STANDARDS OF COMPETENCE:
A MULTI-SITE CASE STUDY OF SCHOOL REFORM

Mary C. Ellwein
Gene V. Glass

CSE Report No. 263

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

1986

The project presented, or reported herein, was performed pursuant to a Grant from the Office of Educational Research and Improvement/Department of Education (OERI/ED). However, the opinions expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

CHAPTER I

INTRODUCTION

> Typically, the students are passed from one teacher to another at the end of the particular academic term, and the errors developed in the student's learning in one term are compounded with the errors made in subsequent academic terms. The errors in this system are eventually built into the student, and only rarely is he able to fully recover from them. (Bloom, 1976, p.212)

> Average achievement of high school students on most standardized tests is now lower than 26 years ago when Sputnik was launched. (National Commission on Excellence in Education, 1983, p. 8)

> Between 1975 and 1980, remedial mathematics courses in public 4-year colleges increased by 72 percent and now constitute one-quarter of all mathematics courses taught in those institutions. (Ibid, p. 9)

> . . . examinations should do much to indicate those academic illiterates who too often become teachers . . . (Vold, 1985, p. 6)

These and similar statements are examples of the rhetoric recently sounded across the nation. In comparison to the past or to the ideal, today's standards are said to be lax. According to this rhetoric, inadequate learning is the problem and testing can solve the problem. Consequently, a growing demand is heard for the use of competency tests to raise educational standards. Specifically, there are calls to use tests to make critical educational decisions such as who should graduate from high school, who should be promoted to the next grade, who should be admitted to college, and who should be certified as a teacher. As educational leaders strive to meet these demands, they eventually confront the problem of setting test standards. What scores must students

earn if they are to pass, receive a diploma, be admitted, be certified?

There are at least two ways to think about the problems of setting standards: as exercises in psychometrics or as episodes in political compromise. Measurement technologists have devised a number of methods (i.e., Nedelsky, Ebel, and Angoff) to yield the standards, test scores that separate the competent from the incompetent. However, when applied to the same test, different methods yield different scores. Which of these scores will be the standard? While academicians investigate these technological discrepancies, decision-makers grapple with political realities. Policymakers such as legislators and school boards decide by fiat that standards must be set. Decision-makers such as school district administrators are left with the task of setting the standards and administering the program. They must set standards that are reasonable and defensible. How do they do it? How are standards set in the face of disparate methods and external pressures for accountability and reform? This multi-site study sought answers to questions like these.

Research Problem

The actors and processes involved in setting educational standards were examined in this study. The research was focused on decision-makers who were charged with determining the test scores that students must earn if they are to pass through gates of the system. Grade promotion, high school

graduation, college admission and teacher certification were the system gates that provided the settings. Although the settings represented different levels of the educational system, all shared the act of determining a standard.

How are standards established? This was the major research question addressed by the study. Supporting research questions were embedded in this query and were petitioned. Why are standard-setting policies created? Who is responsible for articulating the standards, and what meanings do they attach to them? For whom, and for what purposes, are the standards intended? What problems are encountered in the standard-setting process and how are they resolved? What consequences follow from the standard-setting process?

Five sites were selected for study. These sites represented activity in local school districts, state university systems, state-level teacher certification, and state education agency (SEA) based testing programs. Data were collected from middle-level managers and other participants responsible for devising and articulating the test standards. Data were gathered by means of interviews and document collection.

Conceptual Framework

One could focus on several aspects of the standard-setting process. For example, one might choose to study the origins of the broad reforms of which competency tests and standards are merely one part. The research could examine

these origins through the eyes of the policymakers (e.g., why legislators were prompted to mandate competency testing). In contrast, consequences of the standards could be focused upon (e.g., did the imposition of standards increase achievement, increase attrition, and the like). The standard-setting process could be viewed through a narrow lens: which psychometric methods were used and which were not. The focus of this study was not found in only one of these lenses. Instead, this study concentrated primarily on the decision-makers and processes involved in setting standards in terms of test scores. These other lenses were employed to enhance and broaden the primary focus.

The purpose of a conceptual framework is to "identify the main facts and events of interest in the subject of study, as well as the main features of the context in which these facts and events are occurring" (Greene & David, 1984, p. 75). According to Miles and Huberman (1984), the conceptual framework is the foundation upon which the research questions and sampling plans are laid. A central assumption underlying this study's conceptual framework was that the participants charged with determining standards were neither high-level policymakers nor individuals for whom the standards were intended. Rather, middle-level management was responsible for producing the standards for the state or local educational agency. Further, it was assumed that these managers had discretionary powers in delegating the standard-setting task to other actors, like consultants, teachers, and the like.

Ultimately, however, managers were the ones responsible for delivering the final product to upper-level management or policymakers. Hence, their roles and actions were assumed to be central to the standard-setting process.

Given these assumptions, the study focused on middle-level managers in the agency who determined the standards. In addition, attention was paid to participants who were called into the decision-making process by these managers such as teachers and consultants.

The standard-setting process included events and actions taken by the participants. The methods employed, as well as the decisions that led to using these methods, were of interest. Intentions, as well as actions, of the participants were studied to the extent possible. Problems associated with actions and intentions sometimes occurred during the derivation of standards and examination of these lent greater insight into the decision-making process and result.

The outcome of the decision-making process, the standard, was a score attached to performance on a particular test or collection of tests. For the purpose of this study, standards expressed as increased course requirements and the like were noted and examined only when they were added as a result of the competency testing program and standard.

Consequences stemming from the testing program and standard-setting process were examined as well. Consequences to the students addressed questions like "Do minorities in the tested population fail at rates higher than other groups?".

Other outcomes included organizational changes and socio-political consequences. Organizational consequences included the implementation of tutoring programs or adjustment of the standards when actual pass rates become clear. Mobilization of minority groups to protest or support the tests and standards was one example of socio-political responses.

As the above actors, processes, and outcomes were examined, attention was directed to origins of the policy that precipitated the setting of standards. Such attention served to provide a more complete context of the situation in which the participants operated. Specifically, those responsible for initiating the standards policy were noted, as well as individuals or groups who exerted influence on these policymakers. To the extent possible, intentions and actions of these policymakers were explored. Policymakers included legislators, government officials, school boards, and upper-level administrators in the educational agency.

Theoretical Background

The primary outcome of the study was a set of detailed narrative descriptions of the testing program and standard-setting process in five settings. Nevertheless, theoretical considerations informed these five descriptions. Since the conceptual framework was focused on the standard-setting process as conducted by middle-level management, two types of bureaucratic decision-making theories were identified as potentially salient sources for working hypotheses. The two

theories, rational decision-making and bureaucratic politics, differ in their units of analysis, organizing concepts, and dominant inference patterns. Following discussions of the two theories, hypothetical portrayals of each are presented to provide contrast between the two theories.

Rational Decision-Making Theory

Rational decision-making theory extends notions of individual rationality to questions of collective decisions. Embodied in this theory is the assertion that a decision made by an individual or group is one that maximizes utility or strategic goals and objectives (Allison, 1971; Merritt & Coombs, 1977). The basic unit of analysis in this theory is bureaucratic action as choice. Rational theory serves as a framework for studying and explaining how officials choose from among several possible courses of action. In a series of reforms, the following questions could be investigated (Merritt & Coombs, p. 262):

1. Which policy alternatives were decision-makers aware of?
2. What did decision-makers think the probable consequences of the alternative were?
3. How did officials evaluate the consequences (e.g., utility) of the alternatives?

According to Allison (1971), organizing concepts in the rational decision-making theory include an actor, a problem, selection, and finally, action as rational choice. The bureaucracy as a unitary decision-maker represents the actor. It is assumed that the actor has one set of goals, perceived options, and a single estimate of each alternative's

consequences. (p. 32) A decision or action is selected in response to the problem faced by the actor and the problem's solution is defined as the sum of all activity on the part of bureaucratic members. Hence, selection describes the decision as "a steady-state choice among alternative outcomes." (p. 33) Finally, action as rational choice follows from the evaluation of goals and objectives, options, and consequences, leading to a value-maximizing choice.

The dominant inference pattern in the rational decision-making theory combines the organizing concepts described above. According to Allison, the inference pattern follows these lines: "If a group [actor] made a particular decision, that group must have had ends toward which the actions constituted a maximizing means." (p. 33)

Rational decision-making theories assume consensus on goals, options, and perceived consequences. Choices or decisions that follow emphasize efficiency and effectiveness. Technical control is seen as a vital means in achieving those ends (Collins, 1979; Karabel & Halsey, 1977), hence technical methods or solutions are preferred over less sophisticated alternatives.

Bureaucratic Politics Theory

In contrast with the rational decision-making theory, decisions or actions result from power struggles in the bureaucratic politics theory. According to this theory, a monolithic decision-maker with one set of goals, options, and estimates of consequences does not exist. Rather, groups and

individuals with different goals, alternatives, and consequences struggle with the problems at hand (Allison, 1971; Merritt & Coombs, 1977). It is assumed that groups and individuals within a bureaucracy share power and differ on what must be done. Consequently, bureaucratic decisions and actions must result from a political process (Allison, p. 145). According to Allison, action as a political result is the basic unit of analysis. He defined result as "what happens following compromise, conflict, and confusion of officials with diverse interests and unequal influence." (p. 162) Thus, the bureaucratic politics theory is useful in studying and displaying the game that yielded the decision under investigation.

Allison (1971) detailed organizing concepts around four interrelated questions as an aid to understanding the game that produced the political result (pp. 164-173):

1. Who plays?
2. What determines each player's stand or position?
3. What determines each player's impact on results?
4. How are stands, influences and moves combined to yield decision and action?

The first question is similar to the first organizing concept in rational decision-making theory as both identify the actor. From the bureaucratic politics theory, however, the actor is not an individual or group with unitary interests and goals. Rather, the actors or players are individuals in different positions within and outside of the bureaucracy, such as assistant superintendents, testing specialists,

consultants, and the like. These positions define what the players can and must do (Allison, p. 165).

Allison (1971) identified four determinants of players' stands; determinants that influence the perceptions and interests that lead to a stand (pp. 166-168):

1. Parochial priorities and perceptions influence players in their views of the problem and its solution;
2. Professional and personal goals and interests affect players' stands;
3. Stakes, what a player risks to gain or lose, undoubtedly affect the stand a player takes;
4. The deadlines accompanying problems and their resolutions affect the context of the situation, including a player's stand.

Moreover, Allison (1971) asserted that power ultimately determines a player's impact on results. He described power as a blend of bargaining advantages, skill and will in using these advantages, and other players' perceptions of the first two elements.

Players' stands, influence, and moves combine to produce bureaucratic action through "action-channels" and rules of the game (Allison, 1971, pp. 169-171). Action-channels are standardized means of taking action on a particular issue. Positions of players determine many of the action-channels. Rules of the game, implicit and explicit, may be derived largely from laws of the organization or systems within which the bureaucracy is subsumed. The players' use of action-channels and attention to rules of the game constitutes the

politics of the situation from which actions or decisions made.

The dominant inference pattern in the bureaucratic politics theory combines the organizing concepts to display and explain the game that yielded the result (Allison, 1971, p. 173). The action taken by the bureaucracy was a result of political bargaining and positioning, not a rational, value-maximizing choice selected after careful examination of options and consequences.

Vignette of a Standard-Setting Process

The board of education for the Springfield School District expressed concern about social promotion practices in the middle schools (grades 7, 8, and 9). After consulting with the district's superintendent, the board issued a mandate for the implementation of minimum competency tests in each subject area. Specifically, tests for each course in mathematics, English, social studies and science were to be developed and administered. According to the board's mandate, students who fail to earn a passing score on a particular test will be denied credit for the associated course. These students will be allowed to retake sections of the test until they earn a passing score, at which time they will receive credit for the course.

The superintendent was charged with carrying out the board's mandate. In turn, the superintendent assigned the responsibility to an assistant superintendent, under whom the Office of Research and Testing operated. The assistant

superintendent directed managers in this office to coordinate test development and to set a passing score for each test. With the first task complete, the managers turned their attention to determining the standards for performance on the tests. At this point, measurement consultants and teachers representing each subject area were engaged to participate in the standard-setting process.

Rational Decision-making Illustrated. Managers, consultants and teachers acted and made decisions as a unitary decision-making group. Implicitly or explicitly, they agreed on a goal of setting reasonable and defensible test score standards that separated students into two groups, competent and incompetent. Further, they agreed to accomplish their task in the most efficient and effective manner possible. To achieve this objective, participants considered the existing cut-off score methods, focusing on technical methods rather than the more subjective or normative procedures. The former were expected to produce scientifically-sound standards that would be less subject to criticism or dispute. The decision-making group evaluated the advantages and disadvantages of each technique in terms of time needed to implement the method, prerequisite knowledge, and available resources. After careful review, participants selected the method perceived to yield a standard in the most efficient and effective manner. Considerations of efficiency was judged, in part, by the ease involved in applying the method. An effective method was one that would not appear arbitrary and

capricious, hence defensible in court. In this case, the decision-makers concluded that the Angoff method would be used to derive standards in the most efficient and effective way.

The interactions and decisions of decision-makers were characterized by consensus. In addition to agreeing on the goal for setting the standards, participants concurred on what alternatives would be considered and the perceived consequences of each. For instance, everyone in the decision-making group agreed that the Nedelsky method would not produce high enough standards and that the Contrasting Groups method would demand too much work from the district teachers. Disagreements or conflicts were uncommon and aberrant occurrences in the decision-making group.

Bureaucratic Politics Illustrated. Goals for the standard-setting process were numerous and diverse. Even within the Office of Research and Testing, managers differed on goals. One administrator was concerned with producing the standards quickly, and to the satisfaction of the higher-level administrators and school board. A freshman staff member hoped that the standard-setting process would provide a showcase for his technical skills. The measurement consultant wished to use the experience as the basis for a publication in a scholarly journal. The teacher representing the sciences envisioned standards that would demonstrate the high expectations held for students in the science courses. And finally, the English teacher worried that high standards would

result in high fail rates which would lead to allegations of inadequate instruction.

In addition to disparate goals and expectations, the decision-makers also favored different methods and standards. The measurement consultant and younger manager contended that a combination of cut-off scores methods should be used. They argued that the technical methods were superior to less rigorous methods. "All methods are a waste of time and produce arbitrary standards," claimed the other manager. This manager maintained that counting down from 100% was the only reasonable way to derive the standards. The science teacher felt that a score of 90% for each test should be the standard for student achievement and the English teacher thought that separate standards should be set for general and advanced courses with scores of 50% and 75%, respectively.

Bargaining took place among the decision-makers and individuals exercised their powers of persuasion. The science teacher threatened to inform the press that the standards would be embarrassingly low. Managers countered that standards set too high would reflect poorly on district teachers. The consultant argued for time and opportunities to pilot the standards, but the teachers complained that pilot studies would take too much of their time and effort. The freshman manager agreed, but continued to argue for the use of multiple technical methods. Rejecting the proposal, the senior manager claimed that there was no time for such acrobatics. Ultimately, the method or standard was determined

by political compromise among the decision-makers, under less than desirable time deadlines. In this case, the decision-makers agreed to set standards according to performance on test objectives: 50% of the items in an objective must be correct to receive credit for the objective and to pass the test, students must receive credit for 75% of the objectives.

After reviewing the standards, the assistant superintendent expressed concern that the standards would not fail enough students and thus, would not satisfy the board's desire to toughen standards. The assistant superintendent directed the senior manager to change the standards in this way: pass 75% of the items per objective and pass all of the objectives.

Research Questions and Working Hypotheses

Several questions posed in the research problem will be reiterated here. The primary research question addressed the processes by which standards were established. Answers to this question were not complete without investigating several supporting questions. Who were the decision-makers? For whom, and for what purposes, were the standards intended? What problems were encountered in the standard-setting process and how were they resolved? What consequences followed from the standard-setting process?

The research questions posed in the study were grounded in the conceptual framework and supporting theoretical considerations. The conceptual framework dictated the general

boundaries of the questions while the theories guided their focus. Given the structure of the conceptual framework, primary emphasis was directed to questions about the actors, processes and consequences of setting standards. Detailed descriptions of the actors, processes, and consequences at each site were obtained by asking basic who, what, when, why, and how questions. In addition, the working hypotheses based on theories of bureaucratic decision-making served to focus these descriptions. The first four working hypotheses were based on rational decision-making theory. The last four were drawn from bureaucratic politics theory.

Working Hypotheses Based on Rational Decision-Making

1. Decision-makers use the language (e.g., goals, alternatives, consequences, and consensus) of rational decision-making to portray the standard-setting process.

2. Decision-makers describe their goal as one of selecting and applying a cut-off score method to derive test standards in the most efficient and effective manner.

3. a. Decision-makers assert that the method used to derive the standards is a deliberate and consensual choice made by participants. They portray the choice as the result of a three-part decision-making process: alternative methods were surveyed; anticipated consequences were attributed to each method; and the chosen method was the one perceived to

maximize ends of efficiency and effectiveness (i.e., a method that divided the competent from the incompetent).

b. If multiple methods were employed in deriving standards, the choice of which standard to use followed a similar decision-making process: the passing scores for a particular test were collected; anticipated consequences of applying each score were evaluated; and the score whose consequences most closely match the goal was selected as the standard.

c. For each candidate method [standard], decision-makers explain that the closer the anticipated consequences correspond to the goal, the greater was the likelihood of that method's [standard's] selection.

4. Decision-makers describe the standard-setting process as basically smooth and free of conflict. The goal, alternatives, perceived consequences, and ultimate decision are portrayed as stemming from consensus.

a. Lines of communication are hierarchical and linear.

b. Dissension among decision-makers is difficult to identify. If disagreements are detected, decision-makers report satisfaction with the eventual outcome.

Working Hypotheses Based on Bureaucratic Politics

1. Although all or some of the decision-makers use the language of rational decision-making, the standard-setting

process is best understood as a political compromise, or a series of the same, among decision-makers.

2. This process of political compromise occurs among decision-makers who have different goals, stakes, stands, and power in the situation. Once these characteristics can be identified and described for each decision-maker, their actions will be better understood. Specifically, the chronology of their actions in the standard-setting process resembles a negotiation between adversaries.

a. Decision-makers express conflicting goals and value alternative outcomes for standards.

b. Decision-makers' stakes, potential gains or losses from the standards or standard-setting process, may be actual or perceived.

c. Decision-makers stands are expressed in slogans and propaganda related to their valued outcomes.

d. Decision-makers' power or influence may be derived from their job positions, personal characteristics, or perceptions of other decision-makers.

