is half of a standard deviation unit. So a stanine of 7 is 1 standard deviation above the mean. Table 23.5 shows the percentage of scores in a distribution that falls into each stanine and the percentile range that corresponds to each stanine. The stanine system is convenient, but it sacrifices a good deal of information in order to use a single digit to describe a range of students' scores in a way that lets us compare them to others. A student with a stanine of 4 and a student with a stanine of 5 may be at the 24th and 60th percentiles, respectively, which is a considerable distance apart.

AGE AND GRADE NORMS

For some tests it makes good sense to convert raw scores to *age norms* or *grade norms*. These represent the raw scores obtained, on the average, by students of a given age or grade level. If a representative group of nine-year-olds gets a mean raw score of 37 on a test, then anyone whose raw score on that test is 37 is said to have an age score of 9 on the test.

More common in achievement tests is something called a *grade equivalent score*. This score represents the raw scores obtained, on the average, by students at a given grade level. For example, if the average raw score of all students in the norm group in the fourth grade, first month on a test is 37 (out of 50 for example) 37 becomes the grade equivalent score 4.1. What it tells us is that the student who earns a 4.1 grade equivalent got the same raw score as the average child in the first month of the fourth grade on that particular test.

Grade equivalents can be misleading. Suppose a child in the fourth grade earns a grade equivalent score of 6.1 on a mathematics test. This tells us that her raw score is the same as the average raw score of all students in the first month of the sixth grade who took this particular test. What does this mean? This child is in the fourth grade. The test she took is a fourth-grade test: It covers fourth-grade material. The fact that she answered as many questions correctly as a sixth-grader indicates that she has an excellent grasp of the fourth-grade material. It does not mean that she has mastered sixth-grade material! Many parents and teachers misunderstand this. Parents hearing that their fourth-grade child has earned a 6.1 grade-equivalent score wonder why their child isn't in the sixth grade or, at the very least, learning sixth-grade material. In point of fact this child has probably never

been exposed to the information or skills taught in the fifth grade. And if she had to take a sixth-grade test, she would do very poorly. All this child has done is master the fourth-grade material at a level equivalent to that of the average child in the first month of the sixth-grade. This distinction is subtle, but is important. And it is a good reason for avoiding grade-equivalent scores and using percentile or standard scores in describing test results to people who are not experts in score interpretation.

SUMMARY

Standardized tests allow us to compare the performance of any individual student with that of other students. Because they are constructed by experts, they are usually more carefully designed than are teacher-made tests. They also come with important information--on test reliability and validity and on administration procedures and problems. Standard tests cannot be used for grading students or for evaluating a teacher's performance.

Schools use two kinds of standardized tests: aptitude and achievement tests. Aptitude tests give us information for student guidance and counseling. They are used to predict achievement--a source of controversy among educators, some of whom believe that teachers accept predictions as fact and act accordingly with students. Achievements tests, which may actually be better than aptitude tests as predictors of future learning, are used primarily to measure past learning.

There are many hundreds of standardized tests available. Choosing among them means asking critical questions about their validity, reliability, and appropriateness, among other things. It also means locating the information necessary to evaluate different tests.

Administration is a critical part of the standardized-testing process. Preparation is twofold here. Talk to students about how to take the test--about the rules, format, item types, and time requirements. And prepare yourself: Read the test and the directions, do a trial run, and collect any necessary materials ahead of time.

Your work isn't finished with administration. You have to be able to interpret the results. There are five different ways of looking at students' scores: ranking, percentile ranks, medians and means, standard scores (z scores, T soars, and stanines), and age- or grade-level norms. Whatever the form, remember that scores are not perfectly reliable. Any score on a norm-referenced test must be interpreted as falling within a confidence band, and any decision about a student's mastery or nonmastery of objectives is a hypothesis with a substantial probability of being wrong! All of our measurements in education are, in varying degrees, fallible estimates of an individual's true achievement and true ability.

This chapter on standardized tests and the next chapter on teacher-made tests can give you only the briefest glimpse of a highly technical field. We urge you to take a specialized course on testing and to read some of the many texts on the subject (for

example, Cronbach, 1884; Gronlund, 1985; Hopkins & Stanley, 1981).

APPENDIX B

TESTING AND STANDARDIZED TESTS IN EDUCATION

All teaching involves evaluation where teachers are required to make decisions about student performance and appropriate teaching strategies. Testing is often the heart of an evaluation program and provides the teacher with a means of quantifiable measurement to allow comparisons between a student's performance on a particular task with a set standard or with the performances of other students. Let's look at some of the possible types of tests available to a teacher.

Norm-Referenced Tests vs. Criterion-Referenced Tests

Norm-referenced tests use the test performances of other people on the same testing instrument as a basis to interpreting an individual's relative performance. These norms allow comparison among individuals and can originate from within the same class (called immediate peer norms) to those in a given state or to across the nation (distant peer norms).

In terms of application, norm-referenced measurements are particularly useful for classifying students, selecting students for fixed quota requirements, and making decisions as to how much a student has learned in comparison to others.

In contrast, criterion-referenced measurements do not use norms. Instead, an individual's ability is measured with respect to some criterion or standard already determined in advance by knowledgeable people in the field. It is only through criterion-referenced tests that information regarding whether a student has reached a specified level of achievement is available. No comparison with immediate or distant peer groups is made.

Both norm-referenced and criterion-referenced tests can be teacher-made or be standardized tests. At this time only standardized tests will be studied.

Standardized Tests

A standardized test is a test that has been given to a large representative sample of a similar population to that of the future test-takers of the standardized test so that scores on the test can be compared with those of the people in that sample. This representative sample is known as the norming sample.

Standardized tests are constructed by experts, using the technical, statistical, and research knowledge of the testing field. These tests are

given under uniform conditions and scored according to uniform procedures to ensure all the students everywhere are being administered the test under the same conditions.

There are three broad categories of standardized tests: achievement, diagnostic, and aptitude (including interest). Of these, teachers will encounter achievement and aptitude tests most frequently. Achievement tests, the most common standardized tests given to students, are meant to measure how much a student has learned in a given content area such as reading comprehension, spelling, mathematics, science and social studies.

Standardized diagnostic tests, on the other hand, are used to identify more general learning problems. These types of tests are usually administered individually to students by highly trained professionals. While achievement tests identify weaknesses in academic content areas, the goal of diagnostic tests is to identify weaknesses in the learning processes of students. Diagnostic assessment may focus on assessing a student's abilities in areas of motor skills, auditory discrimination and visual perception -abilities needed by students to receive, process and express information. Most diagnostic testing is administered at the elementary school level.

At the high school level, students are more likely to be given aptitude tests. Like the achievement test, the aptitude test measures developed abilities. However, whereas achievement tests measure abilities developed over shorter periods of time, aptitude tests measure abilities over years and predict how well a student will do in learning unfamiliar material in the future. This type of test is used in the selection of a limited number of candidates for admission to certain programs. Another type of aptitude test given in high school is the vocational aptitude and vocational interest tests.

Interpreting Standardized Tests

There are several ways of comparing a student's raw score (number of correct answers) with the norming sample of a standardized test.

The first measurement to be considered is the frequency distribution. It is simply a listing of the number of people who obtain each score or fall into a range of scores on a test. For example, on a test 15 students made these scores: 100, 97, 92, 87, 87, 80, 76, 76, 76, 70, 68, 65, 53, 50, 42. In the frequency distribution one student made a score of 100, two made scores of 87, and so on. Within a distribution of scores, measurements of central tendency can also be useful.

Measurements of central tendency are typical scores for a group of scores. There are three measures of central tendency: the mean, the median, and the mode. In the frequency distribution example above, the mean is the arithmetic average of all the scores divided by the number of students who took the test (1109/15, or 63.93). The median is the middle score in the distribution, where half the scores are larger and half are smaller. In the same example above, the median is 76. The third measure, the mode, is the score that occurs most often. In the same frequency distribution example, the mode is again 76.

While the mean, median and mode are representative scores of a group of scores, they do not indicate how the scores are distributed. The standard deviation is a measure of how the scores spread out around the mean. The larger the standard deviation, the more spread out the scores are around the mean. The smaller the standard deviation, the more the scores are clustered around the mean.

Percentile rank scores are another form of ranking used in comparing a student's raw score to that of the norming sample. A percentile rank is the percentage of those in the norming sample who scored at or below the individual's score. If a student's score is the same or better than three-quarters of the students in the norming sample, the student would score in the 75th percentile. This does not mean that the student answered 75 questions correctly or that he answered 75 percent of the questions correctly. In this case, the 75 refers to the percentage of people in the norming sample whose scores on the test were equal to or below this student's score.

Percentile ranks do have the disadvantage of distorting units so that a small change in raw score can make a large change in percentile rank. This also makes comparisons between two groups of varying sizes with differing means difficult. Standard scores, also known as z scores, eliminate this problem by converting raw scores into scores that indicate a certain deviation from the mean measured in standard deviation units. To avoid the minus signs and decimals of z scores, z scores can in turn be converted to T scores by multiplying the standard score by 10 and adding 50.

A comparison of scores can also be done according to grade level. Separate norming samples for each grade level are necessary to generate grade-equivalent scores. For example, to obtain the seventh-grade equivalent score take the average of the scores of all the seventh graders in the norming sample. That will be the grade-equivalent score for seventh graders. If a three-grader should obtain a grade-equivalent score of 7, it does not necessarily mean that he is capable of doing advanced work but might only indicate a superior mastery of material at the third-grade level.