3. Because decision-makers have disparate stands, stakes, and power, conflict and disagreements are common to the standard-setting process.

a. Lines of communication are chaotic and non-linear.

b. Dissension is identifiable in newspaper accounts, editorials, public meetings, and the like.

4. The negotiation's outcome, the test standard, is determined largely by those individuals who wielded power or

bargained most effectively.

a. Other decision-makers with less power try to subvert, compromise or delay the implementation of the standards.

b. The "winners" attempt to pacify the most vocal "losers" by assuring review of the standards, adjustments, or other actions. c. The negotiation's outcome is accompanied by additional individual, organizational and political consequences.

CHAPTER II

LITERATURE REVIEW

The literature in minimum competency testing (MCT) and standards is multitudinous. Contemporary writings surfaced in the late 1960's, increased during the 1970's, and have flourished in this decade. The bulk of published articles has addressed the controversy surrounding MCT and standards. On the one hand, Popham (1981), Berk (1980), and Hambleton (1980) championed the responsible use of MCT and standards, as well as argued for the many benefits certain to be derived from such practices. On the other hand, Glass (1978), Madaus (1981), and Wise (1978) challenged the assumptions, uses, and impact of MCT and standards. The debate between these opposing sides (moderated by many standing in the middle) has involved complex and engrossing philosophical, legal, political, practical and technical issues. The purpose of this chapter is not to reiterate these oft-cited debates, which are largely theoretical, episodic and anecdotal rather than empirical. Instead, the literature surveyed is mainly composed of empirical studies related to minimum competency testing and standard-setting. Ultimately, they are the works that must be understood before the armchair debates, recommendations, and proclamations can be fortified, applied or rejected.

Guiding this literature review was a consideration of the study's research questions:

1. For whom and what purposes are standards set?

2. Who are the decision-makers?

3. How are standards established?

4. What problems are encountered in the standard-setting process and how are they resolved?

5. What consequences follow from the standard-setting process?

Although the research questions specifically addressed facets of the standard-setting process, the literature review was not limited to these aspects. The standard-setting process is embedded in the much larger set of activities associated with MCT. Hence, confining the literature review to the standard-setting process alone would likely have produced a simplistic and incomplete picture.

A search of published and unpublished documents since the mid-1970's led to the identification of four related genera that addressed the study's questions. Accordingly, documents were grouped and reviewed by genus: documentation of MCT programs; descriptions of MCT programs and practices within particular sites; studies of standard-setting; and investigations of MCT outcomes. The first type of literature, documentation of MCT programs, was subject to change over time. For the purpose of finding sites for this investigation, only the most current literature was appropriate. Consequently, the most recent documentation available at the time of this study is described and summarized in the Chapter 3 under the section on sampling.

The other three types of MCT literature are treated in this chapter.

Descriptions of MCT Programs

Because this multi-site case study includes a substantial descriptive component, the literature was searched for works that portrayed MCT programs and practices in specific sites. Well over forty documents were located that described various aspects of MCT programs: history, test validation, administrative procedures, passing rates and the like. Eighteen states, five local school districts and three community colleges were the subjects of these descriptions, the majority of which were unpublished documents banked in ERIC, usually in the form of conference papers or agency publications. Only a handful of published descriptions was located. Smartschan (1983) described the five-year effort of a local school district in developing competency tests for high school courses. Popham, Cruse, Rankin, Sandifer, and Williams (1985) provided brief overviews of MCT programs and their results in four states and one local district. Turlington (1985) and Tirozzi, Baron, Forgione and Rindone (1985) testified to effects brought on by MCT in their states. Rosner (1982) and Stanard (1985) described teacher testing programs operating in a number of states.

These published works, as well as unpublished documents, are best characterized as testimonial articles, not as critical and systematic analyses. Invariably, these works were authored by persons associated with the particular sites.

Although informed by the authors' familiarity with the sites, the documents varied by selection of topic and detail. They provided idiographic and partial answers to questions of who, what, and when, but offered little insight into questions of how and why.

Compared to the numerous site testimonials, the number of empirically-based studies was small. Two unpublished studies systematically described and examined MCT programs in local school districts. Hayes (1981) studied five districts within a mid-western state that were said to have successful competency testing programs. Using Brickell's (1978) "Seven Key Notes" as a focusing device, Hayes sought answers to a number of questions:

1. What competencies?
2. How to measure?
3. When to measure?
4. One minimum or many?
5. How high the minimum?
6. Minimums for students or schools?
7. What to do with the incompetent?

Representatives of the five districts were interviewed and in addition, completed a questionnaire designed to address these seven questions. Tabulated responses (frequencies) were used to arrive at seven conclusions, all of which were cast in terms of what schools <u>should</u> do:

1. Test basic skills competencies.
2. Use paper and pencil tests.
3. Measure competencies during the school year.
4. Set one minimum standard.

5. Set standard to coincide with existing pass/fail rates.

6. Set minimums for students, not schools.

7. Remediate the incompetent.

Hayes concluded by offering general suggestions for implementing a MCT program:

1. Conduct a study to determine match between curriculum and district goals.

2. Poll the teachers and school community to ascertain what a high school graduate should be able to do.

3. Develop custom examinations.

Although Hayes' study was based on five different sites, the information he produced was reduced to frequencies and aggregated; no sense of individual site experiences was communicated through portrayals or illustrations.

The second study addressed factors underlying decision to adopt or reject the use of MCT as a graduation requirement (Kerins, 1982). A discriminant analysis indicated that districts using MCT as a high school graduation requirement were large, urban districts with a black student population that was at least 20% of the total. (p. 141) In addition to the quantitative analysis, interviews were conducted with superintendents and board members of ten districts to understand the history, intents, rationale and results of each program. Kerins' produced brief sketches for the ten sites. Further analysis of these sites revealed a number of themes that distinguished MCT districts from non-MCT districts:

internal advocates, minority students, curriculum, high school diploma, and perceived utility of MCT. Internal advocates were responsible for introducing the testing programs to the five MCT districts; no such individual exercised influence in the non-MCT districts. Congruent with the results of the discriminant analysis, Kerins found that minority students accounted for at least 25% of the student population in the MCT districts whereas they accounted for less than 10% in the non-MCT sites. Although MCT district representatives generally agreed that there were major curriculum problems that could be corrected by a MCT, the other type of districts disagreed. Most viewed such a test as a gross screening device and not a panacea for curriculum problems. The meaning of the high school diploma was different for these types of districts. According to MCT districts, the devalued diploma could be upgraded with a testing program. Non-MCT districts had few qualms or fears about the diploma's value and in these districts, the perceived utility was limited and suspect. In contrast, the MCT districts maintained that there were any number of benefits to be derived from MCT.

Kerins' study went further than that of Hayes' by providing a portrayal, albeit brief, of MCT in each of the ten sites. However, little or no attention was directed to the standard-setting process nor observable consequences that followed. To fill this gap, a search was undertaken for empirical studies on each of these issues with the fruits of

the search on the standard-setting process described in the following section.

Standard-Setting Investigations

The literature on standard-setting or derivation of cut-scores was partitioned into two sub-types: those works that documented procedures followed within a site and those that examined the technical or theoretical properties of cut-score methods. Studies in this second sub-type may or may not have been related to specific MCT programs.

The procedural literature addressed standard-setting issues at the state, teacher, college, and local district testing levels. The methods used in three state MCT programs were documented by Chafin (1983), Chafin and Lindheim (1982), Lockwood (1985) and Peters (1981). Bowman, Petry, Rakow, Bowyer, and Nothern (1985), Kapes and Welch (1985), and Popham and Yalow (1984) described standard-setting procedures followed on teacher competency tests. A few documents addressed standard-setting for college course placement, although none focused on setting test standards for admission purposes (Agrella & Powers, 1982; Beavers, 1983; Hector, 1984; New Jersey State Department of Higher Education, 1980; Pearse, 1982). Finally, two paper addressed local school districts' standard-setting procedures (Fillbrandt & Merz, 1977; Ziomek & Wright, 1984). By and large, these procedural documents were unpublished works authored by persons involved in the standard-setting process. Like the literature of site descriptions, these documents were factual reports of what

methods were used-- they did not address how or why methods were implemented and used in the broader context.

The second sub-type in the standard-setting literature included studies of the technical or theoretical properties of cut-score methods. Published articles were prevalent and for the most part, differed from the unpublished works only by their numbers. Consequently, the references in this section refer only to published studies.

Following Andrew and Hecht's (1976) lead, many psychometricians have investigated the relative utility, ease of implementation, error and results associated with different cut-score methods. Frequently, judgmental methods (e.g., Angoff, Ebel and Nedelsky) were compared with empirical techniques including borderline and contrasting groups (Koffler, 1980; Mills, 1983). Others confined their comparative investigations among judgmental methods (Cross, Impara, Frary & Jaeger, 1984; Halpin, Sigmon & Halpin, 1983; and Van der Linden, 1982). Another type of studies included those that addressed particular features or procedures associated with individual cut-score methods (Cross, Frary, Kelly, Small & Impara, 1985; Jaeger, 1982). Studies addressing decision-theory, Bayesian statistics, item response theory and compromise models appeared in the literature (Beuk, 1984; Cangelosi, 1984; Garcia-Quintana & Mappus, 1980; Gruijter, 1985; Huynh, 1982; Huynh & Casteel, 1985), but to a lesser extent when compared with studies of the more common judgmental or empirical methods. Considerably fewer studies

evaluated the standards produced from cut-score methods and only one of the three studies found its way to publication. Lathrop (1986) examined misclassification rates for those barely failing a MCT and made a number of suggestion to minimize such errors.

Without exception, the comparative studies showed that different cut-score methods produced different standards. In addition, equivocal results and recommendations followed from studies related to utility, implementation and error of individual methods. Although the comparative and evaluative studies were quite technical in nature, they paled in comparison with those that addressed alternative methods (e.g., decision theory and Bayesian statistics). Complex statistics and measurement models characterized these studies and offered little insight into actual applications.

Absent from all of these standard-setting studies were descriptions of how the decision-making process unfolded; what conditions precipitated the use of a particular method; and how non-technical issues accompanied the application of cut-score methods. Further, the studies reflected only technical efforts to derive standards; efforts that found approval with editors of psychometric journals or reviewers judging papers for research conferences. From the available evidence (or lack thereof), it would be inappropriate to conclude that these state-of-the-art studies accurately reflect the state of affairs in standard-setting.

Studies Related to Outcomes of MCT and Standards

Empirical studies of outcomes of MCT programs and standards formed the third class of literature reviewed in this chapter. These studies attended to overall student achievement, special group performance (e.g., minority and handicapped students), and agency changes.

General student achievement or performance was examined in two published studies. Serow, Davies, and Parramore (1982) investigated high school student performance over three administrations of a state competency test. Specifically, they examined the "extent to which pupils who initially failed the test showed improvements on subsequent reexaminations and to identify various factors which help to account for those improvements." (p.535) According to the investigators, improvement in the subtests was greatest on the first retake compared to gains between the first and second retakes. Further, students who received two to four hours of remediation on a weekly basis made greater improvements on the first retest than students who received more remediation or none at all. On the second retest, non-remedial students gained more than those who received some kind of remediation. However, the authors cautioned against strict interpretations of remediation benefits as school-to-school variations accounted more for test improvement than the measures of remediation.

In his study on the impact of school district resources and policy changes on basic skills achievement, Walstad (1984)

found that only pre-testing accounted for improvement in MCT scores. According to the regression analyses, curriculum revision and teacher training (other outcomes of the state MCT program) did not explain subsequent improvement on the basic skills test.

The performance of special groups like minority and handicapped students received attention in four published studies and one dissertation. Serow (1984) and Serow, et. al., (1982) found substantial discrepancies between black and white test performance on a state competency test. Specifically, all but a few whites had passed by the test's third administration while the number and proportion of three-time failers were much higher among blacks. Further, a pattern of failure and dropping out was more evident within the black sample (14%) than the 2% observed for whites (Serow & Davies, 1982, p.533). As a group, whites showed greater average gains on the first retest following a semester of remediation despite the authors' finding of equitable remedial opportunities for blacks and whites.

Three studies examined handicapped student performance on state competency tests. McKinney (1983) reported that 79% and 75% of the handicapped students failed the first administration of reading and math tests, respectively. By the tests' second administration, 65% and 72% still had not passed. In her dissertation examining handicapped performance on a MCT in another state, Crews (1986) observed that non-handicapped students fared better than the handicapped on each

of the MCT subtests. Among the different handicap categories, students labeled as emotionally handicapped and specific learning disabled (SLD) more closely approximated non-handicapped student performance. Finally, Hall, Griffin, Cronin and Thompson (1985) looked specifically at the performance of SLD students over one year's administration of a state MCT. They found that among the variables of IQ, locus of control, mother's education, home support, student-teacher ratio in remedial programs, and reading score on the test's first administration, only the test score correlated significantly with overall performance on the first retake. This result was congruent with Walstad's (1984) finding for overall student performance on another state MCT.

The impact of MCT programs and standards on educational agencies has been examined and reported in a few studies. In the study cited earlier by Walstad (1984), the implementation of a statewide MCT was said to precipitate changes in local district curricula, the initiation of teacher workshops and student pre-testing. Maher and Thomas (1982) surveyed students and school personnel in one state to assess the impact of the MCT and remediation practices on vocational enrollments. They found that vocational education enrollments had increased while total school enrollments had declined. Other studies of perceived agency changes were conducted by Gray (1981) and Hare (1984). Using survey and case study methods, Gray determined that the greatest impact of a state MCT was on curriculum development, followed by improved staff

communication and empathy. No changes were reported in remedial programs as a result of the two year old program. Hare looked specifically at the perceived impact of another state MCT on math curriculum. Results of his survey indicated that math objectives had changed; remedial courses were added; student achievement had increased; tutorial programs were more available; and instructional material selection had changed as a result of the MCT. Finally, Evans (1985) surveyed districts in a western state to assess the impact of the MCT on high school English requirements and curricula. In five years, the number of English credits required for graduation had not changed although the number of electives had decreased. Evans also reported changes in course descriptions and staff development. Course description had an increased emphasis on writing and paragraph construction and contained statements of course objectives. Staff development was said to experience the greatest impact from the state MCT. Workshops were created to share techniques for teaching writing and to show how to follow structured sequences of such instruction.

Summary

Three types of MCT literature have been surveyed in this chapter: descriptions of MCT programs, standard-setting investigations, and studies related to outcomes of MCT programs and standards. Aside from two empirically-based studies of MCT programs, the descriptions were testimonials that provided superficial and limited information on how such

programs operate in the educational agencies. In like manner, the standard-setting procedures within sites were documented, but were unaccompanied by contextual detail and systematic analysis. Theoretical or technical studies of cut-score methods were abundant and reflected the state-of-the-art in testing technology. However, little evidence was uncovered to indicate that such methods were used as intended in actual MCT programs. Finally, the plethora of testimonials and technical investigations contrasted sharply with the relative paucity of studies related to outcomes of MCT programs and standards. Studies related to differential test performance and perceived changes were located, but no studies were found that addressed other issues of impact and value. Further, the majority of the identified studies were in reference to a single state MCT program and many were produced in part by one author. This point is made only to illumine the fact that the outcomes and intended effects of MCT programs and standards surprisingly have not received the attention that such activities warrant. Indeed, the MCT and standards debates have continued without the benefit of sufficient grounded study and investigation. This multi-site case study will contribute substantially to the small body of empirical literature on MCT and standard-setting and should affect the hitherto data-free debates.

CHAPTER III

METHODS

This investigation is best described as a multi-site qualitative case study. Firestone and Herriott (1982) defined a multi-site qualitative study as one that addresses "the same research question in a number of settings using similar data collection and analysis procedures in each setting." (p. 63) Multi-site qualitative studies allow for understanding of sites as individuals as well as cross-site comparison. Understanding of individual sites is gained from thorough description of each bounded context. Cross-site comparison serves as the basis for establishing generalizations of the phenomenon under study (Firestone & Herriott, 1982; Miles & Huberman, 1984). In the words of Miles and Huberman:

> By comparing sites, one can establish the range of generality of a finding or explanation, and at the same time, pin down the conditions under which that finding will occur. (p. 151)

This multi-site qualitative study was partitioned into six major components. The conceptual framework served as a focusing and bounding device, a tool whose structure remained flexible throughout the course of the study. Second, research questions were derived from the conceptual framework. Both the conceptual framework and research questions were used to make sampling decisions about sites and informants. The fourth component included the methods used to generate data for the individual case studies: interviews and document collection. Individual and cross-site analyses, the fifth

component, proceeded during and following data generation for the individual sites. Interpretations of the analyses were made in the sixth component of the study. The conceptual framework and research questions were presented in chapter one. The research questions are reiterated below with succeeding sections devoted to the next three study components: sampling, data generation and analysis.

1. For whom and what purposes are standards set?
2. Who are the decision-makers?
3. How are standards established?
4. What problems are encountered in the standard-setting process and how are they resolved?
5. What consequences follow from the standard-setting process?

Sampling

Sampling concerns in this study centered on three issues: site, informant, and document selection. In the following paragraphs, the selection procedures are described for each.

Site Selection

Five sites were selected for the study. Two sites were drawn from school districts involved in competency testing for grade promotion or high school graduation. The third was chosen from colleges that used tests in admission decisions. States that employed competency tests for prospective teacher certification were represented by the fourth site and the

fifth was selected from among statewide competency testing programs.

This last type of site was not included in the original proposal to OERI. Rather, private schools were to be represented in this study. Six national organizations were contacted; organizations that represented Catholic, Protestant, and secular non-public schools:

National Catholic Education Association

Association of Christian Schools International

Council for American Private Education

National Association of Independent Schools

Friends Council of Education

Office of Christian Day Schools of the American Lutheran Church

Individuals from these organizations could not name a single school that used competency tests for purposes of reform. Although testing is a common practice in non-public school (e.g., standardized, grade level testing on an annual basis), officials were hard pressed to identify schools that used tests for promotion or graduation decisions. Given the failure to find one school from these major organizations and time constraints on site selection, the search for a non-public school was discontinued and a different type of site was included in the study. Specifically, a state-sponsored testing program was sought which would allow for comparison with a locally-sponsored elementary or secondary testing program.