Reliability and Validity

There are two major criteria for judging any test used: reliability and validity. A good test is both reliable and valid. Reliability refers to how consistent or stable the test performance of a student is while validity refers to the extent the test measures what it was intended to measure.

One way of determining reliability of a test is to retest the student using the same test or a parallel form of the same test (alternate-form reliability). In a norm-referenced situation, the degree to which the rank ordering of the individuals will be the same over time can be used to satisfy the requirement for stability. If the ranks obtained on both tests is exactly the same, the correlation is equal to 1, which means the test is perfectly reliable. Correlations of .80 are usually deemed sufficient evidence for reliability.

In dealing with criterion-referenced tests, reliability is concerned with the degree of dependability of the decision made as to the student's mastery or nonmastery of a specific unit. Here the alternate-form retest method can be used to confirm the decisions made.

Although the test-retest method is available, it is not always a feasible option, especially in the school environment because of time constraints. Reliability from one test administration is necessary. This type of reliability is known as split-half reliability and is calculated by comparing performance on half of the test questions with performance on the other half.

Reliability correlations are always less than 1.00 and often less than .80. The lack of precision in scores is referred to as the standard error of measurement. Scores will fluctuate from one occasion to another because of such factors as a student's health, emotional state, motivation, coordination, memory, and fatigue. The score obtained on a test is made up of true score (hypothetical average of all scores if repeated testings under ideal conditions were possible) and error score. After the standard error of measurement is calculated, a confidence band or interval around the observed score can be developed that will include a person's true score within this area. This reflects a person's true ability range.

One way to improve reliability and thereby reducing the standard error of measurement is to increase the length of the test (both for norm- and criterion-referenced tests).

A test must be reliable to be valid but reliability will not guarantee validity. Although there are various kinds of validity only three will be discussed here: content validity, construct validity, and criterion validity.

Content validity requires the test to contain questions that pertain to the material taught. If, for example the subject matter in a social studies class changes, then the test must also reflect this change of focus by testing the student's knowledge of this new subject matter.

A more difficult type of validity to understand is construct validity. It deals with the question of whether a test measures the attribute or characteristic it claims to measure. There are two ways of checking for construct validity. The first is to use correlations between an aptitude test and a criterion measure of the corresponding kind of performance. The second way is to test hypotheses about how high-scorers and low-scorers should act and if the hypotheses are proved correct, then construct validity applies.

The final type of validity to be discussed is criterion validity. It occurs when the test scores are fairly accurate predictors of outcomes. For example, does the test actually select students who will benefit more from being admitted to a special program or school? To estimate the criterion validity, a group of students need to be tested and allowed into the program in question. Scores on the selection test would be correlated with scores on some criterion measure that reflects success in the instructional program.

Testing is definitely an involved process, not only for the test developers but also for the teachers who use standardized tests. It is necessary to have some basic understanding of standardized testing and of some of the concepts related to it in order to select the appropriate type of test for each situation.

TESTING AND STANDARDIZED TESTS IN EDUCATION

All teaching involves evaluation where teachers are required to make decisions about student performance and appropriate teaching strategies. We evaluate students with tests since they give us a way of comparing students' performance to others or to a set standard. Let's look at some of the possible types of tests available to a teacher.

Norm-Referenced Tests vs. Criterion-Referenced Tests

In norm-referenced testing you use the test results of other people on the same test to compare and interpret someone's score. These norms allow comparison among individuals and can originate from within the same class (called immediate peer norms) to those in a given state or to across the nation (distant peer norms).

In terms of application, norm-referenced measurements are particularly useful for classifying students, selecting students for fixed quota requirements, and making decisions as to how much a student has learned in comparison to others.

Criterion-referenced tests don't use norms. Instead, an individual's ability is measured with respect to some criterion or standard already determined in advance by knowledgeable people in the field. It is only through criterion-referenced tests that information regarding whether a student has reached a specified level of achievement is available. No comparison with immediate or distant peer groups is made.

Both norm-referenced and criterion-referenced tests can be teacher-made or be standardized test. At this time only standardized tests will be studied.

Standardized Tests

A standardized test is a test you give a large group of the same type of people you'll be giving the test to because you'll want to compare future test scores to the sample scores. This representative sample is known as the norming sample.

It's experts with their technical, statistical, and research knowledge of the testing field who make standardized tests. These tests are given under uniform conditions and scored according to uniform procedures to ensure all the students everywhere are being administered the test under the same conditions.

We'll look at the three general categories of standardized tests: achievement, diagnostic, and aptitude (including interest). Of these,

teachers will encounter achievement and aptitude tests most frequently. Achievement tests, the most common standardized tests given to students, are meant to measure how much a student has learned in a given content area such as reading comprehension, spelling, mathematics, science and social studies.

Standardized diagnostic tests identify more general learning problems. We usually let highly trained professionals administer these tests individually to students. While achievement tests identify weaknesses in academic content areas, the goal of diagnostic tests is to identify weaknesses in the learning processes of students. Diagnostic assessment may focus on assessing a student's abilities in areas of motor skills, auditory discrimination and visual perception -abilities needed by students to receive, process and express information. Most diagnostic testing is administered at the elementary school level.

At the high school level, students are more likely to be given aptitude tests. You'll use the aptitude tests to measure how much a student has developed in his abilities. However, whereas achievement tests measure abilities developed over shorter periods of time, aptitude tests measure abilities over years and predict how well a student will do in learning unfamiliar material in the future. This type of test is used in the selection of a limited number of candidates for admission to certain programs. Another type of aptitude test given in high school is the vocational aptitude and vocational interest tests.

Interpreting Standardized Tests

There are several ways of comparing a student's raw score (number of correct answers) with the norming sample of a standardized test.

We'll look at frequency distributions first. It is simply a listing of the number of people who obtain each score or fall into a range of scores on a test. For example, on a test 15 students made these scores: 100, 97, 92, 87, 87, 80, 76, 76, 76, 70, 68, 65, 53, 50, 42. In the frequency distribution one student made a score of 100, two made scores of 87, and so on. Within a distribution of scores, measurements of central tendency can also be useful.

When you've got a group of scores, measures of central tendency give you the typical scores within the group of scores. There are three measures of central tendency: the mean, the median, and the mode. In the frequency distribution example above, the mean is the arithmetic average of all the scores divided by the number of students who took the test (1109/15, or 63.93). The median is the middle score in the distribution, where half the scores are larger and half are smaller. In the same example above, the median is 76. The third measure, the mode, is the score that occurs most often. In the same frequency distribution example, the mode is again 76.

All the measures of central tendency: the mean, median, and mode don't show us how the scores are distributed. The standard deviation is a measure of how the scores spread out around the mean. The larger the standard deviation, the more spread out the scores are around the mean. The smaller the standard deviation, the more the scores are clustered around the mean.

You can also use percentile rank scores to compare raw scores to the norming sample. A percentile rank is the percentage of those in the norming sample who scored at or below the individual's score. If a student's score is the same or better than three-quarters of the students in the norming sample, the student would score in the 75th percentile. This does not mean that the student answered 75 questions correctly or that he answered 75 percent of the questions correctly. In this case, the 75 refers to the percentage of people in the norming sample whose scores on the test were equal to or below this student's score.

Percentile ranks do have the disadvantage of distorting units so that a small change in raw score can make a large change in percentile rank. This also makes comparisons between two groups of varying sizes with differing means difficult. You can use standard scores or z scores to change raw scores into standard deviation units and they show their deviation from the mean. To avoid the minus signs and decimals of z scores, z scores can in turn be converted to T scores by multiplying the standard score by 10 and adding 50.

We're also able to compare scores according to grade level. Separate norming samples for each grade level are necessary to generate grade-equivalent scores. For example, to obtain the seventh-grade equivalent score take the average of the scores of all the seventh graders in the norming sample. That will be the grade-equivalent score for seventh graders. If a three-grader should obtain a grade-equivalent score of 7, it does not necessarily mean that he is capable of doing advanced work but might only indicate a superior mastery of material at the third-grade level.

Reliability and Validity

A good test is both reliable and valid. Reliability refers to how consistent or stable the test performance of a student is while validity refers to the extent the test measures what it was intended to measure.

You can retest a student with the same test or a similiar one (alternate-form reliability) to see how reliable the test is. In a norm-referenced situation, the degree to which the rank ordering of the individuals will be the same over time can be used to satisfy the requirement for stability. If the ranks obtained on both tests is exactly the same, the correlation is equal to 1, which means the test is perfectly reliable. Correlations of .80 are usually deemed sufficient evidence for reliability.

When we use criterion-referenced tests, we've got to check how dependable our decision about the student's mastery of a particular unit is. Here the alternate-form retest method can be used to confirm the decisions made.

Although the test-retest method is available, it is not always a feasible option, especially in the school environment because of time constraints. You need reliable results from one test. This type of reliability is known as split-half reliability and is calculated by comparing

performance on half of the test questions with performance on the other half.

Reliability correlations are always less than 1.00 and often less than .80. Standard error of measurement is what we call the lack of precision in scores. Scores will fluctuate from one occasion to another because of such factors as a student's health, emotional state, motivation, coordination, memory, and fatigue. The score obtained on a test is made up of true score (hypothetical average of all scores if repeated testings under ideal conditions were possible) and error score. After the standard error of measurement is calculated, a confidence band or interval around the observed score can be developed that will include a person's true score within this area. This reflects a person's true ability range.

One way to improve reliability and thereby reducing the standard error of measurement is to increase the length of the test (both for norm- and criterion-referenced tests).

You need a test to be reliable but this doesn't mean it's guaranteed to be valid. Although there are various kinds of validity only three will be discussed here: content validity, construct validity, and criterion validity.

If you want your test to have content validity, you've got to relate the test questions to what you've taught. If for example the subject matter in a social studies class changes, then the test must also reflect this change of focus by testing the student's knowledge of this new subject matter.

A more difficult type of validity to understand is construct validity. Here we consider whether a test measures the attribute or characteristic it claims to measure. There are two ways of checking for construct validity. The first is to use correlations between an aptitude test and a criterion measure of the corresponding kind of performance. The second way is to test hypotheses about how high-scorers and low-scorers should act and if the hypotheses are proved correct, then construct validity applies.

The final type of validity to be discussed is criterion validity. We know we've obtained it when the test scores fairly accurately predict the outcomes. For example, does the test actually select students who will benefit more from being admitted to a special program or school? To estimate the criterion validity, a group of students need to be tested and allowed into the program in question. Scores on the selection test would be correlated with scores on some criterion measure that reflects success in the instructional program.

Anyway, you now see testing is really an involved process for both test developers and teachers. It is necessary to have some basic understanding of standardized testing and of some of the concepts related to it in order to select the appropriate type of test for each situation.

TESTING AND STANDARDIZED TESTS IN EDUCATION

All teaching involves evaluation where teachers are required to make decisions about student performance and appropriate teaching strategies. Testing is often the heart of an evaluation program and provides the teacher with a means of quantifiable measurement to allow comparisons between a student's performance on a particular task with a set standard or with the performances of other students. Anyways, you can see how important it is to use tests in teaching and you need to know something about the tests available to you.

Norm-Referenced Tests vs. Criterion-Referenced Tests

Norm-referenced tests use the test performances of other people on the same testing instrument as a basis to interpreting an individual's relative performance. You're able to use these norms to compare individuals' scores and you do this when you use scores originating locally from within the same class to those across the nation (immediate vs. distant peer norms).

In terms of application, norm-referenced measurements are particularly useful for classifying students, selecting students for fixed quota requirements, and making decisions as to how much a student has learned in comparison to others.

In contrast, criterion-referenced measurements do not use norms. Instead, an individual's ability is measured with respect to some criterion or standard already determined in advance by knowledgeable people in the field. Now, it's through criterion-referenced tests you determine whether someone has achieved a specified level or not. No comparison with immediate or distant peer groups is made.

Both norm-referenced and criterion-referenced tests can be teacher-made or be standardized tests. At this time only standardized tests will be studied.

Standardized Tests

A standardized test is a test that has been given to a large representative sample of a similar population to that of the future test-takers of the standardized test so that scores on the test can be compared with those of the people in that sample. We call this representative sample the norming sample.

Standardized tests are constructed by experts, using the technical, statistical, and research knowledge of the testing field. You're to give these tests under set conditions and you mark them according to set

procedures because you want to make sure all the students everywhere take the same test exactly the same way.

There are three broad categories of standardized tests: achievement, diagnostic, and aptitude (including interest). Of these, teachers will encounter achievement and aptitude tests most frequently. You'll find the former type of test the most common of the standardized tests given to students; you can measure how much a student has learned in a given content area such as reading comprehension, spelling, mathematics, science and social studies.

Standardized diagnostic tests, on the other hand, are used to identify more general learning problems. These types of tests are usually administered individually to students by highly trained professionals. While achievement tests identify weaknesses in academic content areas, the goal of diagnostic tests is to identify weaknesses in the learning processes of students. These tests focus on how well a student hears, sees and how good his motor skills are--all abilities needed by students to receive, process and express information. Most diagnostic testing is administered at the elementary school level.

At the high school level, students are more likely to be given aptitude tests. Like the achievement test, the aptitude test measures developed abilities. However, whereas achievement tests measure abilities devekoped over shorter periods of time, aptitude tests measure abilities over years and predict how well a student will do in learning unfamiliar material in the future. We use this type of test when we've got to admit a limited number of candidates to certain programs. Another type of aptitude test given in high school is the vocational aptitude and vocational interest tests.

Interpreting Standardized Tests

There are several ways of comparing a student's raw score (number of correct answers) with the norming sample of a standardized test.

The first measurement to be considered is the frequency distribution. It is simply a listing of the number of people who obtain each score or fall into a range of scores on a test. For example, on a test 15 students made these scores: 100, 97, 92, 87, 87, 80, 76, 76, 76, 70, 68, 65, 53, 50, 42. In the frequency distribution one student made a score of 100, two made scores of 87, and so on. Within a group of scores, you're able to also use measures of central tendency so you can interpret standardized test scores.

Measurements of central tendency are typical scores for a group of scores. There are three measures of central tendency: the mean, the median, and the mode. We're able to find the mean in the frequency distribution example above when we divide the number of students you gave the test to into the total of all the scores (1109/15, or 63.93). The median is the middle score in the distribution, where half the scores are larger and half are smaller. In the same example above, the median is 76. The third measure, the mode, is the score that occurs most often. In the same frequency distribution example, the mode is again 76.

While the mean, median and mode are representative scores of a group of scores, they do not indicate how the scores are distributed. The

standard deviation is a measure of how the scores spread out around the mean. Well, this means we'll have scores spread out more around the mean when the standard deviation is larger. And we'll see clusters of scores around the mean when the standard deviation is smaller.

Percentile rank scores are another form of ranking used in comparing a student's raw score to that of the norming sample. A percentile rank is the percentage of those in the norming sample who scored at or below the individual's score. If a student's score is the same or better than three-quarters of the students in the norming sample, the student would score in the 75th percentile. This doesn't mean the student answered 75 questions correctly and it doesn't mean he answered 75 percent of the questions correctly. In this case, the 75 refers to the percentage of people in the norming sample whose scores on the test were equal to or below this student's score.

Percentile ranks do have the disadvantage of distorting units so that a small change in raw score can make a large change in percentile rank. This also makes comparisons between two groups of varying sizes with differing means difficult. Standard scores, also known as z scores, eliminate this problem by converting raw scores into scores that indicate a certain deviation from the mean measured in standard deviation units. So you don't have to worry about minus signs and decimals of z scores you can change them to T scores; and you do this by multiplying the standard score by 10 and adding 50.

A comparison of scores can also be done according to grade level. Separate norming samples for each grade level are necessary to generate grade-equivalent scores. For example, to obtain the seventh-grade equivalent score take the average of the scores of all the seventh graders in the norming sample. That will be the grade-equivalent score for seventh graders. If a three-grader obtains a grade-equivalent score of 7, it doesn't really mean he's able to do advanced work but may only mean he's got a superior mastery of material at the third-grade level.

Reliability and Validity

There are two major criteria for judging any test used: reliability and validity. A good test is both reliable and valid. You've got a reliable test when a student's performance is consistent or stable and you've got a valid test when the test measures what it's supposed to.

One way of determining reliability of a test is to retest the student using the same test or a parallel form of the same test (alternate-form reliability). In a norm-referenced situation, the degree to which the rank ordering of the individuals will be the same over time can be used to satisfy the requirement for stability. If you get exactly the same ranking order on both tests you've got a correlation of 1 and this means the test is perfectly reliable. Correlations of .80 are usually deemed sufficient evidence for reliability.

In dealing with criterion-referenced tests, reliability is concerned with the degree of dependability of the decision made as to the student's mastery or nonmastery of a specific unit. Here we use the alternate-form retest method to confirm the decisions we've made.

Although the test-retest method is available, it is not always a feasible option, especially in the school environment because of time constraints. Reliability from one test administration is necessary. We call this split-half reliability and we calculate it when we compare performance on half of the test questions with performance on the other half.

Reliability correlations are always less than 1.00 and often less than .80. The lack of precision in scores is referred to as the standard error of measurement. Scores will fluctuate from one occasion to another because of such factors as a student's health, emotional state, motivation, coordination, memory, and fatigue. The score obtained on a test is made up of true score (hypothetical average of all scores if repeated testings under ideal conditions were possible) and error score. After you figure out the standard error of measurement you're able to develop a confidence band or interval around the observed score that includes a person's true score in this area. This reflects a person's true ability range.

One way to improve reliability and thereby reducing the standard error of measurement is to increase the length of the test (both for norm- and criterion-referenced tests).

A test must be reliable to be valid but reliability will not guarantee validity. Today I'll look at only three kinds of validity: content, construct, and criterion validities.

Content validity requires the test to contain questions that pertain to the material taught. If we change the subject matter in our social studies class, we have to change the test so we can measure the students' knowledge of what we've taught them.

A more difficult type of validity to understand is construct validity. It deals with the question of whether a test measures the attribute or characteristic it claims to measure. There are two ways of checking for construct validity. The first is to use correlations between an aptitude test and a criterion measure of the corresponding kind of performance. The second way is to test hypotheses about how high-scorers and low-scorers should act. If we prove the hypotheses correct, then we've got construct validity.

The final type of validity to be discussed is criterion validity. It occurs when the test scores are fairly accurate predictors of outcomes. For example, does the test actually select students who will benefit more from being admitted to a special program or school? We're able to estimate the criterion validity when we test a group of students and let them into the program. Scores on the selection test would be correlated with scores on some criterion measure that reflects success in the instructional program.