The first step in the site selection process involved the identification of individual sites. Data bases compiled by Education Commission of the States (ECS) and the College Entrance Examination Board (CEEB) were used to locate sites representing the college, teacher and state levels. Similar data bases were not available for local school districts. In addition, nominations were solicited from individuals familiar with testing at four of the five levels.

After lists of potential sites were developed, the next task in the site selection process involved trimming these lists to include only those sites with characteristics congruent with the study's design and constraints. Specifically, five considerations influenced the sample trimming process:

1. Age of testing program;
2. Purpose of testing program;
3. Mandating agency;
4. Other characteristics of testing program;
5. Researcher contact with sites.

The age of the testing program was considered when eliminating sites from the initial sampling framework. Sites with very old or very new programs would pose difficulties. In sites with well-established programs, recall of the standard-setting process would likely be hampered. The further the standard-setting process was removed in time, the harder it would be for informants to reflect on the situation. Indeed, key actors in the standard-setting process might no

longer have been at the site. At the other extreme, sites currently involved in setting standards would not have had adequate opportunity to observe or deal with consequences stemming from the process. Consequently, sites between these two extremes were seen as the most desirable for this study. Hence, targeted programs were those that had been implemented between 1979 and 1984.

Sites that used tests in making decisions about students would meet the second requirement in the trimming process. Promotion or graduation decisions were of interest for local district and state testing programs. Hence, sites using competency tests for curriculum assessment or individual remediation were eliminated. Admission decisions were of interest for the college level and so, sites only using competency test results for information purposes were eliminated.

The third consideration was the mandating agency. Testing programs at the state, college, or teacher certification level may have been mandated by state Boards of Education, legislatures or through some kind of joint effort. If one mandating agency created the majority of programs within each of the state, teacher and college levels, sites from those agencies would be retained.

Other site characteristics would affect which sites would be retained. Sites that received considerable national attention or that had weathered extensive litigation were eliminated from consideration. The logistics of sites and

site access were considered as well. Districts that served large populations or state post-secondary systems with many institutions were viewed as too complex for study in the time available. Districts that served rural communities posed problems for physical access.

Researcher contact with potential sites was a final consideration in the trimming process. If either of the two researchers had contact with sites as consultants, those sites were eliminated to prevent any conflict of interest.

Once the lists were trimmed according to these considerations, at least two sites were identified for each type. Officials at the 'semi-finalist' sites were contacted by phone and the purpose of the call was explained. After gaining additional information about the testing programs, officials were asked if they would consider inclusion as a study site. Among consenting sites, choices were made to ensure that different regions of the country would be represented. When the final selections were made, the five sites were contacted again and formal permission for access was sought. In the following paragraphs, the identification and trimming processes are detailed for each type of site. The investigator's pledge to shield the study sites' identities prevents the naming of specific sites considered in the selection process and in all cases, the site names used here are pseudonyms.

For the elementary and secondary levels, seven individuals from across the country were contacted, each of

whom was known to be familiar with district testing practices. Four of these individuals were Directors of Testing in local school districts; one was a former president of the National Council of Measurement in Education (NCME); another expert held a position in a private educational research firm; and the seventh was associated with the National Association of School Boards (NASB). Each of these experts was asked to lend assistance in identifying districts that used competency tests for grade promotion at the elementary level or graduation at the high school level. Expert were simply asked to name any and all districts they had knowledge of; care was taken to ensure that the experts were not only nominating exemplar or notorious district testing programs. Ultimately, thirty-four school districts of varying size were nominated, eight of which received multiple mention. Table 3.1 shows the distribution of the thirty-four districts grouped by community population.

Table 3.1

Community Populations of Nominated School Districts

Population	Number of districts
< 100,000	18
100-199,999	2
200-299,999	2
300-399,999	1
400-499,999	2
500-599,999	1
600-699,999	1
700-799,999	1
800-899,999	1
900-999,999	1
> 1,000,000	4

Eighteen districts were located in communities with populations fewer than 100,000; eight of these districts served communities smaller than 20,000. Four nominated districts were in cities greater than one million. Twelve districts were located in cities greater than 100,000 but less than one million. Among these twelve districts, five were contacted to learn more about their testing programs.

Ultimately, two districts were selected which varied in size: Lucerne Public Schools served a community of 160,000 and Waterford School District was located in a community of over 500,000.

Four documents produced by ECS served as the primary sources for identifying states involved in competency testing. As of November, 1985, forty states had some type of competency testing programs and over 75% of these states tested both elementary and secondary students. In addition to identifying sites from ECS data bases, expert nominations were sought. A representative of a private research firm and an individual associated with a metropolitan district nominated seven states, two of which were mentioned by both experts. States were arrayed by type of testing program (e.g., elementary, secondary or both) and mandating agency (state Board of Education, legislature or joint action). Table 3.2 is a display of the numbers of state testing programs arrayed by type of program and mandating agency.

Table 3.2

Potential State Testing Programs

Type of program	Mandating Agency: State board of education	State legislature	Joint action
Elementary	0	4	0
Secondary	1	2	0
Both	16	11	6

State Boards of Education and legislatures created equal numbers of state testing programs-- each type of authority mandated seventeen programs. Joint agency efforts produced testing programs in six states. Since decidedly fewer programs were created by joint mandate, only testing programs authorized by one type of authority were eligible. These remaining sites were then arrayed by test use (e.g., promotion, graduation or other). The final list included nineteen states, seventeen of which used competency tests for high school graduation. This list was compared to the list of nominated states. One state appeared on both lists; this state tested for high school graduation and was selected along with an elementary testing program in another state for possible inclusion. Since the state with the elementary

program would not allow access until mid-1986, the secondary testing program in Victoria was selected for the study.

A slightly different procedure was followed in identifying potential sites at the college level. Information was obtained from ECS and CEEB. The former organization provided a summary of college admission requirements in the 50 states. The latter organization provided a detailed summary of state policies for college admission. From these two sources, five state university systems were identified as using competency tests as part of their admission requirements. Expert nomination yielded additional college systems and further information about sites identified from the CEEB document. Two individuals were contacted, one of whom was associated with the American Association of College Registrars and Admissions Officers. The second individual was the Director for Admissions and Guidance Services at one of CEEB's regional offices. Each expert contributed a new state to the list of five. These post-secondary institutions were examined for type (e.g., SAT, ACT, or basic skills) and purpose of test (e.g., admission or placement) as well as date instituted. Table 3.3 displays the college testing programs by these characteristics.

Table 3.3

Potential College Sites

State	Type of test	Purpose of test	Year implemented
A	SAT or ACT	Admission	1986
B	Basic skills	Admission	1981
C	Basic skills	Placement	-
D	Basic skills	Placement	-
E	ACT	Admission	1983
F	Basic skills	Placement	-
G	-	Admission	-

Note. Unreported data indicated by -.

Post-secondary institutions in states 'B' and 'E' were contacted. The latter was using the test a part of a prediction formula and no standards were set on the test itself. State 'B', Brittany, appeared to be the best choice for the study's purposes; implemented in 1981, the basic skills test was used for admissions.

States involved in testing for initial teacher certification were identified from another ECS data-base. As of November, 1985, thirty-one states had mandated such testing programs. Since all of the states' activities were well documented in this data-base, nominations were not solicited.

The thirty-one states that tested prospective teachers for certification were arrayed by mandating agency and date of program mandate (pre-1983 or post-1983). Table 3.4 shows the numbers of states sorted by these characteristics.

Table 3.4

Potential Teacher Certification Sites

	Mandating agency	
Year mandated	State board of education	State legislature
Before 1983	3	8
Since 1983	10	10

Of the programs mandated before 1983, the majority were created by state legislatures. Among these eight states, six had been created after 1979. Two of these sites were contacted and ultimately, Granada was selected for the study.

In summary, five sites were chosen: Waterford, Lucerne, Victoria, Brittany, and Granada. The trimming criteria and processes by which these sites were chosen affect the extent to which results from this study may be generalized. Specifically, extremely large or small districts, very old or new programs, those created by joint agency efforts, nor well-known sites were included in this study.

Informant Selection

At the study's outset, the sampling frame for actors and informants was necessarily incomplete. However, some initial choices were made within the boundaries determined by the conceptual framework. Actors sampled were primarily those persons directly involved with the testing program and standard-setting process. Their specific positions necessarily varied by site and are identified within each of the site narratives. Informants included those actors with direct involvement as well as individuals familiar with the history and consequences of the testing programs and standards. For parsimonious purposes, actors and informants will not be distinguished from one another in sections and chapters to come. Henceforth, references to informants include both actors well as non-actors.

Key contacts at each site were asked to provide names of informants involved in or familiar with the testing program and standard-setting process. The generated lists were compared to check for common and unique names. As individuals were contacted and interviewed, they were asked to identify others who did not appear on the lists.

Document Selection

Since site visitation time was limited, informants were relied upon to provide documents relevant to the origins and descriptions of the testing programs, the standard-setting processes, and related consequences. Such documents consisted

of publicly-distributed literature, internal reports, professional papers, published articles, newspaper articles, memos and correspondences. Informants at all five sites generously contributed pertinent documents. For two sites, additional documentation was obtained independently. These documents were published articles that appeared in local newspapers and state professional journals.

Data Generation and Analysis

Site Visits

Two visits were made to four of the five sites. Informants from the fifth site were interviewed once at home and a second time at a meeting during a professional conference. Each site visit lasted from two to four days. Before each of the initial site visits were made, appointments were scheduled with key informants and to the extent possible, other actors. Two of the sites sent relevant documents prior to the first site visit; the other three sites had documents available upon arrival. Following initial interviews, additional informants were contacted and interviews were scheduled at their convenience.

Informants were advised of the study's purpose in broad terms and the nature of the final report was described. Names would be kept confidential, but informants were warned that they may be identified by their positions within a particular site. Efforts would be made whenever possible to change the names of departments and titles to prevent outside

identification. All interviews were audio-recorded after receiving permission from each informant. Field notes were taken during these semi-structured interviews, which lasted from 20 to 90 minutes each. Since the interviews would be preserved in their entirety on cassette tape, the field notes served primarily as aids in guiding the interviews and tailoring them to the specific issues at hand. Documents were used to stimulate memories of actions, event, attitudes, and the like among informants.

Upon return from the initial site visits, the audio-taped interviews were transcribed. The transcripts and documents were read carefully and repeatedly. Preliminary coding, analysis, and interpretations were made before second-round visits were scheduled.

The purpose of the second site visits was three-fold: 1) address questions left unanswered from the first visit; 2) corroborate data provided by informants and documents; and 3) gather information that was not available earlier. During second-round visits, many of the informants were interviewed again, some new informants were contacted and additional documents were collected.

Instrumentation

An interview protocol was developed prior to the site visits. The major and supporting research questions, as well the working hypotheses, guided the protocol's development. The original interview protocol is included in Appendix A. In the following paragraphs, the nature of the information sought

from the protocol is described. Further, these sections include explanations of what types of generated data would be viewed as evidence for the competing hypotheses.

For whom and what purposes are the standards set. Answers to the major research question, "How are standards established?", could not be complete without understanding the purpose of the test and standards. Section A of the interview protocol included questions that addressed these concerns. In addition to addressing the purpose of the test and standards, this section of the protocol included questions related to the history of the testing program. Answers to these questions provided the context for understanding the standard-setting process that followed.

Who are the decision-makers. Answers to the major research question, "How are standards established?", necessitated information about the decision-makers involved in the process. Although the conceptual framework identified managers as the leading actors with other participants in supporting roles, such assumptions did not remain unsubstantiated. Questions from sections B and C of the interview protocol focused on the decision-makers: their positions, responsibilities, and how each became involved in the standard-setting process.

Questions in section D addressed the second and sixth working hypotheses posited from the rational decision-making and bureaucratic politics theories, respectively. The second

hypothesis asserted that decision-makers had at least one common goal: to select and apply a cut-off score method to derive test standards in the most efficient and effective manner. Questions that focused on decision-makers' goals and expectations in the standard-setting process informed this hypothesis. For example, "Reflecting on the time right before the standard-setting process was to begin, what did you think the testing program would accomplish? With that end in mind, what did you expect the standard-setting process to entail?". Answers to questions like these would aid in determining whether: 1) decision-makers had at least one common goal and 2) the goal resembled the one stated. If decision-makers expressed disparate goals, the second hypothesis would be discredited and the sixth would gain partial support. Additional support for the sixth hypothesis would come from answers to questions that addressed decision-makers' stakes, stands, and power.

According to the working hypotheses, the stakes held by decision-makers in the standard-setting process were influenced by their goals and interests. One way of determining decision-makers' stakes was to ask them to elaborate on how they expected the standard-setting process and results to affect their professional, organizational and personal interests. Another line of questioning included soliciting views on what other participants had to gain or lose. Stands taken by decision-makers influenced their actual or perceived stakes in the standard-setting process. Stands

would be reflected in opinions expressed by decision-makers about what they wanted to occur during the standard-setting process and how they felt about particular standards or outcomes. Catch phrases, slogans and propaganda would be evidence of stands. Comments by decision-makers on their perceptions of other's stands were solicited as corroborating evidence. The various job positions of decision-makers contributed information about the amount of formal authority held by each. Finally, decision-makers' powers were determined by perceptions of others and so, questions were included to learn about the influence wielded by others during the process and on the resulting standards.

How are standards established. The preceding paragraphs described the protocol questions about the purposes of the tests and standards as well as information about decision-makers. Section E of the interview protocol addressed the primary research question: "How do decision-makers set standards?".

The third working hypothesis provided a rational decision-making model for the standard-setting process: a value-maximizing choice made after consideration of alternatives and their perceived consequences. Probable alternatives included various methods for deriving cut-scores. Therefore, some questions dealt with decision-makers' knowledge of methods. What cut-off score methods were familiar to participants? What were the perceived consequences attached to these methods? Which method was

selected for use? How was that method viewed as a means of maximizing valued outcomes held by the decision-makers? Decision-makers may not have considered cut-score methods; rather, they may have selected the standard from a pool of possible scores. Hence, questions were included to address those scores, their perceived consequences, the standard ultimately selected, and how it maximized valued outcomes. As questions were formulated to test the rational decision-making model in the third hypothesis, queries were also developed to collect evidence for the first hypothesis: the use of rational decision-making language. References to a unitary goal, options, and consequences would be viewed as supporting evidence.

The fifth and sixth hypotheses contrasted with the rational decision-making model. A political compromise or negotiation was posited to describe the standard-setting process. Evidence of disparate goals, stakes, stands, and power would lend support to a bureaucratic politics model. As questions addressed the standard-setting process, individual goals, stakes, stands, and power were linked to events and outcomes of the standard-setting process. To support the bureaucratic politics model, respondents would have to speak of how their own or others' actions affected the process or outcomes. References to bargaining, persuasion, and use of authority were seen as further evidence for the bureaucratic politics model.

What problems are encountered in the standard-setting process and how are they resolved. In describing and explaining how standards were established, questions in sections E and F were directed to possible problems. The fourth hypothesis, based on rational decision-making, asserted that few, if any, problems occur during the standard-setting process. Supporting evidence would include descriptions of decision-makers' actions and interactions as linear and hierarchical. Moreover, respondents' would describe of the process as moving through established channels of authority and communication supported rational decision-making. If there were any conflicts, they would be rare instances and difficult to identify. Nevertheless, respondents would not express dissatisfaction with the standard-setting process or its outcomes.

In contrast with the fourth hypothesis, the seventh was based on bureaucratic politics theory and asserted that conflict would be common during the decision-making process. Respondents' accounts of conflict with other decision-makers during the standard-setting process would provide supporting evidence for this hypothesis. Respondents would describe the process as chaotic and episodic, where decision-makers struggled for power and influence. Further evidence would include accounts of factions within the decision-making group that tried to undermine the positions of others or conflicts about the resulting standards. Protocol questions addressed

the parties involved in any conflicts as well as steps taken to resolve them.

What consequences follow from the standard-setting process. The first consequence was the standard(s) derived by participants and it was addressed in section F of the protocol. Respondents were asked to describe their reactions to the standard. Expressions of satisfaction would lend support to the fourth hypothesis. However, the eighth hypothesis posited that there would be individuals and groups that expressed dissatisfactions with the standard(s). Accounts of actions meant to discredit influential decision-makers, compromise the standards, or delay implementation would be evidence for this hypothesis.

Consequences also followed from disclosure of the standard. These were addressed in the penultimate section, G. Questions were included to address possible consequences that stemmed from the disclosure of the standard to policymakers and other stakeholders. Again, if groups responded favorably, the fourth hypothesis would gain support. However, if respondents provided accounts of disagreement and conflict, the eighth hypothesis would gain support.

Finally, questions in section G addressed outcomes following the application of the standard. Failure rates in the total population and by various segments (e.g., ethnic groups) were solicited from informants. Information on changes in achievement levels and attrition rates were solicited as well. Questions in this section also dealt with

organizational consequences, such as implementation of tutoring programs, media attention, and alterations in curricula requirements.

Although the interview protocol included specific questions and probes, it was anticipated that they would not exhaust the realm of potential queries. Indeed, site-specific and other relevant questions became apparent during the course of the fieldwork. Hence, efforts were made to cover the basic spirit of the protocol while remaining sensitive to emerging issues.

Utility of the Working Hypotheses

As data generation progressed, it became apparent that the focus provided by the working hypotheses was too restrictive. The constraints of this focus was made evident in two inter-related ways: 1) the nature and use of questions that stemmed from the hypotheses, and 2) temporal constraints imposed by the study's design.

The nature and use of questions derived from the working hypotheses proved to be problematic. Many of the questions related to bureaucratic politics addressed issues that commonly carry negative connotations: self-interest, wielding of power, stakes, and conflict. Pursual of such sensitive issues was impeded by the investigator's role. Dictated by the study's design, the role of the investigator was like a stranger passing through town. As such, the investigator had access to the more neutral or objective evidence, but was prevented from exploring the more sensitive issues in-depth.

Indeed, informants alluded to issues of conflict, but they were reluctant to speak in great detail about specific individuals or situations. Several reasons may account for this closing of ranks: distrust of the investigator; informant suspicion about potential use of data; informant desire to present events in the best light possible; fear of disloyalty to the organization or profession; or the relative absence of conflict and other negative aspects. For whatever reasons, unchecked use of questions related to the working hypotheses was seen as inappropriate for gaining a preliminary understanding of the standard-setting process.