Testing is definitely an involved process, not only for the test developers but also for the teachers who use standardized tests. It's necessary you understand about standardized testing and related concepts so you'll be able to pick the right type of test for each occasion.

APPENDIX C

ID _____
Date _____

REVIEW QUESTIONS

Please answer the following multiple choice questions with the best response.

1. Which of the following is NOT a use for a norm-referenced test?

   (a) to determine when students may move to more advanced material
   (b) to screen applicants for a limited admission program
   (c) to determine overall achievement compared to other students
   (d) to measure how well students in a particular district are doing

2. Criterion-referenced tests measure

   (a) a student's performance compared to other students
   (b) the mastery of general educational goals
   (c) the range of abilities in a large group
   (d) the mastery of specific objectives

3. Standardization of a test implies standard methods of all of the following EXCEPT

   (a) scoring the test
   (b) reporting the scores
   (c) use of the scores
   (d) administering the test

4. Since a norming sample will be used as a comparison group for all students who will later take the test, the sample should be

   (a) completely random
   (b) similar to future test-takers
   (c) limited in size
   (d) large and diverse

5. The arithmetic average is the

(a) mode
(b) mean
(c) median
(d) standard score

6. The standard deviation is a measure of

(a) how closely scores cluster around the mode
(b) the typical score for a particular group
(c) How far the scores tend to spread around the median
(d) how far the scores tend to spread from the mean

7. A percentile rank score of 50 means that the student

(a) answered half the questions correctly
(b) had 50 correct answers
(c) scored better than 50% of all the test-takers
(d) scored at the 5th grade level

8. A z-score tells

(a) the comparison between a current and a past score
(b) how a percentile rank translates into a raw score
(c) how many standard deviations from the mean a score is
(d) the achievement level equivalent of a raw score

9. T-scores are calculated

(a) from percentile ranks
(b) directly from raw scores
(c) from grade equivalents
(d) from z-scores

10. Kathy took the Miller Analogies Test on Monday and again on Friday. Her two scores differed by only one point. This may be an indication of a good level of

(a) split-half reliability
(b) true score reliability
(c) alternate form reliability
(d) test-retest reliability

11. The standard error of measurement represents

(a) how much scores could vary on retesting

(b) the true score plus the hypothetical reliability
(c) how half the test compares with the other half
(d) the numbers of errors a student is likely to make

12. One of the most effective ways to increase the reliability of a test is to

(a) keep the test brief
(b) administer the test to many students
(c) lengthen the test
(d) allow ample response time

13. On standardized tests, a difference of a few points between two scores is likely to be insignificant due to the

(a) test-retest reliability
(b) confidence interval
(c) true score theory
(d) possibility of chance

14. If a test accurately predicts what it is intended to predict, we say that there is what kind of evidence for validity?

(a) content-related
(b) construct-related
(c) criterion-related
(d) cumulative-indicated

15. The connection between validity and reliability can be best expressed by the statement that

(a) validity requires only a limited reliability
(b) validity is essentially the same as reliability
(c) validity requires and may be assured through reliability
(d) validity requires, but cannot be assured through reliability

16. Mary has scored in the top 10% among piano students under 16 years old in her area, Cass County, North Dakota. She is very happy but also a little perplexed. This norm-referenced test doesn't tell her

(a) where she needs improvement
(b) her chances of getting into the local honors music programs
(c) how her abilities compare with her peers
(d) how well her competition can play

17. Mean, median and mode are all ways of measuring

(a) standard deviation
(b) frequency distribution
(c) arithmetic average
(d) central tendency

18. The larger the standard deviation

(a) the narrower the distribution
(b) the higher the central tendency
(c) the greater the variability
(d) the lower the variability

19. Samuel, a 7th grader, got a grade equivalent score of 9.2 on a standardized vocabulary test. This means

(a) Sam scored about the same as a 9th grader in the norming sample
(b) Sam is as advanced as a 9th grader
(c) Sam's grade should be slightly better than an "A"
(d) Sam did not quite reach the 7th grade level of performance

20. Which of the following best describes "true score"?

(a) raw score
(b) raw score minus testing error
(c) hypothetical score on the student's best day
(d) hypothetical score if the test were completely valid

APPENDIX D

TESTING AND STANDARDIZED TESTS IN EDUCATION

All teaching involves evaluation where teachers are required to make decisions about student performance and appropriate teaching strategies. **We evaluate students with tests since they give us a way of comparing students' performance to others or to a set standard.** Let's look at some of the possible types of tests available to a teacher.

Norm-Referenced Tests vs. Criterion-Referenced Tests

In norm-referenced testing you use the test results of other people on the same test to compare and interpret someone's score. These norms allow comparison among individuals and can originate from within the same class (called immediate peer norms) to those in a given state or to across the nation (distant peer norms).

In terms of application, norm-referenced measurements are particularly useful for classifying students, selecting students for fixed quota requirements, and making decisions as to how much a student has learned in comparison to others.

Criterion-referenced tests don't use norms. Instead, an individual's ability is measured with respect to some criterion or standard already determined in advance by knowledgeable people in the field. It is only through criterion-referenced tests that information regarding whether a student has reached a specified level of achievement is available. No comparison with immediate or distant peer groups is made.

Both norm-referenced and criterion-referenced tests can be teacher-made or be standardized test. At this time only standardized tests will be studied.

Standardized Tests

A standardized test is a test you give a large group of the same type of people you'll be giving the test to because you'll want to compare future test scores to the sample scores. This representative sample is known as the norming sample.

It's experts with their technical, statistical, and research knowledge of the testing field who make standardized tests. These tests are given under uniform conditions and scored according to uniform procedures to ensure all the students everywhere are being administered the test under the same conditions.

We'll look at the three general categories of standardized tests: achievement, diagnostic, and aptitude (including interest). Of these, teachers will encounter achievement and aptitude tests most frequently. Achievement tests, the most common standardized tests given to students, are meant to measure how much a student has learned in a given content area such as reading comprehension, spelling, mathematics, science and social studies.

Standardized diagnostic tests identify more general learning problems. **We usually let highly trained professionals administer these tests individually to students.** While achievement tests identify weaknesses in academic content areas, the goal of diagnostic tests is to identify weaknesses in the learning processes of students. Diagnostic assessment may focus on assessing a student's abilities in areas of motor skills, auditory discrimination and visual perception -abilities needed by students to receive, process and express information. Most diagnostic testing is administered at the elementary school level.

At the high school level, students are more likely to be given aptitude tests. **You'll use the aptitude tests to measure how much a student has developed in his abilities.** However, whereas achievement tests measure abilities developed over shorter periods of time, aptitude tests measure abilities over years and predict how well a student will do in learning unfamiliar material in the future. This type of test is used in the selection of a limited number of candidates for admission to certain programs. Another type of aptitude test given in high school is the vocational aptitude and vocational interest tests.

Interpreting Standardized Tests

There are several ways of comparing a student's raw score (number of correct answers) with the norming sample of a standardized test.

We'll look at frequency distributions first. It is simply a listing of the number of people who obtain each score or fall into a range of scores on a test. For example, on a test 15 students made these scores: 100, 97, 92, 87, 87, 80, 76, 76, 76, 70, 68, 65, 53, 50, 42. In the frequency distribution one student made a score of 100, two made scores of 87, and so on. Within a distribution of scores, measurements of central tendency can also be useful.

When you've got a group of scores, measures of central tendency give you the typical scores within the group of scores. There are three measures of central tendency: the mean, the median, and the mode. In the frequency distribution example above, the mean is the arithmetic average of all the scores divided by the number of students who took the test (1109/15, or 63.93). The median is the middle score in the distribution, where half the scores are larger and half are smaller. In the same example above, the median is 76. The third measure, the mode, is the score that occurs most often. In the same frequency distribution example, the mode is again 76.

All the measures of central tendency: the mean, median, and mode don't show us how the scores are distributed. The standard deviation is a measure of how the scores spread out around the mean. The larger the standard deviation, the more spread out the scores are

around the mean. The smaller the standard deviation, the more the scores are clustered around the mean.

You can also use percentile rank scores to compare raw scores to the norming sample. A percentile rank is the percentage of those in the norming sample who scored at or below the individual's score. If a student's score is the same or better than three-quarters of the students in the norming sample, the student would score in the 75th percentile. This does not mean that the student answered 75 questions correctly or that he answered 75 percent of the questions correctly. In this case, the 75 refers to the percentage of people in the norming sample whose scores on the test were equal to or below this student's score.

Percentile ranks do have the disadvantage of distorting units so that a small change in raw score can make a large change in percentile rank. This also makes comparisons between two groups of varying sizes with differing means difficult. **You can use standard scores or z scores to change raw scores into standard deviation units and they show their deviation from the mean.** To avoid the minus signs and decimals of z scores, z scores can in turn be converted to T scores by multiplying the standard score by 10 and adding 50.

We're also able to compare scores according to grade level. Separate norming samples for each grade level are necessary to generate grade-equivalent scores. For example, to obtain the seventh-grade equivalent score take the average of the scores of all the seventh graders in the norming sample. That will be the grade-equivalent score for seventh graders. If a three-grader should obtain a grade-equivalent score of 7, it does not necessarily mean that he is capable of doing advanced work but might only indicate a superior mastery of material at the third-grade level.

Reliability and Validity

A good test is both reliable and valid. Reliability refers to how consistent or stable the test performance of a student is while validity refers to the extent the test measures what it was intended to measure.

You can retest a student with the same test or a similiar one (alternate-form reliability) to see how reliable the test is. In a norm-referenced situation, the degree to which the rank ordering of the individuals will be the same over time can be used to satisfy the requirement for stability. If the ranks obtained on both tests is exactly the same, the correlation is equal to 1, which means the test is perfectly reliable. Correlations of .80 are usually deemed sufficient evidence for reliability.

When we use criterion-referenced tests, we've got to check how dependable our decision about the student's mastery of a particular unit is. Here the alternate-form retest method can be used to confirm the decisions made.

Although the test-retest method is available, it is not always a feasible option, especially in the school environment because of time constraints. **You need reliable results from one test.** This type of

reliability is known as split-half reliability and is calculated by comparing performance on half of the test questions with performance on the other half.

Reliability correlations are always less than 1.00 and often less than .80. **Standard error of measurement is what we call the lack of precision in scores.** Scores will fluctuate from one occasion to another because of such factors as a student's health, emotional state, motivation, coordination, memory, and fatigue. The score obtained on a test is made up of true score (hypothetical average of all scores if repeated testings under ideal conditions were possible) and error score. After the standard error of measurement is calculated, a confidence band or interval around the observed score can be developed that will include a person's true score within this area. This reflects a person's true ability range.

One way to improve reliability and thereby reducing the standard error of measurement is to increase the length of the test (both for norm- and criterion-referenced tests).

You need a test to be reliable but this doesn't mean it's guaranteed to be valid. Although there are various kinds of validity only three will be discussed here: content validity, construct validity, and criterion validity.

If you want your test to have content validity, you've got to relate the test questions to what you've taught. If for example the subject matter in a social studies class changes, then the test must also reflect this change of focus by testing the student's knowledge of this new subject matter.

A more difficult type of validity to understand is construct validity. **Here we consider whether a test measures the attribute or characteristic it claims to measure.** There are two ways of checking for construct validity. The first is to use correlations between an aptitude test and a criterion measure of the corresponding kind of performance. The second way is to test hypotheses about how high-scorers and low-scorers should act and if the hypotheses are proved correct, then construct validity applies.

The final type of validity to be discussed is criterion validity. **We know we've obtained it when the test scores fairly accurately predict the outcomes.** For example, does the test actually select students who will benefit more from being admitted to a special program or school? To estimate the criterion validity, a group of students need to be tested and allowed into the program in question. Scores on the selection test would be correlated with scores on some criterion measure that reflects success in the instructional program.

Anyway, you now see testing is really an involved process for both test developers and teachers. It is necessary to have some basic understanding of standardized testing and of some of the concepts related to it in order to select the appropriate type of test for each situation.

TESTING AND STANDARDIZED TESTS IN EDUCATION

All teaching involves evaluation where teachers are required to make decisions about student performance and appropriate teaching strategies. Testing is often the heart of an evaluation program and provides the teacher with a means of quantifiable measurement to allow comparisons between a student's performance on a particular task with a set standard or with the performances of other students. **Anyways, you can see how important it is to use tests in teaching and you need to know something about the tests available to you.**

Norm-Referenced Tests vs. Criterion-Referenced Tests

Norm-referenced tests use the test performances of other people on the same testing instrument as a basis to interpreting an individual's relative performance. **You're able to use these norms to compare individuals' scores and you do this when you use scores originating locally from within the same class to those across the nation (immediate vs. distant peer norms).**

In terms of application, norm-referenced measurements are particularly useful for classifying students, selecting students for fixed quota requirements, and making decisions as to how much a student has learned in comparison to others.

In contrast, criterion-referenced measurements do not use norms. Instead, an individual's ability is measured with respect to some criterion or standard already determined in advance by knowledgeable people in the field. **Now, it's through criterion-referenced tests you determine whether someone has achieved a specified level or not.** No comparison with immediate or distant peer groups is made.

Both norm-referenced and criterion-referenced tests can be teacher-made or be standardized tests. At this time only standardized tests will be studied.

Standardized Tests

A standardized test is a test that has been given to a large representative sample of a similar population to that of the future test-takers of the standardized test so that scores on the test can be compared with those of the people in that sample. **We call this representative sample the norming sample.**

Standardized tests are constructed by experts, using the technical, statistical, and research knowledge of the testing field. **You're to give these tests under set conditions and you mark them according to**

set procedures because you want to make sure all the students everywhere take the same test exactly the same way.

There are three broad categories of standardized tests: achievement, diagnostic, and aptitude (including interest). Of these, teachers will encounter achievement and aptitude tests most frequently. **You'll find the former type of test the most common of the standardized tests given to students; you can measure how much a student has learned in a given content area such as reading comprehension, spelling, mathematics, science and social studies.**

Standardized diagnostic tests, on the other hand, are used to identify more general learning problems. These types of tests are usually administered individually to students by highly trained professionals. While achievement tests identify weaknesses in academic content areas, the goal of diagnostic tests is to identify weaknesses in the learning processes of students. **These tests focus on how well a student hears, sees and how good his motor skills are--all abilities needed by students to receive, process and express information.** Most diagnostic testing is administered at the elementary school level.

At the high school level, students are more likely to be given aptitude tests. Like the achievement test, the aptitude test measures developed abilities. However, whereas achievement tests measure abilities devekoped over shorter periods of time, aptitude tests measure abilities over years and predict how well a student will do in learning unfamiliar material in the future. **We use this type of test when we've got to admit a limited number of candidates to certain programs.** Another type of aptitude test given in high school is the vocational aptitude and vocational interest tests.

Interpreting Standardized Tests

There are several ways of comparing a student's raw score (number of correct answers) with the norming sample of a standardized test.

The first measurement to be considered is the frequency distribution. It is simply a listing of the number of people who obtain each score or fall into a range of scores on a test. For example, on a test 15 students made these scores: 100, 97, 92, 87, 87, 80, 76, 76, 76, 70, 68, 65, 53, 50, 42. In the frequency distribution one student made a score of 100, two made scores of 87, and so on. **Within a group of scores, you're able to also use measures of central tendency so you can interpret standardized test scores.**

Measurements of central tendency are typical scores for a group of scores. There are three measures of central tendency: the mean, the median, and the mode. **We're able to find the mean in the frequency distribution example above when we divide the number of students you gave the test to into the total of all the scores (1109/15, or 63.93). The median is the middle score in the distribution, where half the scores are larger and half are smaller.** In the same example above, the median is 76. The third measure, the mode, is the score that occurs most often. In the same frequency distribution example, the mode is again 76.

While the mean, median and mode are representative scores of a group of scores, they do not indicate how the scores are distributed. The standard deviation is a measure of how the scores spread out around the mean. **Well, this means we'll have scores spread out more around the mean when the standard deviation is larger. And we'll see clusters of scores around the mean when the standard deviation is smaller.**

Percentile rank scores are another form of ranking used in comparing a student's raw score to that of the norming sample. A percentile rank is the percentage of those in the norming sample who scored at or below the individual's score. If a student's score is the same or better than three-quarters of the students in the norming sample, the student would score in the 75th percentile. **This doesn't mean the student answered 75 questions correctly and it doesn't mean he answered 75 percent of the questions correctly.** In this case, the 75 refers to the percentage of people in the norming sample whose scores on the test were equal to or below this student's score.

Percentile ranks do have the disadvantage of distorting units so that a small change in raw score can make a large change in percentile rank. This also makes comparisons between two groups of varying sizes with differing means difficult. Standard scores, also known as z scores, eliminate this problem by converting raw scores into scores that indicate a certain deviation from the mean measured in standard deviation units. **So you don't have to worry about minus signs and decimals of z scores you can change them to T scores; and you do this by multiplying the standard score by 10 and adding 50.**

A comparison of scores can also be done according to grade level. Separate norming samples for each grade level are necessary to generate grade-equivalent scores. For example, to obtain the seventh-grade equivalent score take the average of the scores of all the seventh graders in the norming sample. That will be the grade-equivalent score for seventh graders. **If a three-grader obtains a grade-equivalent score of 7, it doesn't really mean he's able to do advanced work but may only mean he's got a superior mastery of material at the third-grade level.**

Reliability and Validity

There are two major criteria for judging any test used: reliability and validity. A good test is both reliable and valid. **You've got a reliable test when a student's performance is consistent or stable and you've got a valid test when the test measures what it's supposed to.**

One way of determining reliability of a test is to retest the student using the same test or a parallel form of the same test (alternate-form reliability). In a norm-referenced situation, the degree to which the rank ordering of the individuals will be the same over time can be used to satisfy the requirement for stability. **If you get exactly the same ranking order on both tests you've got a correlation of 1 and this means the test is perfectly reliable.** Correlations of .80 are usually deemed sufficient evidence for reliability.

In dealing with criterion-referenced tests, reliability is concerned with the degree of dependability of the decision made as to the student's mastery or nonmastery of a specific unit. **Here we use the alternate-form retest method to confirm the decisions we've made.**

Although the test-retest method is available, it is not always a feasible option, especially in the school environment because of time constraints. Reliability from one test administration is necessary. **We call this split-half reliability and we calculate it when we compare performance on half of the test questions with performance on the other half.**

Reliability correlations are always less than 1.00 and often less than .80. The lack of precision in scores is referred to as the standard error of measurement. Scores will fluctuate from one occasion to another because of such factors as a student's health, emotional state, motivation, coordination, memory, and fatigue. The score obtained on a test is made up of true score (hypothetical average of all scores if repeated testings under ideal conditions were possible) and error score. **After you figure out the standard error of measurement you're able to develop a confidence band or interval around the observed score that includes a person's true score in this area.** This reflects a person's true ability range.