Temporal constraints affected the appropriateness of the working hypotheses in two ways: detail of recall and availability of actors. Informants were asked to reflect on a decision-making process that had occurred anywhere from two to six years ago. As time passes, the ability to recall details diminishes or at the least, becomes more selective. The passage of time also affected the availability of actors. In all of the sites, a number of individuals involved in the standard-setting process had left the organization for one reason or another. As a result, there were not always enough actors from which to sample for the range of stakes, stands, use of power and the like.

The working hypotheses were set aside as a result of the nature and use of related questions as well as temporal constraints. Consequently, increased attention was given to other issues surrounding the standard-setting process.

Specifically, greater emphasis was placed on the research questions associated with contexts and consequences. The resulting shift from a narrow to a broadened focus affected subsequent data generation, analysis and ultimately, the five narratives.

Data Analysis

Data analysis was not a discrete and independent set of activities. Indeed, data analysis began with the first interview, continued well past data generation, and ended with the final preparation of site narratives and cross-site discussion. As discussed in preceding sections, the conceptual framework, research questions, and working hypotheses guided initial data generation and analysis efforts. However, as fieldwork progressed, the data began to shape those initial guiding forces. Moreover, the reflexive nature of data generation and analysis facilitated a sensitivity to unanticipated or emerging issues.

Data analysis occurred at the intra- and inter-site levels. As stated in the introductory chapter, the primary product of this inquiry is a collection of narratives that describe the standard-setting process, context, and consequences in five sites. To produce these narratives, data analysis was necessarily descriptive and site-specific. Intra-site analysis closely followed the research questions to provide a chronological illustration of the genesis of a testing program, unfolding standard-setting process, and subsequent consequences. Analytic observations were made for

each site and used to focus the narratives. Inter-site analysis was based largely on these analytic observations; observations were compared to each of the five sites, modified when warranted by the data, and presented with any necessary qualifications.

The generated data were subjected to a reduction process during the first stage of analysis. A computer software program, The Ethnograph (Seidel, 1983), was used to facilitate the reduction process and later analyses. Interviews were transcribed onto floppy disks using a word processor. Each interview file was formatted into text forty columns wide. Speaker identifiers (e.g., interviewer and informant) were included in these files. The interview files were transformed into machine-readable (ASCII) files so they could be read by The Ethnograph. Using this program, each line of text in an interview file was numbered sequentially and a hard copy of each numbered file was printed.

After repeated readings of each interview file in a site, descriptive codes were assigned to the data. Segments of test were marked using the section and major questions from the interview protocol as codes. For example any response that addressed question E2, "What role did cut-off scores methods play in determining standards?", was coded "ROLEMETH". Figure 3.1 is an example of one page of an interview file with codes mapped alongside the text. Many of the denser or richer-meaning codes extended beyond one page of text and seldom overlapped. In order to communicate the flexibility of code-

mapping capabilities of The Ethnograph, the more nominal codes are shown in Figure 3.1.

Figure 3.1 is a copy of the first page of informant 307's interview transcript. The initial questions centered on 307's position at Waterford School District. The text coded "TITLE" identified her position in Waterford, "PASTPOS" identified her past positions, "RESPONS" marked her position responsibilities, and "BKGRD" indicated her training or background. "RESPONS" extended onto the next page of the interview file and so line #43 did not have a horizontal marker.

NUMBERED VERSION OF FILE B:307.ETH

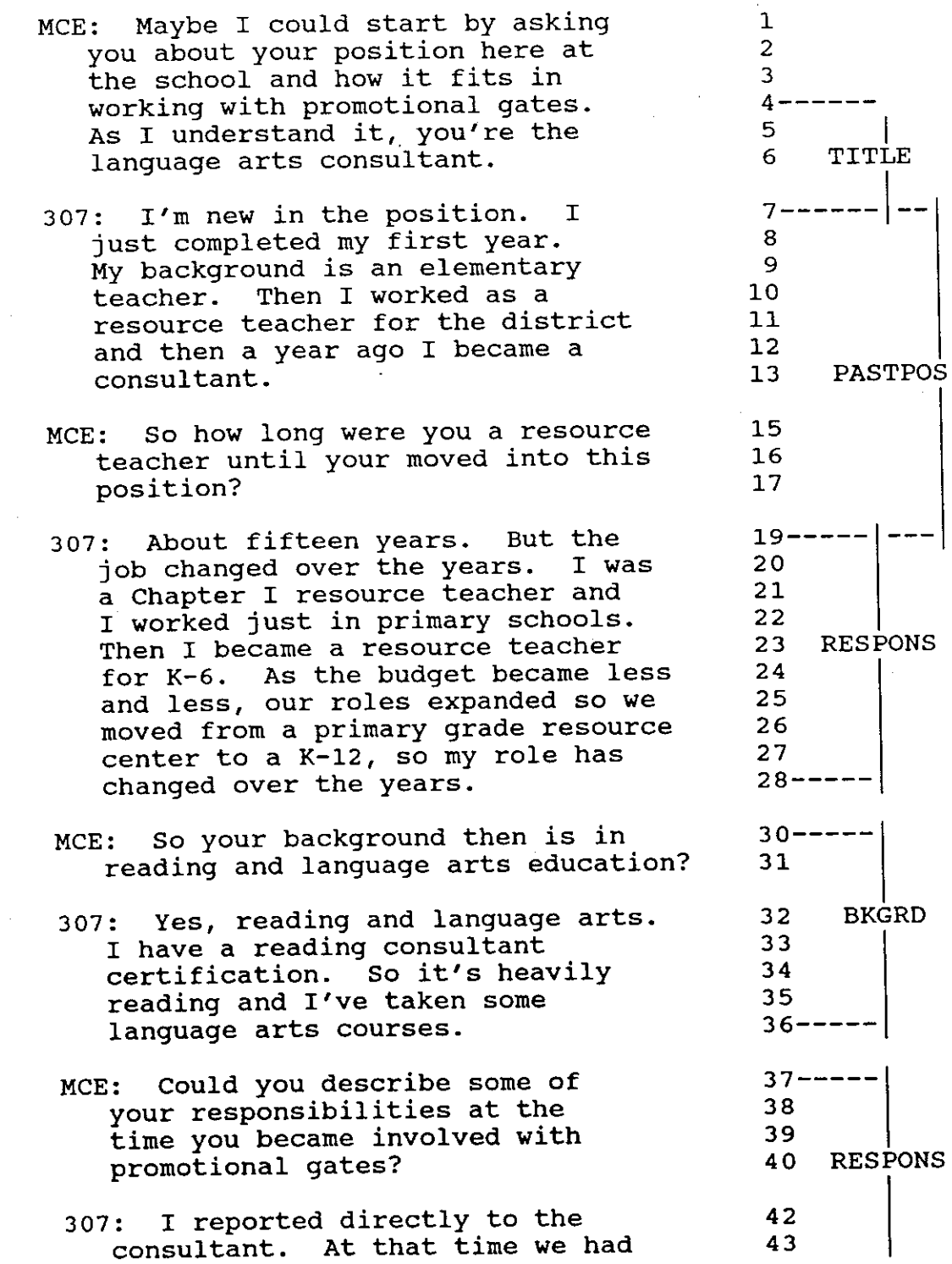

Figure 3.1. Example of numbered and coded interview file

Once the descriptive codes were mapped by hand onto each interview file, the codes and corresponding text lines were entered into The Ethnograph. After all files had been coded for a particular site visit, codes were sorted, printed and organized into looseleaf notebooks. Figure 3.2 is an example of text sorted by the code "ROLEMETH" ("What role did cut-off score methods play in determining the standards?").

```
SORTED OUTPUT FOR FILE B:E3012                          cp. 109
SORT VARIABLE:  ROLEMETH

B:3012   301
C: #-ROLEMETH
SV:  ROLEMETH

                                                  # $ % * @ ^
301:  We basically told them in the          664-#
  report the other ways of doing             665 #
  it.  We gave them a summary of             666 #
  other ways of doing it and they            667 #
  agreed with us.  It would have             668 # O.N.
  been the unusual person who had            669 #  authority
  enough knowledge with other kinds          670 #    of
  of methods.  It was basically one          671 #  technical
  of trust in the professional               672 #  knowledge
  judgment of people who have                673 # undisputed
  explored it.  Saying your job is           674 #
  to go out and explore and to give          675 #
  us a recommendation and a rationale        676 #
  for why it seems to make sense to          677 #
  us and therefore go ahead and do it        678 #
```

Figure 3.2. Excerpt of transcript coded "ROLEMETH"

Coded text were read repeatedly and finer distinctions were made among similarly-coded text. For example, text were sorted by supporting protocol questions (e.g., the code "METHCONS" referred to question E2a and "METHEVAL" referred to

E2b). Interpretations and observer notes (O.N.) were penciled in text margins and segments that addressed site observations were tagged as supporting or disconfirming evidence. In addition, disparate or conflicting accounts were identified and checked against information available in documents or in later site visits. Documents could not be managed by The Ethnograph as they were not transcribed onto floppy disks. Nevertheless, documents were hand coded and sorted using the same codes applied to the interview files.

After careful examination and coding of the data, a chronology of each site and relevant issues was constructed. Events and activities were mapped temporally starting with the history of the testing program and ending with known consequences. From this map, a draft of each site was written. Gaps and ambiguities prompted re-examination of both the coded and raw data. When such gaps and ambiguities were not rectified after re-examination of the data, follow-up questions were prepared. These questions were posed either during the second site visit or in a phone call to informants. The fruits of the individual site analyses are found in Chapters 4-8. The narratives are structured chronologically so that readers may understand the evolutionary or unfolding process of program initiation, standard-setting and consequences. Supporting evidence is in the form of direct quotes from informants and documents which are woven throughout the chronologies.

Cross-site analyses overlapped considerably with within-site analyses. Observations made at an individual site were checked against data from the other four sites. For example, the opportunities for additional test retakes was noted during the first site visit to Granada. A tentative observation of the use of safety nets was made and recorded following return from the site. As analysis of Granada progressed, it became apparent that there were other mechanisms that served as safety nets. As this observation was translated into a theme, "safety nets", the other four sites were examined for application. Although other sites were found to have similar mechanisms, cross-site investigation led to the identification of others. Eventually, a proposition based on this theme was advanced and a search was initiated for confirming and disconfirming instances. Similar procedures were followed in the evolution of the observations, themes, and propositions generated from this study. The propositions and their evidence are detailed in the concluding chapter.

Validity and Credibility Checks

One major threat to the validity or credibility of this type of study involved subject bias which could have taken a number of forms. Memory limitations may have affected the generated data. But, generally, such limitations were minimal. The standard-setting events were recent and salient enough to informants that recall was seldom hindered. Frequently, informants would refer to documents to revive faltering memories of detail and when possible, other

informant responses would be employed by the investigator to stimulate recall.

Providing data that cast only a good light on the agency or institution was another potential form of subject bias. Pledges of confidentiality, both site and informant, was presumed to help minimize this threat. Theoretically, informants could comment without fear that they or the institutions would suffer any adverse effects or repercussions. This presumption held true for most of the investigation. However, the questions derived from the bureaucratic politics hypotheses dealt with sensitive topics and it became apparent that promises of confidentiality would not sufficiently quell fears enough to produce blunt or frank responses.

A third type of subject bias included the range and type of documents provided by informants. Due to the visiting role of the investigator, access to documents was regulated by study informants. Inadvertently or intentionally, informants may have failed to provide documents relevant to the testing program, standard-setting process, or observed consequences.

Triangulation methods were employed to minimize or eliminate threats to credibility. Data-source triangulation efforts included comparing evidence provided by different informants, checking informant data against document information, and respondent validation. When topics or phenomena under study related to events and activities within the agency, the first two types of triangulation were used.

Comparing evidence generated by different informants and between informants and documents allowed for rigorous tests of validity. However, these types of triangulation were less helpful when perceptions and attitudes of informants were examined. In these instances, informant perceptions and attitudes were checked for consistency within each interview and when possible, between first and second-round interviews. Respondent validation was sought with key contacts from the sites. Copies of the narratives were sent to these contacts and they were asked to circulate them to other informants or interested parties. The key informant was contacted by phone a few weeks following receipt of the drafts. They were asked to share readers' reactions and to correct misstatement of facts about the testing program and standard-setting process. Further, respondent validation was used to fill in any gaps and to generate additional data. Due to time constraints, the use of this type of triangulation was limited; under ideal conditions, respondent validation could have been used with every informant to substantiate and elaborate.

Investigator bias served as a second potential threat to validity and credibility. The conceptual framework set forth at the start of the study could have produced an inflexible mind set for the investigator. However, the apparent utility of the working hypotheses diminished sharply during the site visits and as other issues emerged, the conceptual framework was modified. Other variations of researcher bias were addressed by respondent validation and investigator

triangulation. Respondent validation was used to ensure that portrayals were fair, balanced, and corresponded to their recollections. If informants expressed concerns of investigator bias (e.g., writing for 'good' press), the sections under question were reconsidered in light of the data and if necessary, modified. The design of this OERI-supported study necessitated the use of more than one investigator. The other investigator, the project director, participated in data generation in two sites and was briefed during and after the other site visits; read interview transcripts; independently composed site observations; compared and discussed this investigator's site observations; discussed progress of data analysis; and read drafts of the five narratives. When propositions were advanced to explain cross-site phenomena, the project director was asked to examine and judge the chain of evidence presented to support the propositions.

Aids in Reading the Narratives

To protect their identities, the five sites were assigned pseudonyms and identification codes: Waterford (3), Lucerne (4), Victoria (5), Brittany (2), and Granada (1). In like manner, titles of positions, tests, and agency documents were changed whenever possible. Informants were assigned identification numbers corresponding to the site codes. For example, the Director of Curriculum Services in Lucerne was informant #403. Second-round interviews with informants were identified by a "2" in the second digit place. For instance, the second interview with the Lucerne Director of Curriculum

Services was coded as #423. Documents were coded in numeric-alpha form. For example, document 5A referred to the first document obtained from Victoria.

CHAPTER IV

COMPETENCY TESTING AND ELEMENTARY GRADE PROMOTION IN WATERFORD

A Glimpse of the Present

Each spring, the Waterford School District administers competency tests to students in grades K-8. The tests are locally-developed, criterion-referenced examinations that cover grade-level skills and material. In this case study, we will focus only on tests at four of the elementary grades: K, 2, 5, and 7. Unlike other grades where the tests are used as indicators of progress, the tests in grades K, 2, 5, and 7 bear a heavier burden: they are to be the primary criteria for making decisions about promotion and retention. Specifically, students in these grades must pass the competency tests, known as promotional gates, in order to be promoted to the next grade. To all outward appearances, social promotion in Waterford is dead.

As in the other four case studies, we will examine the standard-setting process and its consequences. The central storyline to Waterford, however, will not be found in backroom political maneuvers nor hard-nosed technical efforts to derive standards for the promotional gates. Rather, a description of the consequences of this testing program, together with preceding events, will convey the theme of Waterford: the loose coupling between the intent of testing reform policy and actual practice.

A Brief History of Events Preceding the Promotional Gates

Since the early 1970's, a city-wide evaluation committee met to discuss and make recommendations about testing and related issues in the Waterford School District. The committee was composed of representatives from teaching, counseling and administration from around the district, which serves a population of nearly 500,000. The committee faced a number of issues during the 1970s; two such concerns were expressed by the district math specialists and teachers as a whole. The former group wished to implement criterion-referenced tests, a task normally confined to the district testing office. The latter group expressed concern that the norm-referenced tests used by the district didn't measure what they taught. A number of committee individuals were intrigued enough with these issues that they began to meet on a informal basis for over two years to discuss possible directions that the district could take. Criterion-referenced testing came up time and again during their discussions. While members of this informal committee agreed that locally-developed, outcome-based objectives must precede the development of district criterion-referenced tests, members disagreed about the use of such tests. Some felt that the tests would be appropriate only for monitoring curriculum needs. More felt that the tests could be used for assessing student progress, even for making promotion-retention decisions.

In the early 1980's, a new superintendent took office. Charismatic and competent, the new superintendent was seen as the answer to many of the problems plaguing the Waterford School District. One such problem was voiced by the business community; high school graduates were not able to fill out basic application forms. In the words of the Deputy Superintendent,

> The business community wanted the schools to ensure that students would reach their maximum academic potential and then business could take over from there and train students to serve its purposes. The superintendent is an unusual visionary type of person who can conceptualize and he believes that every person can reach their maximum academic potential. (308)

Superintendent Williams felt that there was a need to monitor the progress of students systematically; a conviction that well coincided with beliefs held by the informal testing committee. In fact, members of this informal committee met with the superintendent's cabinet to present some of the issues and concerns they had been dealing with for the past two years. The grade-level, criterion-referenced tests intrigued Superintendent Williams and his cabinet. In the following months, Superintendent Williams announced that social promotion in Waterford schools would no longer be condoned nor practiced; quantifiable measures of progress should be used in making promotion decisions. His announcement was formally incorporated into the district's five-year plan:

> Addressing the needs of students who are not achieving up to expectations is an important component of the mission of the Waterford School District. The following commitments will be made as the school district

> proceeds to develop policies and procedures for intervention and for promotion or retention of students . . . . #12 Test data will be used as primary information to be considered in decisions about promotion or retention. (3A)

The Deputy Superintendent echoed the commitment in the district's five-year plan:

> The Board [of Education] had made a statement that the district was not going to promote students socially. We needed some measures as checkpoints to assure that students wouldn't slip through the cracks. (308)

To this end, a committee was formed to make recommendations about achievement standards, promotional gates and curriculum interventions. The committee was composed of over forty individuals from teaching, administration, curriculum and other special programs. The committee met from June, 1982, through February, 1983, with three subcommittees also convening during that time. A final report of the committee's work was produced in March, 1983. A number of the committee's recommendations for policies and procedures are discussed in the following section.

Blueprint for Excellence

Recommendations from this district-wide committee dealt with issues of achievement standards, promotional gates, curriculum interventions, retention policies, parent notification and due process, implementation timeline, position responsibilities, staff development, evaluation and budget estimates. A few of the recommendations will be discussed here; recommendations that would set the stage for

the actual determination of test standards. The remaining recommendations will be discussed in following sections.

Criterion-referenced tests had been developed in reading and mathematics for grades K-9, with tests of writing in four of these grades (3, 5, 7, and 9). Rather than basing promotion decisions on each grade's test performance, the committee selected four grades to serve as the promotional gates. According to the committee's report:

> In order to provide students the experiences that accommodate the readiness of five year olds, the first promotional gate is at kindergarten. Here the focus is not so much upon achievement as it is upon readiness for school learning. Some students need additional experiences and/or maturing time before going on to first grade; that is, they need opportunities to build readiness. The gates at grades 2, 5, and 7 organizationally occur at about the middle of the K-3, 4-6, and the 7-8 schools. There is still enough time before the passage to the next school for necessary skill acquisition. The gate placed one year ahead of the normal "end point" for that level of schooling allows for reassessment of the impact of the interventions during the repeated grade year with an additional year to monitor progress and support achievement gains of the students before leaving the environment of the primary, intermediate, or junior high settings. (3A)

In addition to recommending the grades at which the promotional gates would be placed, the committee identified the means by which standards would be determined:

> Standards of performance for the benchmark tests will be established centrally. The process to be used correlates test results with teacher-determined criteria of performance. (3A)

Another standards issue was dealt with by the committee. Instead of having one score to separate students, two scores would be determined for each promotional gate test. According to the committee's report,

> Test scores that are closest to the standard that separate proficiency from non-proficiency may not accurately reflect performance due to error of measurement present in all test. Teachers may make errors in their judgment about student proficiency. Because those errors exist, scores from each test will be divided into three bands: satisfactory, questionable, and unsatisfactory. (3A)

The process by which the committee decided upon the standard-setting method and the multiple cut-scores, as well as the method's implementation, will be discussed more thoroughly in the following section on standard-setting. The committee's recommendations are mentioned here for two reasons. The first is to convey the understanding that many policy and procedure decisions were made before the actual setting of standards; they were not by-products of the standard-setting process. Second, another recommendation made by the committee in their 1983 report can be understood better in light of these two previous recommendations. This recommendation was actually a set of decision rules and an accompanying flow chart by which promotion decisions would be made. These decision rules were made necessary by the existence of two (or three) subtests at the 2nd, 5th, and 7th grade promotional gates. Since a student would score in one of three areas on each of the subtests (satisfactory, questionable, or unsatisfactory), a student could have any number of score combinations on the entire promotional gate test. In the words of a committee member who prepared the committee's final report,

> One of the issues we dealt with was what would happen if a kid fails one test and passes the other(s). We came up with some decision rules and we fought and argued over

> those damn things, too. In the back of the whole argument were those people who were basically against retaining anybody and I think those people would argue that you should only retain those kids who fail all of the subtests. [Others would argue,] 'We have a superintendent who says kids won't pass if they fail a promotional gate, so how can you go ahead and promote a kid who has failed even one of the subtests?'. (303)

After considerable debate, the committee finally agreed on a set of decision rules and approved a flow chart by which promotion decisions would be made in the Waterford schools. According to the committee's final report,

> At kindergarten, there is one benchmark test which assesses reading and mathematics readiness. Students who have scores in the satisfactory range generally will be recommended for promotion. Questionable scores will lead to a review of total performance to determine promotion or retention. Unsatisfactory scores will generally lead to retention. (3A)

The decision rules for 2nd, 5th and 7th grade test performance were considerably more complicated. For all three grade levels, students were to be classified into one of three recommendation categories on the basis of subtest performance: promotion, review, or retention. Table 4.1 shows the possible combinations of subtest performance for each category. As can be seen in the table, there are six combinations of performance on the two subtests (T1 and T2) in grade 2. Two of these combinations slot students into the promotion category: 1) pass both T1 and T2 or 2) pass one subtest and score in the gray area on the second. For grades 5 and 7, there are ten performance combinations for the three subtests which place students in one of the decision categories.

Table 4.1

Combinations of Subtest Performance for Promotion Decisions

	Grade 2		Grades 5 and 7		
Subtest	T1	T2	T1	T2	T3
Promotion recommendation:					
C_1[a]	S	S	S	S	S
C_2	S	Q	S	S	Q
C_3	-	-	S	S	U
Review recommendation:					
C_4	S	U	S	Q	Q
C_5	Q	Q	S	U	U
C_6	-	-	S	Q	U
C_7	-	-	Q	Q	Q
Retention recommendation:					
C_8	Q	U	Q	Q	U
C_9	U	U	Q	U	U
C_{10}	-	-	U	U	U

Note. Where **S**=satisfactory subtest performance

Q=questionable subtest performance

U=unsatisfactory subtest performance

[a] possible combinations of subtest performance

Once the decision rules were formulated, rules that sorted students into recommendations for promotion, review or retention, the committee drafted a procedural flow chart for promotion-retention decision-making. Students recommended for promotion would automatically be promoted to the next grade. Students recommended for retention would normally be retained. The committee's final report described the fate of students recommended for review:

> Students recommended for a review of total performance will have their scores examined by a review team consisting of the principal, the classroom teacher plus at least one other staff member (e.g., social worker, counselor, reading or math specialist, or intervention teacher). A review of additional objective data, in addition to staff judgments, will lead to either a case conference, a retention recommendation or a promotion. (3A)

A case conference would be called when there had been no delivery of intervention or parent notification. At this time, staff and parents would meet to determine whether the student would be promoted or retained. Although a consensual decision was seen as most desirable, the final decision to promote or retain a child would be made by the building principal (3A, p. 24).

An addendum to the committee's recommendations gave fleeting reference to some non-test considerations in promotion decisions:

> A student should be retained for no more than one year at any single grade level. The next consideration for a repeated retention would normally occur at the next promotional gate. Any student recommended for more than one retention, excluding kindergarten, can be retained only after a case conference. (3A)

This set of recommendations addressed the inevitable questions of how to deal with multiple retentions during the course of a student's time in elementary school. Not only did the committee speak against repeated retention in any one grade, but they indicated that retention should generally occur only at the promotional gates of 2, 5, and 7.

A committee member instrumental in the writing of the final report described the decision rules, flow chart and multiple retention policies as means for gaining cooperation from those committee members expressing concerns about the promotional gates program:

> We were screaming about [these issues] and we finally came up with a flow chart which said here's how we'll handle promotion-retention decisions. The checks and balances built into the chart, like the opportunity for review, finally sold those folks who were against the whole concept from the beginning. (303)

We will note one other set of recommendations made by this committee before we describe the standard-setting process. Plans were carefully laid for implementation of the four promotional gates. A timeline was created for test piloting, standard-setting and implementation of those standards. Tests would be piloted along a staggered schedule beginning in Spring, 1983, with standards set on the city-wide pilot data, and implemented the following year. By Spring, 1986, all four promotional gates would be in place and guarded.

Setting Standards: Where Reason Prevails

The Committee on Achievement Standards, Promotional Gates and Curriculum Interventions recommended the method by which standards on the promotional gates would be determined:

> Standards of performance for the tests will be established centrally. The process to be used correlates test results with teacher-determined criteria of performance. (3A)

Central to this recommendation were efforts made by individuals in the district testing office, professionals judged by the standards subcommittee as those most informed on standard-setting methods and issues. Rather than wading through and interpreting the plethora of information on standard-setting as a whole, the subcommittee looked to these three individuals to recommend the most appropriate standard-setting procedure for Waterford.

One of these individuals, the current Director of Testing, described their efforts:

> [We asked ourselves,] 'So we have to set a standard. How do we go about it?' We talked to people from Educational Testing Service (ETS), read as much literature as we could and then after reading the literature, we decided that we would use the contrasting groups method. (301)

Another of these three individuals elaborated on the preceding account by her colleague:

> We did a lot of reading: papers, articles, and a book on [setting] passing scores. When we finished going through all of the reading, we came to the conclusion that the contrasting groups methods was the best method to use for our purposes . . . . The basic difference between the methods we didn't select and the contrasting groups method is that you evaluate the student as to what his or her skills are, rather than evaluating the [test] items. I think teachers know kids better than they know

> items . . . it's harder for teachers to use the [item-judgment] methods than it is to tell whether the kid is competent. (302)

The recommendation to use the contrasting groups method, as well as an explanation of the decision-making process, was brought before the standards subcommittee. The other methods were summarized briefly and reasons for selecting the contrasting groups method were presented. The Director of Testing reflected on the subcommittee's response:

> . . . they agreed with us. It would have been the unusual person who had enough knowledge of other [cut-score] methods. It was basically an issue of trust in the professional judgment of people who have explored [the various methods]. They said to us, 'Your job is to go out and explore; give us a recommendation and a rationale for why it makes the most sense [to use]'. (301)

With the standard-setting method approved, the responsibilities for implementing the procedure were delegated to the district testing office. This office, along with the curriculum department, had been in the throes of objective writing, test development, bias reviews and were ready to begin pilot testing. The kindergarten test was the first to be piloted in Spring, 1983. The standard-setting procedure described in the following paragraphs was much the same for all of the promotional gates, with variations in people involved and the number of subtests for which standards had to be set.

The district testing office sent letters to teachers involved in the city-wide pilot of each test. Teachers were asked to assess each of their students as a master or non-master of the skills measured on the test before administering

the test. The following definition was provided to help teachers differentiate between the two types of students:

> A master is defined as a student who know enough of the basic skills in this area that if you had to make a decision about promotion or retention solely on this skill level as it will be measured by this test at this time, you would elect to promote the student. A non-master is defined as a student whose reading [math or writing] are such that if you only had two choices-- to promote or to retain-- you would elect to retain the student at this time. (3A)

Teachers were urged to draw upon any and all of the information they had about each student: classroom performance; other test scores; and intuitive assessment of how the student would perform in the areas measured by the test. The Director of Testing noted that even with encouragement to draw upon other sources of information, teachers expressed some concerns with identifying students as masters and non-masters:

> I think the concept of 'master' bothered teachers more than the 'non-master' concept. To many [teachers], a master really must mean someone who is really good; the problem was what was good enough for a student to be labeled a master? (301)

After teacher judgments had been collected and test results were available, the district testing office merged the two sets of data and displayed them in tabular form. Table 4.2 is a portion of the kindergarten data from which the standards would be determined.

Table 4.2

Kindergarten Master/Nonmaster Distributions

Raw score	# master	# non-m	% master	% master smoothed	% scoring at or below
41	38	0	100	100	-
40	74	0	100	100	-
39	148	0	100	99	-
38	156	2	99	99	-
37	214	1	99+	99	-
36	183	3	98	99	-
35	173	2	99	97	-
34	147	11	93	95	-
33	110	12	90	91	-
32	98	11	90	87	-
31	102	24	81	85	-
30	94	18	84	84	37.9
29	83	13	86	81	32.9
28	63	24	72	78	29.1
27	55	20	73	72	25.4
26	50	21	70	68	22.1
25	20	18	53	61	19.3
24	22	21	51	48	17.3
23	17	24	41	44	15.3
22	13	22	37	40	13.4
21	17	25	40	39	11.8
20	13	20	39	36	9.9
19	7	22	24	28	-
18	4	20	17	20	-
17	3	13	19	15	-
16	2	20	9	11	-
15	1	15	6	13	-
14	4	10	29	17	-
13	2	10	17	20	-
12	3	15	17	14	-
11	1	11	8	11	-

The first column in Table 4.2 shows the raw scores on the kindergarten test, which ranged from 5 to 41 on the pilot administration. The next two columns display the numbers of students rated as masters and non-masters, respectively, who earned the corresponding raw score. Column four is the ratio of numbers of students identified as masters relative to the total number of students earning a particular raw score. Column five, labeled as "Percent Master Smoothed" is an average based on the numbers (masters and total) for a particular raw scores and numbers on either side of that score. For example, a raw score of 27 corresponds to a percent master smoothed of 72%. This figure is based on the numbers of students identified as masters for raw scores of 26, 27, and 28 relative to the total number of students earning these three scores. The last column displays information only at the middle of the frequency distribution; each entry is the percent of all students who scored at or below the particular raw score. For example, 25.4% of the kindergarten students tested in the pilot administration scored at or below 27.

For each subtest in each promotional gate test, similar frequency tables were constructed by the testing office. These tables, and others, were used during the standard-setting process by committees created for that purpose. Specifically, a committee of 10 to 15 members was created for each promotional gate. Each committee was composed of district teachers and principals, as well as individuals from central

administration: curriculum and testing offices (3C). For each promotional gate subtest, the committee would inspect the data shown in the frequency tables and apply the contrasting groups method to determine the test standards. As applied in Waterford, the standard identified by the contrasting groups method would be the score at which 50% of the examinees were masters.

Before each committee convened, individuals from the curriculum and testing offices sat down to examine the test performance and teacher judgment data. According to the Director of Testing, they would identify a couple of possible standards; scores at which nearly 50% of the students were masters. When each standard-setting committee convened, these people delivered their recommendations at which time the committee discussed those and other possible cut-scores. By examining the frequency distributions, the committee members could see the proportion of masters students who earned a particular score (columns four and five), as well as an idea of how many students would fail the promotional gate for any given score (column six).

These frequency tables were not the only sets of information available to the standard-setting committees. The testing office also prepared tables for each of the recommended cut-scores, displaying test performance by ethnic group and sex of mainstream students, as well as breakdowns by other students groups: special education and limited English proficiency. Another table was prepared to indicate the

percent of students (by ethnic group and sex) who would be sorted into one of three recommendation categories: promotion, review, or retention.

The language arts resource teacher served on a number of the standard-setting committees. She described the general decision-making process followed by the committees:

> In the first part of the meeting, people from the testing office would explain how to read the data. They would explain where committee members might want to look in the data for possible cut-scores. Then the committee would discuss those scores in light of their implications. (304)

A number of considerations were brought to bear on the committee's decision-making process; considerations that may be grouped into three categories: technical issues, practical realities, and appearances of the standards. Considerations in the technical issues category included adherence to the cut-score method and acknowledgement that there was no absolute standard. According to the Director of Testing,

> As long as the committee members didn't violate what I thought was the intent of the contrasting groups method, then I basically took a position that they knew more about the curriculum and expectations to set a reasonable standard. To me, a violation would be if they set the standard where 80% of the kids were rated as masters--any extreme [percentage] would be a violation of the method. (301)

Another technical consideration was the committee's acknowledgement that there was no absolute standard. The Math Consultant who served on a number of the standard-setting committees described the prevailing attitude:

> Any standard setting is arbitrary. I mean, the bottom line is that kids above the cut-score pass; kids below the cut-score do not. And the reaction is that the standard-setting is going to be arbitrary, but not

> arbitrarily arbitrary. You don't just go and draw a line. The contrasting groups method seemed to be a good enough process. You could move the cut-score up or down a few points and I don't know that doing so would make that much difference. When you get down to it, the kids who are marginal-- a few points higher or lower than the cut-score-- are going to get into some kind of [intervention] program anyway. (305)

The Director of Testing echoed the Math Consultant's view:

> The committees worked very well . . . . It was like 'We have a job to do, so let's do what we think is the most rational and reasonable thing to do . . . . If I'm holding out [on group consensus] and my preferred cut-score is a point different, I'm not going to argue it'. Let's face it, the cut-score is not perfect. There's no magic point. (301)

Practical concerns form another category of considerations taken into account by the standard-setting committees. One such example comes from the Director of Testing when addressing concerns about particular cut-scores and their corresponding fail rates:

> And we weren't entirely dumb about what the district could afford. We never went and asked [central administration], but we knew what the common sense approach was. (301)

When asked if the tacit understanding of what the district could afford influenced the recommendations given to the standard-setting committees, he explained: "Sure. I'm not supposed to say that, but it was. I wasn't going to recommend a standard that would produce a 50% fail rate." (301)

The Deputy Superintendent lent a slightly different perspective:

> When you consider that all of the people working on the tests and standards were seasoned teachers, their

> experience would tell them that a standard that identifies only a small percent of students is utopic. That's not a reality-- it's not what teachers face in the classroom. The same thing is true if a standard failed 60% of the students. Kids are doing much better than that. There's a certain amount of intuitiveness to a standard. (308)

Committee members came from a number of different areas (curriculum, testing, administration, and teaching) and so, many concerns stemmed from these different vantage points. According to the language arts resource teacher,

> Different people had different agendas when they set the standards. I think that when statisticians look at the data, they see the numbers in the distributions and rely on [those numbers] to pick out the high risk kids. And then someone from a curriculum viewpoint looks and says, 'Oh my god. What are we going to do? This is an 80% minority group and there are all these kids, all in different places. How are we going to design a curriculum that is going to help these kids?' And then there's the administrative viewpoint: 'We don't have any classrooms. How are we going to schedule these kids? What are we going to do with them? Who is going to want to teach them when we put them all together-- a lot of them are going to be behavior problems'. (304)

The third set of considerations taken into account by the standard-setting committees dealt with the appearances of the standards. In the words of the Math Consultant:

> The idea is that the passing score has to be high enough and the number of students failing low enough . . . . If you take a look, the cutoff scores range between 60% and 76% correct. I think our feeling was that if the cut-score is lower than 60%, it is pretty low for a competency test. (305)

Taken together, the considerations brought to bear on the standard-setting process would seemingly prevent the committee from arriving at any consensus about the test standards. However, each committee was able to cooperate and reach

consensus on the cutscores. According to the Language Arts Consultant:

> The one thing was that everyone was willing to listen. Everything that anybody said was important, whether you were a resource teacher, classroom teacher or a principal. (307)

The language arts resource teacher echoed the consultant's account:

> We just reached consensus; standards that we could agree with. The Director of Testing was very good about saying, 'Can you live with this cut-score?' I think everybody left feeling really good about it. I mean, we hadn't gone too high nor too low [for the cut-scores]. (304)

The kindergarten cut-score was easier to achieve consensus on than the other promotional gate tests. According to several committee members (301, 302, 305), the kindergarten data were such that the cut-scores were obvious; the breaks between the master and non-master distributions were clearer. The data for the higher grades were more ambiguous and cut-scores could have been set in a number of places. In those instances, normative information was referenced; information about the percent of students failing was used to decide between possible cut scores.

Passing scores were set on each promotional gate subtest. Another score for each subtest was set in relation to the passing score. This latter standard was usually two standard errors of measurement (SEM) below the passing score. Students who scored below the passing score and above the latter standard were said to be in the 'gray' or 'questionable' area. Students scoring below the latter standard were in the 'fail'

or 'unsatisfactory' area. In describing the purpose of the gray area, the Math Consultant echoed the concerns of the city-wide committee which recommended two standards for each promotional gate test:

> I think people have all seen situations where a child's test performance is not in agreement with their performance apart from the test. People didn't want to have an inflexible situation, one that nobody could work with. (305)

The final standards for the promotional gate subtests are displayed in Table 4.3. Passing scores are included only for the multiple-choice subtests; the writing passing scores for grades 5 and 7 were determined by a different method, not detailed in this case study.