One way to improve reliability and thereby reducing the standard error of measurement is to increase the length of the test (both for norm- and criterion-referenced tests).

A test must be reliable to be valid but reliability will not guarantee validity. **Today I'll look at only three kinds of validity: content, construct, and criterion validities.**

Content validity requires the test to contain questions that pertain to the material taught. **If we change the subject matter in our social studies class, we have to change the test so we can measure the students' knowledge of what we've taught them.**

A more difficult type of validity to understand is construct validity. It deals with the question of whether a test measures the attribute or characteristic it claims to measure. There are two ways of checking for construct validity. The first is to use correlations between an aptitude test and a criterion measure of the corresponding kind of performance. The second way is to test hypotheses about how high-scorers and low-scorers should act. **If we prove the hypotheses correct, then we've got construct validity.**

The final type of validity to be discussed is criterion validity. It occurs when the test scores are fairly accurate predictors of outcomes. For example, does the test actually select students who will benefit more from being admitted to a special program or school? **We're able to estimate the criterion validity when we test a group of students and let them into the program.** Scores on the selection test would be correlated with scores on some criterion measure that reflects success in the instructional program.

Testing is definitely an involved process, not only for the test developers but also for the teachers who use standardized tests. **It's necessary you understand about standardized testing and related concepts so you'll be able to pick the right type of test for each occasion.**

APPENDIX E

QUESTIONNAIRE

Please complete the following questions.

1. Age _____

2. Sex:
   Male ____ Female ____

3. What year do you expect to graduate?

   _____

4. Have you taken other college or university psychology courses?

   _____

5. Did you speak English when you started elementary school?

   Yes _____ No _____

6. Do you and at least one of your parents speak English when you converse at home?

   Yes _____ No _____

APPENDIX F

PARTICIPATION FORM

A study investigating reading comprehension of university age students is being conducted at Simon Fraser University. The subjects of the study will be drawn from first and second year students.

Participants of this study will be asked to read a passage and answer questions based on the passage read. This should take approximately 20 minutes to complete.

Participation is voluntary and strict confidentiality will be observed in reporting the data. Use of the data will be restricted to the study.

I have read the above and agree to participate in this study. I also understand that strict confidentiality will be observed in reporting the data.

_____ _____

Signature of Participant Date

APPENDIX G

SIMON FRASER UNIVERSITY

VICE-PRESIDENT, RESEARCH

BURNABY, BRITISH COLUMBIA
CANADA V5A 1S6
Telephone: (604) 291-4152
FAX: (604) 291-4860

September 16, 1991

Ms. Julia Wellinger
7389 Montecito Drive
Burnaby, B.C.
V5A 1R4

Dear Ms. Wellinger:

Re: The Effects Of Text-Type Mixing On Reading Comprehension In University Age Students

This is to advise that the above referenced application has been approved on behalf of the University Ethics Review Committee.

Sincerely,

for William Leiss, Chair
University Ethics Review
Committee

cc: R. Barrow
G. Sampson

APPENDIX H

A L L Y N & B A C O N

Simon & Schuster Education Group
160 Could Street
Needham Heights, MA 02194-2310
617-455-1250
Fax: 617-455-1220

TO: Julia Wellinger

FROM: Barbara Tsantinis, Permissions Department

DATE: October 8, 1991

Dear Ms. Wellinger:

Thank you for your inquiry regarding permission to use material from one of our publications.

As I indicated to you on the phone, Allyn and Bacon acquired the Woolfolk publication from Prentice-Hall, and as it is the case with most transfers important copyright information regarding previous editions often gets lost.

The only records I have on the TEST ITEM FILE to EDUCATIONAL PSYCHOLOGY by Anita Woolfolk apply to the most recent edition (4th.) Since I cannot reference the third edition in my files, I am only able to release a conditional permission.

You may proceed with your research and publishing objectives to use twenty questions from the third edition TEST ITEM FILE in your forthcoming Masters thesis providing that the material which appears in our publication is not credited to another source. Please check all end notes and credit lines throughout the chapter as well as those at the end of the book. Permission to use such material must be obtained from the original source.

All Allyn and Bacon material must include a credit line: author, title and edition, copyright year date©, and Allyn and Bacon as publishers.

If in the future you decide to publish your thesis, please reapply to this office for a renewal of this agreement.

Thank you for your interest in our publication. I wish you much success with your paper.

A L L Y N & B A C O N

Simon & Schuster Education Group
160 Gould Street
Needham Heights, MA 02194-2310
617-455-1250
Fax: 617-455-1220

July 7, 1993

Julia Wellinger
ATTN: Graduate Studies
Faculty of Education
Simon Fraser University
Burnaby, B.C.
Canada

Dear Ms Wellinger:

Thank you for your letter of June 27, 1993. We will be happy to grant you permission to include pages 488-521 in the appendix of your unpublished thesis for Simon Fraser University.

This permission does not extend to any material appearing in our publication with credit to another source. Permission to use such material must be obtained from the original copyright holder. Please refer to the credit lines for the appropriate sources.

If at a future date you decide to have your dissertation published, you must reapply for permission.

Thank you for your interest in Allyn and Bacon, and best wishes for a successful paper.

Sincerely:

Barbara J. Tsantinis
Permissions Specialist

Julia Wellinger
Attn: Graduate Studies
Faculty of Education
Simon Fraser University
Burnaby, B.C.
Canada
Fax (604)291-3203

June 7, 1993

Ms. Laura McCormack
Permissions Dept.
Longman Inc.
White Plains, N.Y.
U.S.A.

Dear Ms. McCormack:

I am a graduate student of Simon Fraser University, Vancouver, Canada working on a Masters thesis. The focus of my investigation is on "The Effect of Text-type Mixing on Reading Comprehension". As one part of my study I have linguistically analyzed passages from three commonly used university texts in Educational Psychology, one of them, "Applying Educational Psychology in the Classroom" (3rd edition, 1988) by Myron H. Dembo, published by Longman Inc. I would appreciate obtaining permission to include a photocopy of Part 4 of the above-mentioned text, namely Chapter 12, pages 429-469 inclusive to be included in the appendix of my thesis.

Thank you for considering my request.

Sincerely yours,

Julia Wellinger

June 8, 1993

Ms. Wellinger:

Permission granted, without fee. But please use following credit line: APPLYING EDUCATIONAL PSYCHOLOGY IN THE CLASSROOM 3/e by Myron H. Dembo. Copyright © 1988 by Longman Publishing Group.

Good luck.

Jack Adams
Rights & Contracts Department

Houghton Mifflin Company

222 Berkeley Street, Boston, Massachusetts 02116-3764
(617) 351-5000 Cable HOUGHTON

College Division
TELEX 4430255
FAX 617-351-1134

June 8, 1993

Dear Sir/Madame:

Thank you for your request to use material copyrighted by Houghton Mifflin Company in your Thesis or Dissertation. We are pleased to grant permission for you to do so, without charge. We do ask, however, that you reapply for permission should you choose to use the same material in a commercial publication.

Please use the following format for your credit line: Author/Title/Edition or Voume/Copyright © Year by Houghton Mifflin Company. Used with permission.

Sincerely,

Jill C. Conway
Rights Associate

Enclosures (includes original request)

* We would prefer that you use material from the latest edition of Gage/Berliner's text, s/c ©1992.

REFERENCES

Ackerman, John M. (1991). Reading, Writing and Knowing: The Role of Disciplinary Knowledge in Comprehension and Composing. Research in the Teaching of English, 25(2), 133-178.

Adams M.J. (1982). Models of Reading. In Jean-Francois and Walter Kintsch (Eds.) Language and Comprehension (pp. 193-206). Amsterdam: North-Holland Publishing Company.

Aijmer, Karin, and Altenberg, Bengt (Eds.). (1991). English Corpus Linguistics. New York, N.Y.: Longman Inc.

Anderson, J.R. and Bower, G.H. (1973). Human Associative Memory. Washington, D.C.: Winston.

Athey, Irene. (1983). Language Development Factors Related to Reading Development. Journal of Educational Research, 76(4), 197-203.

Atkinson, Martin, Kilby, David, and Roca, Iggy. (1982). Foundations of General Linguistics. London, U.K.: George Allen & Unwin.

Ben-Amos, Dan. (1976). Folklore Genres. Austin: University of Texas Press.

Berns, Margie. (1990). Contexts of Competence: Social and Cultural Considerations in Communicative Language Teaching. New York, N.Y.: Plenum Press.

Besnier, N. (1988). The Linguistic Relationships of Spoken and Written Nukulaelae Registers. Language, 64(4), 707-736.

Biber, Douglas. (1988). Variations across Speech and Writing. Cambridge, Great Britain: Cambride University Press.

Biber, Douglas. (1989). A Typology of English Texts. In Linguistics, 27(1), 3-43.

Biber, D. and Finegan, E. (1989). Drift and the Evolution of English Style - A History of 3 Genres. Language, 65(3), 487-517.

Butler, Christopher S. (1985). Systemic Linguistics: Theory and Applications. London, England: Batsford Academic and Educational.