Table 4.3

Promotional Gates Test Standards

Test (Form A)	No. of items correct (% correct)	
	Passing score	Gray area score
Kindergarten	26 (63%)	21 (51%)
Grade 2		
Reading	45 (75%)	40 (67%)
Mathematics	42 (75%)	37 (67%)
Grade 5		
Reading	49 (70%)	44 (63%)
Mathematics	41 (63%)	36 (55%)
Grade 7		
Reading	36 (60%)	31 (52%)
Mathematics	51 (60%)	46 (54%)

Once the committees had set the standards, the Director of Testing circulated the scores to the Superintendent's cabinet for approval. According to the Deputy Superintendent:

> We would share the figures with the cabinet and our superintendent. They raised a lot of question because they would have to answer to the Board of Education and the public. [Some of the concerns included] the different standards in terms of percent of correct items; the projected fail rates for each grade level; and the percent of minority students who would fail. Their concern was something like this: 'I accept this standard, but what are we going to do about it? How can we make a difference so that the children who fail this year are not the ones who fail next year. (308)

After the standards were approved by the cabinet, a presentation was made to the Board of Education. No standard was changed or rejected by the Board. According to the Math Consultant,

> If we had come up with a standard by which 50% of the kids would fail, I am sure the cabinet would have told us to go back to the drawing board. But, I think the standards weren't changed because as we were setting them, we were looking at the political realities. (305)

Before consequences related to the promotional gate program are discussed, the remaining policies and procedures will be outlined in the following section.

Policies and Practices of the Promotional Gates Program

1. All regular education students are expected to take the promotional gates tests. The Individual Education Program (IEP) for special education students will indicate which tests the students will take, if any.

2. If a student fails a promotional gate test, teacher judgment may overrule the results--paving the way for promotion to the next grade.

3. Retesting is allowed only after a special request by the teacher has been submitted.

4. Students who fail the promotional gate tests and are retained are eligible to receive retention services. In kindergarten, retention services are delivered in a K-1 transition curriculum. Other grades offer pull-out programs for skills work during the retained year.

Consequences of the Promotional Gates Program

Table 4.4 summarizes the passing rates for each promotional gate test since the inception of the piloting phase in Spring, 1983. These data were provided by the Director of Testing.

Table 4.4

Promotional Gates Passing Rates

Test	1983	1984	1985	1986
Kindergarten	81%[a]	89%	87%	86%[b]
Grade 2				
Reading	--	78%[a]	81%	87%[b]
Mathematics	--	85%[a]	90%	90%[b]
Grade 5				
Reading	--	--	80%[a]	87%
Mathematics	--	--	76%[a]	82%
Writing	--	--	74%[a]	87%
Grade 7				
Reading	--	--	83%[a]	84%
Mathematics	--	--	78%[a]	84%
Writing	--	--	70%[a]	82%

Note. Not applicable indicated by --.

[a]projected passing rates from pilot test data

[b]passing rates from parallel form B

Passing rates on the grades 2, 5, and 7 promotional gate subtests do not convey the whole picture. Rather, the percent classified in each of the three recommendation categories (pass, review, retain) via the decision rule gives a much clearer picture of who would pass and who would not. Table 4.5 shows the percent classified in each category by grade and year. Kindergarten is not included in this table as the classification of students into pass, review, and retain categories was not based on a similar decision rule. Instead, students who scored 1) at or above the cut-score; 2) below the gray area cut-score; or 3) between the pass and gray area cut-scores were recommended for promotion, retention or review, respectively.

Table 4.5

Pass, Review and Retain Classifications

Grade	Category	1984	1985	1986
2nd	Pass	81%[a]	86%	85%[b]
	Review	10%[a]	9%	7%[b]
	Retain	9%[a]	5%	7%[b]
5th	Pass	--	75%[a]	81%
	Review	--	13%[a]	9%
	Retain	--	13%[a]	7%
7th	Pass	--	78%[a]	74%
	Review	--	11%[a]	9%
	Retain	--	11%[a]	7%

Note. Figures do not add to 100% in the last column. The remaining percent of students were not classified due to missing subtest scores.

[a]derived from pilot test data

[b]based on parallel form

When asked whether the fail rates on the promotional gate test were seen as reasonable, the Director of Testing responded: "The education community and community in general buys it. It seems rational." (301)

The metropolitan newspapers published test results each spring, ranking individuals schools by passing rates. Among those interviewed, district officials felt that the reporting was generally even-handed. According to the Deputy Superintendent:

> In retrospect, I think the media was very fair. We spent a lot of time going over the test results and I think the reporters sensed the seriousness of an error in reporting the results. And so, they were not alarmists. They didn't have control over their headline, so the headliner would tend to shock people with percentage of failures and percentage of minority students. It didn't seem to be detrimental; in fact, it was helpful. If what we were doing was for a political motivation, it would have been a good strategy because the common response was 'It's good to see a school district insist the kids learn and it's not fair to kids to pass them if they don't know what they should'. (308)

What actually happened to students after taking the promotional gate tests? While the Office of Testing could cite how many students passed the promotional gates tests, as well as how many failed, the department was not responsible for tracking retention decisions. Neither was the task the responsibility of the Office of Research and Evaluation. However, this department was charged with evaluating the services delivered to students held back in grades K and 2. Drawing upon these evaluation reports, as well as earlier testing reports, we will examine the connection between promotion recommendation and practice.

Table 4.6 is a 2 X 3 table that displays information about the 1984-85 Kindergarten class. The columns indicate the numbers who passed and failed the Spring, 1985 promotional gate. The rows indicate the actual decisions made about these students: promoted to first grade and retained (with or without retention services) in kindergarten.

Table 4.6

Kindergarten Test Performance and Promotion Decisions

Decision		Promotional gate performance		
		Failed[a]	Passed	Total
Promoted		(117)	(2,709)	(2,926)
Retained	retention services	278	1	279
	no retention services	11	7	18
		406	2,717	3,123

Note. Numbers in parentheses inferred from other table cells.

[a]Includes those who scored below the passing score of 26

In Spring, 1985, 3123 regular education students took the kindergarten promotional gate test. As shown earlier in Table 4.4, 87% of these students (n=2,717) scored at or above the

cut-score, thereby passing the exam. The remaining 13% (n=406) scored in either the gray or unsatisfactory area, thereby placing them into the review or retain categories, respectively. (3F) Of retained kindergarten students receiving transition services in 1985-86, 278 scored in the gray or unsatisfactory areas of the 1984-85 promotional gate test. An additional 11 students who failed the 1984-85 promotional gate were retained but did not receive transition services. At least eight children were retained even though they passed the 1985 promotional gate; only one of these students received transition services designed for those who failed the test. Approximately 117 students (29%) of the 406 who failed the 1984-85 promotional gate tests were not found in 1985-86 kindergarten classrooms. It is likely that all of these students were promoted to first grade. According to the evaluation report released in Fall, 1986, more than 300 kindergarten students who failed or barely passed the 1985 promotional gate were promoted to the first grade. (3P) Tracking the fate of students within a large system is difficult for any district office to do. When two offices monitor test performance and promotion decisions separately, comparisons are sure to be disparate. Nevertheless, one observation can be made: test information was used in an asymmetrical manner when promotion decisions were made. With a few exceptions, students who passed the promotional gate test were promoted to first grade. The contribution of test information seems to be weaker when considering retention for

children who failed the kindergarten gate. Nearly 30% of the students who failed the test were promoted to first grade.

Information about ethnicity and sex may have influenced retention decisions about those who failed the kindergarten test. Minorities accounted for 59% of the students who failed the 1985 kindergarten test, however, 69% of the retained students who received transition services were minorities. Among those failing the kindergarten test, 56% were male, but they accounted for 65% of the retained students receiving transition services.

The connection between test performance and promotion decisions is even more obscure at grade two. We tried to construct a table similar to Table 4.6, but one with three columns for the categories: pass, review and retain. From the information provided by the two offices, we could not fill in the table cells completely. While the district has indicated interest in knowing what happens to every student, the information has not been compiled in a way to link recommendation and practice. Nevertheless, a few points can be made. In Spring, 1985, 2,521 regular education students took the promotional gate test. As indicated earlier in Table 4.5, 86% (n=2,171) would be recommended for promotion to third grade and 450 would be reviewed or retained. According to the evaluation report (3Q), 169 were retained, nearly 38% of the eligible 450 students. By our best estimate, it appears that over 60% of the 450 students were <u>not</u> retained in second grade. Further, the evaluation report indicated that roughly

400 students were promoted to third grade even though they were placed in the retain or review categories. The observation drawn for kindergarten seems to hold for grade two: promotion decisions are most congruent with recommendations when students have been slotted into the pass category; the link is much weaker between retention decisions and the review or retain recommendations. One of the informants alluded to the weak link (the informant's i.d. number is not provided-- it was apparent that this comment was a sensitive one to make):

> Kids are passed or retained on a larger agenda, and yet, I don't dare say that very loudly. We don't have the statistics to confirm anything, but I know there is a different number of kids in retention compared to those who were eligible for retention. What we don't know is what happened to those kids who weren't retained, nor why that happened. (30X)

As with kindergarten, students' ethnicity may have influenced decisions to retain in second grade. Of those failing the reading and math subtests, minorities accounted for 61% and 62%, respectively. Of those receiving retention services, minorities accounted for 70%-- disproportionate to the percent failing either. Males accounted for 60% and 47% of those failing the reading and math subtests, respectively, only 56% received retention services. Because figures were not available on the entire promotional gate (subtests combined) by ethnicity and sex, conclusions about either's influence on the decision-making process are speculative at this point.

We were unable to examine the connection between test performance and retention decisions at grades five and seven. Retention services for these grades had yet to be implemented as the first administration for each of these promotional gates was Spring, 1986.

Another view of the loose coupling between district promotion policy and actual retention decisions was provided by the language arts resource teacher. Elementary building principals met unofficially to make certain modifications to existing policy:

> There are some schools who prefer to retain after grade one, because they feel that the sooner the retention, the better. So if a student is retained in first grade and then fails the grade two promotional gate, he or she will not be retained again. (304)

Their informal understanding was not translated into official policy. Indeed, the notion of promotion bands was introduced in the 1983 committee report. However, this informal agreement serves to empower building principals and weaken the centralized policy that retentions occur only at the promotional gates. Because of this agreement, principals may exercise wider discretion in promotion decisions.

Other consequences are related to the Waterford promotional gates testing program. Students receive remediation or retention services at grades K and 2. At kindergarten, retained students enroll in a regular kindergarten class for half of the day and in an intensive skill development class the other half of the school day. (3P) In second grade, retained students are provided supplemental

reading or math instruction outside of the regular classroom. (3Q) Both evaluations cited teachers' perceptions that the programs had positive impacts on achievement and achievement-related behaviors. Attempts were made to compare retained students with promoted students on a number of issues, especially future promotional gate performance. Unfortunately, students were not tested along a common metric and so, comparisons would not be appropriate here.

Superintendent Williams appears pleased with the promotional gates program. Two informants indicated that the Superintendent has yet to hear parental complaints about the program (301, 308). When asked last spring to comment about public relations and the promotional gates program, the Director of Testing remarked:

> The program is in great shape. I would never in my fondest expectations have thought that it would be in as good of shape as it has been. The program has been well done. (301)

The Math Consultant concurred: "I think it looks like the program is going to be around for a long time." (305)

According to many, Waterford's promotional gates testing program is strong and running smoothly. However, unless more is learned about the coupling between testing and promotion, the competency tests will resemble an end, rather than the means to it.

CHAPTER V

COMPETENCY TESTING FOR HIGH SCHOOL GRADUATION IN LUCERNE

A Glimpse of the Present

The Lucerne School District requires high school students to take two competency exams in the spring of their freshman year. The reading test, Senior High Assessment of Reading Proficiency (SHARP), is a nationally-normed test developed by CTB. The SHARP has 120 questions covering thirty objectives. Lucerne developed the mathematics test, Inventory of Mathematics Competencies (IMC), which examines 25 district objectives with 75 items. Examinees must pass 100% of the objectives on each test. The criteria for passing an objective varies by exam: three of four items per objective must be answered correctly on the SHARP and two of three items per objective on the IMC. If an examinee fails to meet these standards, an alternate standard is applied: pass 70% of the total number of items on each test. Whatever standard is applied, students must pass both exams by their senior year to receive a high school diploma.

The minimal competency testing program in Lucerne is probably typical of many elsewhere and the following description of Lucerne could mirror any one of a number of testing programs across the nation. Indeed, elements of political maneuvering and technical sophistication are absent from our illustration of Lucerne. The testing program did not

spring from a politically-turbulent setting nor did standard-setters cling tenaciously to canons of technical excellence. What makes Lucerne interesting and informative, however, is the symbolic nature of the minimum competency testing program which can be discerned in the end result: the negligible number of students who fail to receive a high school diploma.

Forerunner to the Lucerne High School Graduation Exam

In February, 1970, the Lucerne Board of Education adopted a requirement that all sophomores take competency examinations in reading and mathematics. Students who failed either of the tests were required to enroll in a one-semester remedial course during the fall of the eleventh grade. This class, termed a lab course, was designed to improve skills in the two subjects. According to the Director of Curriculum Services, "The only requirement subsequent to enrolling in the lab course was that students pass that course-- period. They were not examined again." (401)

Virtually all of the students successfully completed the lab courses, and by doing so they fulfilled the district's minimum competency requirement. This practice continued until the late 1970s, at which time the Board of Education expressed dissatisfaction about existing graduation requirements, including the testing program:

"Our board felt that simply requiring youngsters who had failed either of those two tests to take a one-semester course wasn't stringent enough." (421)

As a result of this concern, a city-wide committee of citizens, employers, and university academics was formed in February, 1977, to examine the current graduation testing program. At the same time, an internal committee of district staff was created to work with the citizens' committee. The internal committee was composed of the mathematics and reading coordinators, the Director of Testing, a representative from the special education department, principals from the three levels, the Director of Curriculum Services and the Director of Elementary Education. No teachers served on the internal committee.

Together, the citizen and internal committees made two recommendations. The first of which was to make changes in the current graduation testing program: select new tests and determine rigorous standards for them. The second recommendation included a plan to implement a K-8 competency testing program. Both committees had expressed concern with waiting until the last years of high school before administering competency exams. The proposed elementary testing program was seen as a means of abolishing social promotion and avoiding eleventh-hour notices of student incompetence. In the following sections, we will describe events and activities related to the committees' first recommendation. Descriptions related to the second recommendation will be made only as they contribute to an understanding of the graduation testing program.

Development of the High School Competency Exams

The Curriculum Services department was charged with the development and standard-setting of all competency tests, elementary and secondary. Throughout this department, committees of teachers were formed for elementary grade combinations of K-1, 1-2, 2-3, and so on. Their task was to agree on the minimum objectives necessary for promotion to the next grade. The high school teachers followed a slightly different procedure. A committee of secondary teachers was asked to indicate what they would rather do: use the existing graduation tests or adopt new exams. The teachers elected to adopt new examinations; the reading teachers would choose a commercially-developed examination whereas the math teachers would develop a new competency test.

The Director of Curriculum Services described how the reading test was selected:

> There were meetings of the city-wide language arts council. They decided that the best way to proceed would be to adopt an existing reading test rather than take the amount of time that would have been necessary to write and pilot a test. They looked at several tests and the SHARP was chosen because the council felt it was more closely related to application skills. Also, the objectives on the test were judged to be appropriate. (421)

The Director of Testing explained that the publisher of the SHARP provided the district with a five-page report (4B) which outlined the test's characteristics, technical and otherwise. This report would serve as a guideline for

development of the graduation math test. According to the Director of Testing:

> When the SHARP publishers provided us with that report, it became very apparent to our math people that a similar document should be produced for the math test. At the time, we were pretty concerned about potential legal action and so we should have some comparable documentation for the math test. (423)

The current math coordinator described the development of the Inventory of Mathematics Competencies (IMC):

> All teachers who wanted to be involved participated in the process. They selected 25 objectives that were representative of what students needed to know for competence in mathematics. Having selected those 25 objectives, then the decision was made that a test would be built. We started out by having groups of teachers meet and write a large number of items. These items were combined into pilot tests and sent into the schools. (402)

After the pilot tests had been completed, the math department studied the items to determine which would be most appropriate for inclusion in the test. The math coordinator explained how items were judged:

> We studied items using a foil analysis. For a competency test, we chose to assume that 90% of our students should be able to get the item correct. At the same time, the three foils or suggested answers should have been reasonable; something there for the student to choose from. (402)

The items ultimately selected were piloted a second time and revised before the final form of the test was assembled. After the internal validation process, the IMC was critiqued by consultants from state universities and the department of education. In addition, a bias study was conducted by faculty at a nearby university following the second year's administration. Test results were analyzed for ethnic group

and gender bias. The Director of Curriculum Services recalled the study's conclusions: "The report indicated that the tests were free from bias, as nearly as tests can be." (401)

With the reading test selected and the math test developed, the internal committee would begin the task of determining performance standards.

Setting Standards: By Test Objectives or Items

In addition to setting standards for the graduation exam, the internal committee had to determine performance standards for the elementary tests. The committee's standard-setting process was described by the Director of Curriculum Services:

> We solicited the views of university types and other schools, looked into the literature, and we even asked some of the test manufacturers how they would have gone about it. We arrived at what we felt were fair and equitable kinds of standards for those tests. (401)

For the elementary and graduation exams, the committee decided to require students to pass 100% of the test objectives. The number of correct items necessary to pass an objective would vary from test to test. In the high school math test, each objective was measured by three items. Students were to answer at least two of the three items correctly to pass each IMC objective. On the SHARP, each reading objective was measured by four items and students were to answer at least three items correctly to pass each objective. The decision to require passage of all the objectives on each test stemmed from the committee's conviction of what the term minimum implied: "If in fact the

skills are minimum, the youngster should reach 100%. A score below this, derived from cutoff methods, wouldn't be a minimum." (401)

The same standard would be applied to the elementary competency tests. Students would be expected to pass all of the test's objectives before they could be promoted to the next grade. The internal committee provided an alternative to the standard on the elementary tests; an exception that would not be allowed for the graduation exam. According to the Director of Curriculum Services,

> The elementary teacher was in the best position to know whether or not the youngster knew the objectives by the end of the school year. There were five or six weeks left in the school year after teachers received the competency test results and so they could continue to teach the objectives in which the youngsters were deficient. So we wanted a fail-safe kind of a mechanism by which the teacher could overrule the test results. (401)

In a follow-up interview, the Director of Curriculum Services elaborated on this fail-safe mechanism for the elementary tests:

> Let's put it this way: A youngster comes into your class and takes the test. You are the teacher and you have had this youngster in class now for seven months. You know that his youngster knows a specific objective but he missed it on the test. Do you trust the test? In other words, did he have a bad day? Did he misread the items? Did he mark his answer in the wrong place? There are many places for a kid to make mistakes on a test. I am not talking about the teacher guessing if the kid knew the objective. I am saying the teacher absolutely knows that the kid knows the objective. He has proved that all year, but he just made a mistake on the test. Now which one do you go with? I go with the teacher. What did we do before we had competency test? The teacher's judgment. (421)

Committee members did not extend the same fail-safe mechanism to the graduation exam. They felt that departmentalization at the secondary level prevented high school teachers from being as familiar as an elementary teacher with individual student competence. Secondary teachers were limited to seeing students one period a day whereas elementary teachers spent nearly the entire day with one group of students. In contrast with the elementary grades, no one high school class focused on reading skills and so committee members felt that secondary teachers, in general, would have more difficulty certifying a student's competence in reading. Consequently, the committee devised an alternative fail-safe for the graduation exam.

> Since the teacher was not going to have the opportunity to certify that the youngster knew the objectives for high school, holding the student to the 100% standard did not seem to be fair. That is why we came up with the second criterion. (401)

The alternate standard was set at 70% of the test items correct:

> We arrived at that by considering a couple of things. First of all, most competency programs at the time were using something similar to a 75% criterion. Second, we wanted some even number that would have required youngsters to get a few more items correct than they needed according to the first standard. With that, a student needed to answer 67% of the items correct, so we just said it should be 70% of the items if we are going to disregard the individual objectives. (401)

The publisher's technical summary of the SHARP was available to the committee's during their selection of the alternate standard. A section on cutting scores was included in the 1978 report:

> The publisher suggests that a cutting score of 84 be used, which represents a 70% proficiency level. A similar cutting score was adopted by the Segovia public schools as a result of a review of the test content by curriculum directors, a review of the test by the Segovia School Board, and because a student would have to attempt all three sections of the SHARP to pass. A 70% proficiency level is a widely-used and popularly recognized passing level in American education. (4B)

According to the Director of Testing in Lucerne, the preceding section influenced the committee's choice of 70% as the alternate standard.

> I think that when people on the committee saw that standard in print, a lot of them jumped on it. As we fought our way through the standard-setting process, I think some people felt like they were out on a limb-- setting standards was somewhat of a new concept. (423)

When asked if there was any discussion about setting the standard higher than 70%, the Director of Curriculum Services explained:

> There was some talk of that, but I think we arrived at that one very swiftly on the basis that 70% is better than 67%-- the percent of correct items necessary to pass all of the objectives on the math test. (401)

With the tests developed and standards set, the internal and citizen committees delivered their recommendations to the Superintendent who would make the final decision. The Director of Curriculum Services and testing coordinator provided the Superintendent with the background, rationale, and steps taken in the decision-making process. The Superintendent did not make any changes in the committee's recommendations before approving the standards. The Director of Curriculum Services reflected on these events:

> The Superintendent would have had a difficult time trying to overrule any recommendation the committees made because they were two pretty high-powered committees. I

> think that the work proceeded slowly enough, and yet with enough dispatch, to assure that the committee was producing very good recommendations. (401)

The Lucerne Board of Education listened to reports by the Superintendent, as well as Directors of Curriculum Services and the Testing. The Board expressed concerns about potential fail rates on the graduation examinations, but because both tests had yet to be piloted on the same set of students, the administrators could only offer estimates; estimates that ranged from 0 to 30% of a graduating class. After a number of city-wide meetings were held to inform the public about the elementary and graduation testing programs, the Lucerne Board of Education formally adopted the programs in Spring, 1978. The Class of 1981 would be the first class held to the new graduation requirement. These students would first take the examinations during their sophomore year-- Fall, 1978. Following district re-organization in which the ninth grade was added to the high school grades, later classes would first take the graduation tests during their freshman year. We will describe a number of other policies and practices before we turn our attention to some consequences of the high school testing program.

Policies and Practices of the Graduation Testing Program

1. Students must take the IMC and SHARP the spring semester of their freshman year.

2. Students may retake failed exam(s) once a semester and if enrolled, once during summer school until they pass.

In sum, students may have up to eleven chances to take the graduation tests.

3. If students haven't passed an exam by their junior year, they must enroll in a competency class for that subject. They must continue in the class until they pass the exam.

4. Parents are notified in the fall that students will be held responsible for passing the objectives in reading and mathematics as part of the graduation requirements.

5. Special education students are required to take and pass the competency exams unless their Individual Education Plans (IEP) specify otherwise.

6. Students cannot graduate without passing the competency exams. However, passage of both does not ensure graduation. Students must complete the necessary credit hours and attendance requirement of seven semesters.

Results of Lucerne's High School Testing Program

Following the first administration of high school tests, district administrators estimated the ultimate number of students who would fail the examinations. Looking at the first-round passing rate and guessing that at least 50% of the students would pass on subsequent retakes, these administrators estimated that approximately 1% of a class of 2000 seniors would ultimately fail the graduation examinations-- 20 students. (421)

The Director of Curriculum Services described the actual fail rate of the Class of 1981:

> We projected that we were going to have twelve out of nearly 2000 youngsters who would not receive their diplomas as a result of the competency exam. But before we reported that to the board, we asked the high school counselor or principal to look back and see if those youngsters also had the necessary credits. Ten of the twelve did not have the credits, so it wasn't just the competency test that was holding them back. The other two students were special education students and really shouldn't have been held to the testing question. As a consequence, no one was denied a diploma due to the test alone. (401)

Current pass rate data were provided by the Director of Testing. Table 5.1 displays the pass rates for each class in Spring, 1986.

Table 5.1

Passing Rates for Grades 9-12

Grade	Number in grade[a]	Math	Reading	Both tests
9th	2,491	62%	68%	58%
10th	2,107	81%	86%	80%
11th	1,752	88%	90%	88%
12th	1,413	93%	93%	93%

[a]number includes special education students

While the classes in Table 5.1 reflect the performance of different cohort groups, we can see that pass rates gradually increase in the four years. The Director of Testing indicated

that of the 103 (7%) who had not passed the graduation exams by the end of their senior year, approximately 24 were regular education students. The remaining students were special education and exempt from the requirement. These 79 or so students may or may not have taken the graduation tests and so they are not included in the ultimate pass rate. The district did not keep data about minority pass rates on the high school graduation exams so we are unable to examine performance patterns by ethnic groups.

According to the Director of Curriculum Services, drop-out rates do not appear to have changed appreciably in the last ten years. He indicated that the drop-out ratio among blacks and whites was approximately 60-40. Given the population ratio of 67-33 of blacks to whites, respectively, he felt that if there was a disproportionate drop-out rate, whites were at a slightly higher rate. He indicated further that there was no evidence that the high school testing program affected the drop-out rates:

> The Board of Education continues to express concern that testing may be causing some dropouts, without any data to suggest it. We don't know how to provide that kind of data: how can we say what is cause and what is effect? (421)

The high school graduation tests seem to be woven well into the fabric of high school life and requirements. According to the high school counselor at one of the four high schools, a magnet school with a large bilingual education program, the tests are less imposing than those at the elementary level:

> The tests are not a big threat; they don't hang over anyone's head. Unlike the elementary tests, the high school tests have nothing to do with being promoted from grade to grade. So when youngsters enter the ninth grade, take the tests and fail, they get a little upset when they don't pass. I have to reassure them that this it where it changes. I tell them that they have to be able to pass the test some time before they graduate, but if they don't pass it the first time, they will not have to repeat ninth grade. (424)

Since the initiation of the graduation tests, no student has been denied a high school diploma on the basis of the tests alone. According to this same high school counselor:

> We have not been confronted with holding a student back from graduation because of the test. Usually it is a combination of things: not enough credits or not enough perseverance to blast through. (424)

The Class of 1986 was no exception to this pattern. According to a document produced by the Office of Testing, all 1986 seniors who earned sufficient Carnegie credits to graduate passed both of the competency tests and were awarded diplomas (4D).

CHAPTER VI

SETTING STATE STANDARDS FOR HIGH SCHOOL GRADUATION: REFORM IN VICTORIA

A Glimpse of the Present

The state of Victoria requires students to take and pass the Victoria Examination for High School Graduation (VEHSG) before they receive their high school diploma. The VEHSG was developed by Victoria educators after identifying fifty-seven competencies in three subjects. These competencies were judged to be "critical skills to be mastered by students prior to their being granted a diploma." (5A) Three sub-tests in reading, language, and mathematics were developed to measure these critical skills and students must pass each sub-test to graduate. Mandated by the Victoria State Board of Education, this testing program is one of three programs in operation at the state level. The other two programs involve minimum competency testing of elementary students and prospective teachers seeking certification. While the focus of this case study is on the high school graduation exam, some attention will be directed to the other programs as they contribute to an understanding of the VEHSG.

A number of similarities can be identified among the five sites in this series of case studies. The very nature of competency testing goals and practices limits the possibilities for radically innovative or unique programs. Nevertheless, differences among the sites can be discerned:

one such difference is the manner by which standards were determined. Specifically, some standards were set using psychometric cut-off methods while others were based on prevailing notions of what represented competence. Whatever means were employed in the decision-making process, it is clear that political or practical concerns were mitigating influences. Victoria is the exception among these five sites. Those charged with the task of setting standards on the VEHSG were able to proceed with apparent unparalleled technical and professional discretion; they were mindful of practical implications, but felt unfettered by political demands. Much of this case study will be devoted to a description of how these decision-makers determined the standards for the VEHSG. From this description, it will be seen that the standard-setting process operated in a textbook-like context, where technical cut-score methods were employed with professional discretion. The remainder of the Victoria case study will focus on two issues, the first of which will be a description of the various consequences related to the high school testing program. Second, we will look at some of the circumstances and events external to the VEHSG standard-setting process. We will examine reasons why the standard-setting process and product were free from mitigating circumstances.

A Brief History of Events Leading to the VEHSG

Before the standard-setting process is described, a brief history of events will be presented. In February, 1976, the

Commissioner of Instruction appointed a task force to examine the then current high school graduation requirements and consider possible changes. Over a year later in April, 1977, the Victoria State Board of Education mandated that a statewide committee of educators and lay people develop minimum competencies for high school graduation and the means for measuring them. The committee was to make a final report to the State Board in July, 1978. At this same April meeting, the Board tabled the pending report from the Commissioner's task force on graduation requirements until the newly formed committee produced its final recommendations.

The committee of 50 educators and 50 lay people, known as the Professional and Citizen Committee (PCC), met from August, 1977, through July, 1978. In October, 1977, the steering committee of the PCC met and produced the following recommendations:

1. Create a task force to determine competencies in reading, language and mathematics.

2. Delay high school graduation testing until tests at three grade levels within the K-9 system could be implemented as checkpoints.

3. Develop the tests in-state so they conform to the competencies and expectations specific to the state of Victoria.

From November, 1977, to June, 1978, a sub-committee of the PCC met to draft competencies for the three elementary checkpoints. After securing reactions from local

superintendents and administrators, the competencies were revised and sent to the PCC. Eight public hearings were held across the state before final approval of the competencies was given in October, 1978. Three months earlier, the PCC reported to the State Board on activities to that point. The PCC advanced the steering committee's recommendation that the State Board implement the elementary checkpoints testing before instituting a graduation exam. Further, the committee recommended that progress be assessed initially in grade 11, with a final assessment in the senior year. The committee also recommended consideration of exceptional children: some type of diploma or certificate should be awarded to them.

For the next three years, the checkpoint competencies were piloted throughout the state. In addition, a professional committee was formed to develop corresponding checkpoint tests. The checkpoint tests were ready for initial administration in the fall of the 1981-82 school year.

During the piloting of competencies and test items for the elementary checkpoints, a number of events occurred that were related to high school graduation. In Spring, 1980, the Commissioner's task force entered the picture again, recommending that a minimum number of academic credits and demonstration of acquired knowledge be required for graduation. Taking the recommendations from the task force and those made earlier by the PCC, the Victoria Board of Education adopted revised graduation requirements for students entering the ninth grade in Fall, 1981. Work on the VEHSG

began immediately. The VEHSG competencies were to be the same as those from the ninth grade checkpoint test. From those competencies, item specifications were developed and multiple-choice questions were written and revised. Following a bias review, the VEHSG was ready for pilot-testing in October, 1982. Before the second pilot the following spring, the Victoria Board of Education formally adopted the VEHSG as the official high school graduation examination. Nearly two years had been devoted to the development of the VEHSG; numerous state educators participated in the process; and the resulting product was one in which all could take pride.

The standard-setting process would follow similar lines. Standards were to be set on each of the three subject matter tests and students would have to meet each of those standards to receive the diploma. The assistant director of the state department division responsible for the VEHSG explained:

> If you give students a diploma and put them out in the world, saying that they are competent, they need to be competent in each of the three areas-- not just partially competent in some. If a student can't read, he can't read. If he can't do computational math, he can't do computational math and therefore, he should not get his diploma. (508)

Aware of the many technical concerns involved in minimum competency testing, the Victoria Department of Education formed an advisory committee. (5C) The committee's primary charge was to determine the standards that Victoria high school students would be required to meet before receiving a high school diploma. A description of the committee's efforts to fulfill this charge follows in the next section.

Setting Standards: By the Book

The VEHSG Technical Advisory Committee (TAC) was formed in Fall, 1982. The committee was composed of faculty and administrators from several of the state's universities, as well as employees of the Victoria Department of Education. All had backgrounds in testing and measurement and many had extensive research experience in competency testing and standard setting. When the TAC met for the first time, members were assured that political issues would neither limit nor constrain them in their task of determining standards:

> Most of us took our charge in a scientific manner. These meetings were not political. We were told up front by either the Commissioner or the divisional director in the State Department of Education that they wanted dependable cut scores that would hold up in court. And I argued that courts look for empirical standard-setting processes backed up with technical expertise. Political decisions should not be a part of those processes. (507)

Since the decision-making process would be spared political interference, the TAC members were free to exercise their professional and technical judgment in determining standards on the VEHSG. In the words of one TAC member, "We could apply our professional knowledge in this [non-political] setting, almost unbridled." (503)

One of the committee's early decisions was to review and consider the available techniques for producing cut-scores. Following a presentation of the available methods by various members, the TAC classified them into three types: judgmental, empirical and theoretical. Judgmental methods would produce passing scores by estimating the performance of

a hypothetical minimally competent student on the VEHSG. Empirical procedures would involve studying the performance of examinees on the VEHSG and selecting the passing score from that information so as to produce a desired proportion of passes and failures. A standard derived from theoretical methods would be based on probability theory in which there was a specified (and small) probability that the score was not earned by random guessing.

The committee members agreed that no single standard-setting procedure addressed all of their concerns, and so they decided to use one or two procedures from each category. (5C) There was only one theoretical method and so the committee approved its use. A modified version of the Angoff method was chosen from the judgmental category while two methods were selected from the empirical: contrasting groups and borderline. In the modified Angoff method, judges would examine the items for each skill and estimate the proportion of those items the minimally competent student should answer correctly. (5C) For the two empirical methods, teachers would be asked to identify students as masters, non-masters, or borderline with regard to the competencies tested. After taking the test, distributions of each group's scores would be examined. In the contrasting group method, the test scores of masters and non-masters would be compared and the passing score would be the one that best separated the two groups. In the borderline method, the passing score would be the average of the borderline students' test scores. Using the four cut-

score methods (theoretical, modified Angoff, contrasting groups and borderline) would produce four passing scores for any given test within the VEHSG. The TAC now faced the task of reconciling these various scores to produce a single passing score for each of the VEHSG reading, mathematics and language tests.

After considerable discussion of the advantages and disadvantages of the four cut-score methods, the TAC made two decisions. Since the theoretical method would produce a non-chance passing score, the committee decided to use the resulting score as the benchmark for procedures from the judgmental and empirical frameworks. Any method, empirical or judgmental, that produced a standard lower than the theoretical passing score would not be considered by the TAC. The second decision dealt with combining the results of the judgmental and empirical methods. The committee reasoned that the distinction between the two types was superficial. Both types of method involved judgments of some kind; judgments rendered about test items or examinees. In the words of one TAC member,

> We saw that both [types] were judgmental procedures. It was just that the judgment was being made on the items in one case and on individuals in the other case. We felt that they were both legitimate and valid judgments and we could probably get a better estimate of what the final score should be by taking both types of judgments into account. (505)

Another TAC member elaborated on the committee's concerns:

> I think we were trying to get the teacher judgment about the items and take into account the empirical

> information that we had from teachers about the students. We felt like we could combine the information and come up with a stronger pass score. (504)

To take the two types of information into account, the TAC elected to average the scores produced from the modified Angoff, contrasting groups and borderline cut-score methods. If any method produced a score lower than the theoretical standard, that score would not be included in the average.

In Spring, 1983, the TAC supervised the determination of the cut-score methods. Theoretical passing scores were calculated separately for the reading, language, and mathematics tests. When the items on each test were piloted a second time, teachers were asked to indicate whether each examinee had enough knowledge of the subject to graduate from high school: yes, no or questionable. The test scores of students in the first two groups were analyzed as part of the contrasting groups method. Test scores of those in the questionable group were used in the borderline group method.

For the modified Angoff method, reading, language and mathematics teachers were asked to estimate the percentage of items the minimally competent students should be able to answer correctly. Table 6.1 summarizes the results of the four cut-score methods.

Table 6.1

Results of the VEHSG Standard Setting Methods

Test	Raw scores and percent of items			
	Theo-retical	Modified Angoff	Contrasting groups	Border-line
Reading	49 (61%)	60 (75%)	53 (66%)	64 (80%)
Math	57 (60%)	62 (65%)	40 (42%)	61 (64%)
Language	70 (59%)	96 (81%)	60 (50%)	87 (73%)

During the standard-setting study, members of the TAC identified a number of problems with the contrasting groups method. After inspecting the distribution of scores for the masters and non-masters, the committee determined that there were difficulties in identifying a single score that best separated the two groups. Further, the number of students identified as non-masters was so small that the TAC questioned the stability of any score selected to separate them from the masters. Finally, two of the three passing scores produced from this method were lower than the theoretical standards. As a result, the TAC dropped the contrasting groups method from further consideration and proceeded to average the remaining passing scores for the final VEHSG standards, shown in Table 6.2.