Chafe, Wallace L. (1982). Integration and Involvement in Speaking, Writing, and Oral Literature. In D. Tannen (Ed.) Spoken and Written Language: Exploring Orality and Literacy (pp. 35-54). Norwood, N.J.: Ablex.

Chafe, Wallace L., & Danielicz. (1987). Properties of Spoken and Written Language. In Rosalind Horowitz & S. Jay Samuels (Eds.) Comprehending Oral and Written Language (pp.83-113). San Diego, California: Academic Press, Inc.

Chalon, Yves. (1985). Reading and Communicative Competence. In Philip Riley (Ed.) Discourse and Learning (pp.67-73). Essex, England: Longman Group Limited.

Chaplin, Miriam T. (1982). Rosenblatt Revisited: The Transaction Between Reader and Text. Journal of Reading, 26(2), 150-154.

Chapman L. John (Ed.). (1981). The Reader and the Text. London, England: Heinemann Educational Books Ltd.

Clark, Virginia P., Eschholz, Paul A., and Rosa, Alfred F. (1981). Language: Introductory Readings (3rd ed.). New York: St. Martin's Press.

Couture, Barbara. (1986). Functional Approaches to Writing Research Perspectives. London, U.K.: Frances Pinter (Publishers) Ltd.

Cox, B.E., Shanahan T., and Tinzmann, M.B. (1991). Children's Knowledge of Organization, Cohesion, and Voice in Written Exposition. Research in the Teaching of English, 25(2), 179-218.

Crystal, David. (1987). Cambridge Encyclopedia of Language. Cambridge, U.K.: Cambridge University Press.

Culler, Jonathan. (1975). Structural Poetics. Ithaca: Cornell UP.

Cummings, Katherine. (1987). Test Item File Educational Psychology Anita E. Woolfolk (3rd ed.). Englewood CLiffs, N.J.: Prentice Hall.

D'Angelo, Frank J. (1986). The Topic Sentence Revisited. College Composition and Communication, 37(4), 431-441.

Dembo, Myron H. (1988). Applying Educational Psychology in the Classroom. (3rd ed.) White Plains, N.Y.: Longman Inc.

DeStefano, Johanna S. (1972). Register: Social Variation in Language Use. Elementary School Journal, 72(4), 189-194.

Dixon, John. (1987). The Question of Genres. In Ian Reid (Ed.) The Place of Genre in Learning: Current Debates (pp. 9-21). Deakin University, Australia: Centre for Studies in Literary Education.

Dubin, Fraida, Eskey, David E., and Grabe, William (Eds.) (1986). Teaching Second Language Reading for Academic Purposes. Reading, Massachusetts: Addison-Wesley Publishing Company.

Eckhoff B. (1983). How Reading Affects Children's Writing. Language Arts, 60, 607-616.

Edwards, Derek, and Mercer, Neil. (1987). Common Knowledge. London: Methuen & Co. Ltd.

Ellegard, Alvar. (1978). The Syntactic Structure of English Texts. Goteborg, Sweden: acta Universitatis Gothoburgensis.

Elson, Nicholas. (1984). Reading and Meaning for University Level ESL Students. TESL Talk, 15(3), 26-34.

Englert, Carol Sue. (1984). Children's Developing Awareness of Text Structures in Expository Materials. Journal of Educational Psychology, 76(1), 65-74.

Ferguson, Charles A. (1983). Sports Announcer Talk: Syntactic Aspects of Register Variation. Language in Society, 12(2), 153-172.

Ferrara, K., Brunner, H., and Whittemore, G. (1991). Interactive Written Discourse as an Emergent Register. Written Communication, 8(1), 8-34.

Frow, John. (1980). Discourse Genres. The Journal of Literary Semantics, 9, 73-79.

Gage, N.L., and Berliner, David C. (1988). Educational Psychology. (4th ed.) Boston, Mass.: Houghton Mifflin.

Garside, Roger. (1987). The CLAWS Word-tagging System. In Roger Garside, Geoffrey Leech, and Geoffrey Sampson (Eds) The Computational Analysis of English (pp. 30-41). New York, N.Y.: Longman Inc.

Garside, Roger, Leech, Geoffrey, and Sampson, Geoffrey (Eds.). (1987), The Computational Analysis of English. New York, N.Y.: Longman Inc.

Gates, Rosemary L. (1989). Coherence and Contextuality: Linguistic Features of Register in the Text. Freshman English News 18(1), 12-19.

Ghadessy, Mohsen (Ed.). (1988). Registers of Written English: Situational Factors and Linguistic Features. London, England: Pinter Publishers Ltd.

Gibbons, Jean Dickinson, and Chakraborti, Subhabrata. (1992). Nonparametric Statistical Inference (3rd ed.). New York, N.Y.: Marcel Dekker Inc.

Goodenough, Ward H. (1981). Culture, Language, and Society. Menlo Park, California: Benjamin/Cummings Publishing Company, Inc.

Gordon, Christine J. (1990) Contexts for Expository Text Structure Use. Reading Research and Instruction, 29(2), 55-72.

Graesser, Arthur C. (1981). Prose Comprehension Beyond the Word. New York, N.Y.: Springer-Verlag.

Greenbaum, Sidney. (1975). Language Variation and Acceptability. TESOL Quarterly 9(2), 165-172.

Gremmo, M.J. (1985). Learning a Language -- or Learning to Read? In Philip Riley (Ed.) Discourse and Learning (pp. 74-90). Essex, England: Longman Group Limited.

Gunderson, Doris V. (1970). Language & Reading: An Interdisciplinary Approach. Washington, D.C.: Center for Applied Linguistics.

Halliday, M.A.K. (1974). Language and Social Man. London: Longman

Halliday, M.A.K. (1978). Language as Social Semiotic: the Social Interpretation of Language and Meaning. London: Edward Arnold.

Halliday, M.A.K. (1985). Introduction to Functional Grammar. London: Arnold.

Halliday, M.A.K. (1987). Spoken and Written Modes of Meaning. In Rosalind Horowitz & S.Jay Samuels (Eds.) Comprehending Oral and Written Language (pp. 55-82). San Diego, California: Academic Press, Inc.

Harmer, Jeremy. (1991). The Practice of English Language Teaching. Essex, England: Longman Group UK Limited.

Hasan, Ruqaiya & Martin, J.R. (Eds.). (1989). Advances in Discourse Processes: Vol. 27. Language Development: Learning Culture: Meaning and Choice In Language: Studies for Michael Halliday. Norwood, N.J.: Ablex Publishing Corporation.

Hess, Carla W., Haug, Holly T., and Landry, Richard G. (1989). The Reliability of Type-Token Ratios for the Oral Language of School Age Children. Journal of Speech and Hearing Research, 32. 536-540.

Horowitz, Rosalind. (1987). Rhetorical Structure in Discourse Processing. In Rosalind Horowitz & S.Jay Samuels (Eds.) Comprehending Oral and Written Language. San Diego, California: Academic Press, Inc.

Hudson, Richard. (1990). English Word Grammar. Cambridge, Massachusetts: Basil Blackwell, Inc.

Jones, Janet, Gollin, Sandra, Drury, Helen and Economou, Dorothy. (1989). Systemic-Functional Linguistics and its Application to the TESOL Curriculum. In Hasan, Ruqaiya & Martin, J.R. (Eds.) Advances in Discourse Processes: Vol. 27. Language Development: Learning Culture: Meaning and Choice In Language: Studies for Michael Halliday (pp. 257-328). Norwood, N.J.: Ablex Publishing Corporation.

Kent, Carolyn E. (1984). A Linguist Compares Narrative and Expository Prose. Journal of Reading, 232-236.

Kintsch, Walter. (1982). Role of Rhetorical Structure in Text Comprehension. Journal of Educational Psychology, 74(6), 828-834.

Kintsch, W., Kozminsky, E., Streby, W.J., McKoon, G., and Keenan, J.M. (1975). Comprehension and Recall of Text as a Function of a Content Variable. Journal of Verbal Learning and Verbal Behavior, 14, 196-214.

Kintsch, W., and Keenan, J. (1973). Reading Rate and Retention as a Function of the Number of Propositions in the Base Structure of Sentences. Cognitive Psychology, 5, 257-274.

Kintsch, W. (1974). The Representation of Meaning in Memory. Hillsdale, N.J.: Erlbaum Associates.

Kress, Gunther. (1988). Language as Social Practice. In Gunther Kress (Ed.) Communication and Culture: An Introduction (pp. 79-129). Kensignton, Australia: New South Wales University Press.

Kress, Gunther. (1987). Genre in a Social Theory of Language: A Reply to John Dixon. In Ian Reid (Ed.) The Place of Genre in Learning: Current Debates (pp. 35-45). Deakin University, Australia: Centre for the Studies in Literary Education.

Kress G.R. (Ed.). (1976). Halliday: System and Function in Language. Oxford: Oxford University Press.

Langer, Judith A. (1990). The Process of Understanding: Reading for Literary and Informative Purposes. Research in the Teaching of English, 24(3), 229-260.

Le Ny, Jean-Francois and Kintsch, Walter. (1982). Language and Comprehension. Amsterdam: North-Holland Publishing Company.

Leech, Geoffrey. (1987). In Roger Garside, Geoffrey Leech, and Geoffrey Sampson (Eds.) The Computational Analysis of English (pp. 8-29). New York, N.Y.: Longman Inc.

Lemke, Jay. (1988). Genres, Semantics, and Classroom Education. Linguistics and Education, 1(1), 81-99.

Lemke, Jay. (1985). Ideology, Intertextuality, and the Notion of Register. In Benson, James D., & Greaves, William S. (Eds.) Systemic Perspectives on Discourse, Vol. 1: Selected Theoretical Papers from the 9th International Systemic Workshop (pp. 275-294). Norwood, N.J.: Ablex Publishing Corporation.

Lewy, Arieh. (1977). Planning the School Curriculum. Paris, France: United Nations Educational, Scientific and Cultural Organization.

Littlefair, Alison B. (1989). Register Awareness: An Important Factor in Children's Continuing Reading Development. In Reading, 23(2), 56-61.

Littlefair, Alison. (1991). Reading All Types of Writing: The Importance of Genre and Register. Philadelphia: Open University Press.

Longacre, Robert E. (1989). Two Hypotheses Regarding Text Generation and Analysis. Discourse Processes, 12(4), 413-460.

Longacre, Robert E. (1976). An Anatomy of Speech Notions. Lisse, The Netherlands: Peter de Ridder Press.

Love, Alison. (1991). Process and Product in Geology: An Investigation of Some Discourse Features of Two Introductory Textbooks. English for Specific Purposes (10) 89-109.

Macaulay, Marcia I. (1990). Processing Varieties in English. Vancouver, B.C.: Univeristy of British Columbia Press.

Maggart, Zelda R., and Zint, Miles V. (1992) The Reading Process (6th ed.). Dubuque, Iowa: Brown & Benchmark.

Mann, William C. (1985). The Anatomy of a Systemic Choice. Discourse Processes, 8(1), 53-74.

Marr, Mary Beth and Gormley, Kathleen. (1982). Children's Recall of Familiar and Unfamiliar Text. Reading Research Quarterly, 18(1), 89-104.

Marsh, George, Friedman, Morton, Welsch, Veronica, & Desberg, Peter. (1981). A Cognitive-Developmental Theory of Reading Acquisition. In G.E. Mackinnon and T.Gary Waller (Eds.) Reading Research: Advances in Theory and Practice, Vol. 3 (pp. 199-221). New York, N.Y.: Academic Press.

Marshall, Ian. (1987). In Roger Garside, Geoffrey Leech, and Geoffrey Sampson (Eds.) The Computational Analysis of English (pp. 42-56) New York, N.Y.: Longman Inc.

Marshall, Stewart. (1991). A Genre-Based Approach to the Teaching of Report-Writing. English for Specific Purposes, 10, 3-13.

Martin, J.R. (1982). Process and Text: Two Aspects of Human Semiosis. In Benson, James D., & Greaves, William S. (Eds.) Systemic Perspectives on Discourse, Vol. 1: Selected Theoretical Papers from the 9th International Systemic Workshop (pp. 275-294). Norwood, N.J.: Ablex Publishing Corporation.

Martin, James R. and Joan Rothery. (1986). What a Functional Approach to the Writing Task can Show Teachers about 'good writing'. In Barbara Couture (Ed.) Functional Approaches to Writing: Research Perspectives (pp. 241-265). Norwood, N.J.: Ablex.

McCarthy, Michael. (1991). Discourse Analysis for Language Teachers. Cambridge, U.K.: Press Syndicate of the University of Cambridge.

Melrose, Robin. (1988). Systemic Linguistics and the communicative Language Syllabus. In Robin P. Fawcett & David Young (Eds.) New Developments in Systemic Linguistics: Vol. 2. Theory and Application (pp. 78-93). London, U.K.: Pinter Publishers Ltd.

Miller, Carolyn R. (1984). Genre as Social Action. Quarterly Journal of Speech, 70, 151-67.

Mitchell, D.C. (1982). The Process of Reading: A Cognitive Analysis of Fluent Reading and Learning to Read. New York, N.Y.: John Wiley & Sons.

Munby, John. (1978). Communicative Syllabus Design. Cambridge, Great Britain: Cambridge University Press.

Nystrand, Martin. (1986). The Structure of Written Communication: Studies in Reciprocity Between Writers and Readers. Orlando, Florida: Academic Press.

Ong, Walter, J. (1982). Orality and Literacy. New York, N.Y.: Methuen and Co.

Painter, Claire. (1989). Learning Language: A Functional View of Language Development. In Ruqaiya Hasan and J.R. Martin (Eds.) Advances in Discourse Processes: Vol. 27: Language Development: Learning Language, Learning Culture: Meaning and Choice in Language: Studies for Michael Halliday (pp. 18-65). Norwood, N.J.: Ablex Publishing Corporation.

Perera, Katharine. (1984). Children's Writing and Reading: Analysing Classroom Language. London: Basil Blackwell.

Pugh, A.K. (1981). Construction and Reconstruction of Text. In Chapman, John L. (Ed.) The Reader and the Text (pp. 70-79). London, England: Heinemann Educational Books Ltd.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, & Svartvik, Jan. (1985). A Comprehensive Grammar of the English Language. London: Longman.

Ratcliff, R. and McKoon, G. (1978). Priming in Item Recognition: Evidence for the Propositional Structure of Sentences. Journal of Verbal Learning and Verbal Behavior, 17, 403-417.

Reagan, Timothy. (1984). In Defense of Prescriptivism: The Case for "Linguistic Imperialism" in the Classroom. Journal of Research and Development in Education, 17(3), 8-11.

Richards, Brian. (1987). Type/token Ratios: What Do They Really Tell Us?. Journal of Child Language, 14, 201-209.

Riley, Philip (Ed.). (1985). Discourse and Learning. New York, N.Y.: Longman Group Ltd.

Rosenblatt, Louise M. (1989). Writing and Reading: The Transactional Theory. In Jana M. Mason (Ed.) Reading and Writing Connections. Boston: Allyn and Bacon.

Rothery, Joan. (1989). Learning about Language. In Ruqaiya Hasan and J.R. Martin (Eds.) Advances in Discourse Processes: Vol. 27: Language Development: Learning Language, Learning Culture: Meaning and Choice in Language: Studies for Michael Halliday (pp.199-256). Norwood, N.J.: Ablex Publishing Corporation.

Ryan, Ellen Bouchard. (1977). Linguistic Awareness and Reading Performance Among Beginning Readers. Journal of Reading Behavior, 9(4), 399-400.

Sachs, J.S. (1974). Memory in Reading and Listening to Discourse. Memory and Cognition, 2, 95-100.

Sampson, Geoffrey. (1987). Probabilistic Models of Analysis. In Roger Garside, Geoffrey Leech, and Geoffrey Sampson (Eds.) The Computational Analysis of English (pp. 16-29) New York, N.Y.: Longman Inc.

Sankoff, David. (1976). The Dimensionality of Grammatical Variation. Language 52(1), 163-178.

Saylor, J. Galen, Alexander, William M., and Lewis, Arthur J. (1981). Curriculum Planning for Better Teaching and Learning (4th ed.). New York, N.Y.: Holt, Rinehart and Winston.

Shanahan, T. (1988). The Reading-writing Relationship: Seven Instructional Principles. The Reading Teacher, 41, 636-647.

Sinclair, John. (1991). Corpus Concordance Collocation. Oxford: Oxford University Press.

Singer, Harry and Donlan, Dan. (1980). Reading and Learning from Text. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Smith, Carl B. (1991). The Role of Different Literary Genres. Reading Teacher, 44(6), 440-441.

Smith, Micheal W. and Hillocks, George Jr. (1988). Sensible Sequencing: Developing Knowledge about Literature Text by Text. English Journal, 77(6), 44-49.

Sprent, Peter. (1989). Applied Nonparametric Statistical Methods. London: Chapman and Hall.

Stoddard, Sally. (1991). Text and Texture: Patterns of Cohesion. In Roy O. Freedle (Ed.), Advances in Discourse Process Vol. XL. Norwood, N.J.: Ablex Publishing Corporation.

Stubbs, Michael. (1980). Language and Literacy: The Sociolinguistics of Reading and Writing. London, England: Routledge and Kegan Paul.

Sullivan, Anne McCray. (1991). Basic Students, Linguistic Drift, and the Language of the Future. English Journal, 80(8), 43-47.

Swaffar, Janet, Arens, Katherine, and Byrnes, Heidi. (1991). Reading for Meaning: An Integrated Approach to Language Learning. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Swales, John. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge, Great Britain: Cambridge University Press.

Ulijn, Jan M. and Strother, Judith B. (1990). The Effect of Syntactic Simplification on Reading EST Texts as L1 and L2. In Journal of Research in Reading, 13(1), 38-54.

van Dijk, Teun A. and Kintsch. (1983). Strategies of Discourse Comprehension. New York, N.Y.: Academic Press, Inc.

Ventola, Eija. (1988). Text Analysis in Operation: a Multilevel Approach In Robin P. Fawcett & David Young (Eds.) New Developments in Systemic Linguistics: Vol. 2. Theory and Application. London, U.K.: Pinter Publishers Ltd.

Walsh, V. (1982). Reading Scientific Texts in English. System, 10(3), 231-239.

Warham, Sylvia. (1981). Discourse and Text: A Linguistic Perspective on Reading Skills. In John L. Chapman (Ed.) The Reader and the Text (pp. 91-99). London: Heinemann Educational Books Ltd.

White, Ronald V. (1974). Communicative Competence, Registers, and Second Language Teaching. IRAL 12(2), 127-141.

Woolfolk, Anita E. (1987). Educational Psychology. (third edition) Englewood Cliffs, N.J.: Prentice Hall.