Table 6.2

Final VEHSG Standards

Subject area	Number of items	Passing score	Percent of items
Reading	80	63	79%
Mathematics	95	62	65%
Language	119	92	77%

Until the TAC had determined the ultimate passing scores for the VEHSG, information about potential passing rates (percents of pupils passing) had been suppressed. Once the standards were set, the committee agreed to apply them to the pilot-test results and estimate the passing rates. According to one TAC member,

> We looked at passing rates for various categories of students to see what the impact [of the standards] would be. On the basis of the pilot, it looked like there would be a fail rate of 25% for certain categories of students. (502)

Because no consequences were tied to the pilot test administration, the TAC felt that students probably did not try as hard as they would when real penalties were attached to test performance. As a result, the committee members expected the actual fail rate to be lower than the fail rate of 25% on the pilot test; how much lower was anybody's guess. Whatever the fail rate would be, the TAC was adamant: "At no point did

we consider changing the standards when we saw what impact the scores might have." (505)

Satisfied that reasonable, technically-sound, and legally-defensible standards had been set, the committee's next step was to deliver its recommendations to the Commissioner of Education and eventually, to the Victoria Board of Education. Members of the TAC and division managers from the Department of Education met with the Commissioner of Instruction and the state's attorney in Summer, 1983, to inform them of the committee's procedures and decisions. Step by step, the standard-setting process was detailed and explained; the passing scores were presented; and potential consequences were discussed based on the pilot test data. Following the presentation, Commissioner Zwei had one question for the TAC: 'Could the committee members stand up in court and defend their decisions?'. The answer was 'yes'. In return, Commissioner Zwei dispelled any concerns that the carefully set standards would be changed in response to anticipated fail rates:

> After we showed him what the results would be, Dr. Zwei said, 'I don't want anything based on how many kids are going to fail. Do it right. We're not going to adjust the cut score based on x number failing'. Everyone agreed that was hedging the issue. You either believe it's a minimum competency test and is good for kids, or you play a game with the public by saying that we're going to let this many pass the exam regardless [of competence]. To us, that was philosophically, as well as morally, incorrect. Commissioner Zwei, without any coaching at all, totally agreed with that. (501)

The Commissioner brought the TAC's recommendations bearing his approval to the Victoria Board of Education. The

committee's actions were explained in lay terms, carefully avoiding technical jargon. With Commissioner Zwei as moderator, members of the TAC were present to field any questions the Board had about how the VEHSG standards were set. According to a Department of Education manager present at the meeting, the Commissioner was specific and emphatic--the standard-setting process, while technical in nature, produced fair passing scores; scores that separated the students who did learn the basic skills from those who did not. According to this same person, the Board responded positively:

> We were amazed how few questions the Board members had. There were two or three questions, but they were legitimate kinds of questions. I thought the Board was very positive. I think they were anxious to get on with the graduation exam after the years of work on it. (508)

The TAC recommendations were thus approved by the Board and formally incorporated as the standards for the graduation examination, bringing the one-year standard-setting process to an end. At no time were the standards modified in light of anticipated or actual passing rates. Indeed, political pressures to change the standards were absent or at least, unarticulated. How did the Victoria graduation testing program remain untouched by political maneuvering when all of the other sites in this series of case studies were affected in some way? Before we finish telling the story of the VEHSG and its consequences, we will examine some conditions that helped to set Victoria apart from the other four sites.

Extenuating Conditions Surrounding the VEHSG

Clearly, standards on the VEHSG were set in a technical fashion without political interference or normative modifications. Officials at the Victoria Department of Education and members of the Technical Advisory Committee agreed that one condition primarily contributed to the apolitical environment in which they worked: the mandating agency for the VEHSG. In a document prepared by the Coordinator of Victoria's student testing programs, the mandating agency was said to play an important role in keeping politics at bay.

> Whether a graduation examination is legislatively mandated or state board mandated will have a major effect on the program. Legislatures often demand that the program be implemented in a short period of time; which results in hasty decisions that may have legal ramifications. . . The State Board of Education is the mandating body for the Victoria Examination for High School Graduation (VEHSG). As a result, implementation has had the advantage of being deliberate, enabling the State Department to take the lead with extensive educator involvement. (5D)

While the State Board of Education could not provide substantial remediation dollars, it could give license to the Department of Education to produce a defensible testing program, without the pressures certain to be applied by an impatient legislature. Another Department of Education official, the assistant director of the division responsible for student testing, expressed his thoughts about the mandating agency,

> I've been to various states to talk about the VEHSG. In one state someone asked, 'Which do you think is best: having your program mandated by the legislature or the

> state board?'. My response was this, 'Well, in our case, it was better to be mandated by the state board because it gave us the opportunity to do it the way we wanted it done'. On the other hand, there is an advantage to having a legislative mandate: you get money for remediation. (508)

A TAC member agreed that the Department of Education was free to set its own course,

> We didn't have any pressure from the state department. I think that the people from the state department who served on the TAC felt a nurturing attitude by the state department. They didn't feel any pressure to come up with any score that was going to pass a certain number of people. (505)

By many accounts, the apolitical environment for the VEHSG development, standard-setting and implementation was due to the mandating source: the Victoria Board of Education. Another set of events and conditions may better explain this environment; events and conditions connected to a different state testing program.

In the late 1970's, teacher testing was a popular issue among Victoria legislators, Board of Education members, and the Commissioner of Instruction. The Commissioner made his stand clear when in November, 1978, he stated that teachers should be tested for minimum teaching competencies. Further, he felt that such testing loomed on the horizon. Less than a month after he made these comments, the Board of Education discussed the merits of testing prospective teachers in Victoria but delayed any action until January, 1979. During the January meeting, two legislators spoke with Board members. They announced that if the Board did not make some provision for teacher competency testing, the legislature would. At

this same meeting, the Board reconsidered and passed the tabled resolution. Students graduating from state teacher education programs in 1981 would be required to pass a competency test to be eligible for certification. The particular competency test was not named in the Board's resolution, but sentiments ran high for a state-developed examination.

The teacher testing issue remained in the political arena despite the Board's action. During the next two legislative sessions, bills were introduced that would have thwarted the Board's authority and efforts to implement its mandate. Specifically, the bills called for an earlier start-up date and named the National Teachers Examination (NTE) as the test to be used. Members of the Board of Education and Victoria Education Association (VEA) opposed the use of the NTE, claiming that the test would not reflect curriculum specific to Victoria. While neither of the bills was enacted, sufficient pressure may have been applied to prompt the Board of Education to complete its test.

To this end, the Board and Department of Education contracted with a private testing firm in mid-1980. The firm was to develop a test that measured teaching skills relevant to Victoria. In June, 1981, the test was administered for the first time and approximately 80% passed. Although the overall passing rate was quite high, minority passing rates were disproportionally low; minorities were half as likely as whites to pass. Indeed, graduates from four institutions with

high minority enrollments passed at very low rates. Table 6.3 is a display of initial passing rates at these four institutions.

Table 6.3

Initial Teacher Test Passing Rates at Select Minority Institutions

School	Passing rate
A	0%
B	14%
C	22%
D	34%

Alarmed at such low and disparate rates, members of the Board of Educations sought solutions. Before any action could be taken, the Board of Education and Commissioner of Instruction were greeted with a lawsuit. The plaintiffs claimed that the Victoria teacher test was biased and invalid. Litigation and related activity would proceed for a number of years and thousands of dollars would be spent by the state to defend itself.

Litigation was not the only problem that plagued this State Board mandated test. Examinees were failing a number of subject-matter tests at such a high rate that Victoria faced a shortage of certified teachers in these areas. To correct this shortage, several cut scores were lowered by as many as

ten points. With the hurdles lowered a notch or two, more teachers were eligible for certification and able to fill positions where they were urgently needed.

Although the Victoria teacher test was mandated by the State Board of Education and not the state legislature, it was, nevertheless, plagued by political, legal, and practical problems. As a result, we are wary of assertions that the State Board of Education as the mandating agency kept the VEHSG from falling prey to political interference. Such claims are not false per se; clearly they are credible and reasonable to our informants. However, the assertions do not tell the whole story. They fail to illustrate the complexity of testing in Victoria. We believe that the Board of Education and Commissioner of Instruction learned some very hard lessons from the teacher testing experience, lessons that account for the apolitical environment for the development of the VEHSG. To avoid a situation similar to that associated with the teacher testing program, the VEHSG must be kept from the reaches of political hands and free from the sharp criticisms leveled (in and out of court) at the Victoria teacher test. We contend that the Commissioner of Instruction, with the Board's support, would set the tone for the development of the VEHSG and subsequent standard-setting. He would encourage professional discretion by the Department of Education division responsible for the VEHSG development. He would support technical efforts to derive defensible standards. The division and TAC would be sheltered from

political storms in return for a test and cutscores that would stand up to close scrutiny and possible litigation. Indeed, the VEHSG has not been subject to the same tribulations suffered by the teacher test. In the following sections we describe policies, practices and consequences related to the VEHSG that may well have contributed to its success in avoiding litigation.

Policies and Practices Surrounding the VEHSG

A number of policies and practices involving the VEHSG are outlined below:

1. Every ninth grade student in Victoria's public schools is given a brochure explaining the examination requirement and that students will have their first opportunity to pass the VEHSG the fall of their junior year. Students and their parents must sign the brochure and return it to the school, where it is included in students' cumulative records.

2. If special education students wish to earn a regular high school diploma, they must take and pass the VEHSG. No reading of any portion of the VEHSG to any student is allowed.

3. If students fail any of the three tests in the VEHSG, they may retake the test(s) the following semester. Students have four opportunities to take the VEHSG by the spring of their senior year, but may take the exam as many times as they desire after their senior year.

4. Local school districts must provide remediation at their own expense for students who fail the VEHSG. The Victoria Department of Education requires that school districts document students' test performance and district efforts to remediate.

Consequences of the VEHSG

As the initial test administration date neared, local school districts sought to prepare their students. According to the Coordinator for the K-12 testing programs,

> There was so much tension, so much fear of the unknown, that a lot of high schools set up reviews [for the test]. Many had two or three weeks of reviews before the first administration. (501)

The schools weren't shooting in the dark about what might be on the VEHSG. The Department of Education provided detailed information on each of the test items: the skills tested, sample items, and examples of incorrect options. In addition to test item information, the schools were familiar with the ninth grade checkpoint test, the basis for the VEHSG. As a result, schools were able to provide instruction specific to the VEHSG. Preparatory reviews have dwindled since the first administration of the VEHSG. Tensions have diminished. Teachers know what will be on the test, students are said to know the skills, and nearly every one passes the test.

In general, of one hundred first-time takers, ninety-two will pass all three tests in the VEHSG. Of the eight who fail and retake the exam the spring of their junior year, about four students will pass. The pass rate continues at 50% on

future retakes, leaving one to two students out of one hundred who have not passed the VEHSG by the end of the senior year.

The Victoria School Journal published passing rates on the VEHSG subtests which are detailed in Table 6.4. These data reflect test performance of the Class of 1985 on the first administration of the VEHSG in Fall, 1983, and the last regularly-scheduled administration in Spring, 1985. Passing rates for regular education students are the first figures provided, accompanied by passing rates for the entire test-taking population (regular and special education students) in parentheses.

Table 6.4

1985 VEHSG Passing Rates

VEHSG subtest	Fall, 1983	Spring, 1985
Reading	94% (90%)	99.5% (98%)
Mathematics	89% (85%)	98.9% (97%)
Language	87% (82%)	98.9% (97%)
Entire test battery	-	98.4% (96%)

Note. Unreported data indicated by -.

A Department of Education official described passing rates for the next graduating class,

> The Class of 1986 had about 52,000 students (including special education students). There were about 1,500 [2.8%] students out of the 52,000 who didn't pass

> the graduation exam. Of these 1,500 students, 480 [0.9%] were non-special education. Many of that 480 wouldn't have graduated because of too few Carnegie units. (502)

The specific proportion of regular-education students lacking the necessary Carnegie units among those who failed the VEHSG was unavailable from the Department of Education. According to the Coordinator of the state testing programs, such information is kept only at the local district level and not aggregated at the state level.

Ultimate passing rates for special education students are harder to discern, but this same Department of Education informant provided his best estimate:

> In general, of the [special education] students who start taking the test in the fall of their junior year, about 68% of those students (collapsed across all twelve exceptionalities) wind up passing all three sections. (502)

This individual added that the particular label of the special education student would probably influence the observed passing rates on the VEHSG. Specifically, students classified in less severe categories are likely have higher passing rates than those labeled as having severely limiting exceptionalities.

Minority passing rates on the VEHSG have been lower than Anglo passing rates. The VEHSG had been subjected to bias reviews during its development and during the pilot stages, a study was conducted to examine potentially-biased test items. In this study, item difficulty indices were compared between minority and Anglo students with comparable achievement levels. Items with discrepant difficulties were identified

and then revised or eliminated. TAC members expressed confidence in the various bias checks during test development and piloting; many agreed that the discrepant passing rates were due to curriculum and learning, and not to test bias. (505, 506, 509)

Local school districts are required by the state of Victoria to provide remediation to students who fail any of the three tests in the VEHSG. The Coordinator for the state testing programs explained:

> Victoria had no money for remediation so it was very easy to say [to the local districts] 'You are responsible. You still have the same number of students and if they have a deficiency, it's your problem.' So every high school in the state will remediate differently. In the beginning when we had larger numbers of students failing, the most popular thing in some areas was to set up competency classes; courses that would focus on basic skills. Now, rather than setting up classes, a lot of school systems have set up skills periods during the first thirty minutes of the day or during study hall. (501)

Since students were not failing in large numbers as time went on, competency classes were reduced to skills periods in many schools. Students needing work in specific competencies, say fractions and decimals, would gather for intensive instruction during these skills periods.

As indicated earlier by the testing coordinator, schools do not receive dollars from the state for remediation:

> At first people said, 'How can we do this with no money?'. After the first and second administration of the test, no one questioned us. They found that it wasn't going to be as bad as they thought. We didn't have 40% of our kids failing the math section. As soon as the fear was gone, we never got asked [about money] again (501, cp. 204).

Department of Education officials and TAC members attest to other consequences. Curriculum and instruction viewed as tighter, more directed. Teachers and students are said to be aware of the expectations and perform accordingly. Another consequence of the VEHSG involves strengthened ties between the Department of Education and faculty at the state universities. Implementation of the testing program, standard-setting, and validation studies were all conducted by faculty and their graduate students. The VEHSG project provided valuable experience for students, brought monies to the universities, and introduced new colleagues and publishing opportunities to the faculty. In return, the Department of Education was able to draw upon professional and technical expertise in creating a sound testing program. Indeed, the Victoria Department of Education, State Board of Education and Commissioner of Instruction are satisfied that standards have been raised and students have jumped the hurdle.

CHAPTER VII

SETTING STANDARDS FOR COLLEGE ADMISSION IN THE STATE OF BRITTANY

A Glimpse of the Present

Admission requirements to post-secondary institutions in the state of Brittany resemble criteria in other states. High school graduates seeking admission to any of the seven state colleges must complete fourteen units of college preparatory work in five subject areas. Required high school grade point averages vary by institution, ranging from 2.25 to 3.00. In addition, applicants must submit scores from the Scholastic Aptitude Test (SAT) or American College Test (ACT). No standard or minimum score is required on either of these tests; rather, the tests are used for guidance and placement. There is one other requirement for applicants, a criterion that sets the post-secondary institutions in Brittany apart from those in many other states. Applicants must score at least 30 on the National Written English Test (NWET), a standardized, multiple-choice examination. Only those seeking admission to the state's flagship institutions, the University of Brittany (UB) and Brittany State University (BSU), are held to this requirement. A third institution, a four-year college, requires applicants to pass this test or enroll in a remedial writing course at extra expense to the student; an option not available to university applicants. To those seeking admission to either UB or BSU, the message seems

clear: meet the standard for written English or look for another school.

In the following case study, we will describe and examine Brittany's efforts to raise standards in higher education. Specifically, we will chart the evolution of the NWET's initial use as a placement tool to its somewhat controversial use in admissions. In addition, we will show how economic considerations moderated the extent to which standards were raised in Brittany. The NWET admissions standard was determined largely by consideration of the number of students the institutions could afford to lose without incurring fiscal hardship. A higher standard on the NWET or the imposition of standards for other subjects (e.g., mathematics or foreign language) would have decimated the pool of admissible students and hence, invited economic disaster for the institutions.

Placing Students into Composition Courses: A Better Way

Writing skills and competence had been receiving considerable national attention during the mid-1970's. An increasing number of students across the nation were said to be pre-literate in national reports by Time magazine and other publications. Instructors bemoaned students' inability to write and the ranks of remedial composition courses were swelling. Across the nation, committees and blue-ribbon panels were convened to examine and deal with the writing crisis. Brittany was no exception to the trend.

An ad hoc committee was created to address competencies in English composition for students in the state's two- and four-year post-secondary institutions. This committee was charged with reviewing current institutional practices and recommending improvements. Specifically, the following tasks were assigned:

1. Review the means used by Brittany college and universities to determine the competency in English composition required for graduation.

2. Review the means used to waive requirements in composition.

3. Recommend means to be used by Brittany colleges and universities to determine competency in English composition more effectively and equitably.

4. Recommend means to waive requirements in composition. (2A)

Formed early in 1974, the ad hoc committee was composed of members representing the state's post-secondary institutions, community colleges, private institutions, and public secondary institutions. This committee would meet for well over a year and deliver its recommendations in June, 1976. These recommendations would be advisory only; they would not be binding on any of the state's post-secondary institutions.

Apart from the ad hoc committee's efforts, individual post-secondary institutions in Brittany were addressing issues related to writing skills. One such issue was placement in

university composition courses. Until the mid-1970's, the University of Brittany and Brittany State University had been using scores from the verbal section of the SAT (SAT-V) to place students in the freshman writing courses: Writing (Wr) 111 and 110. The former course was the regular freshman level class, the first of two composition courses required by each university. Students scoring low on the SAT-V were placed in Wr 110, the remedial writing course. By 1974, the two universities had formed different opinions about the SAT-V's utility in making placement decisions. Officials at BSU were satisfied with the SAT-V. According to the BSU Registrar:

> We had a lot of years of experience and success with using the SAT-V. We found a very good correlation over the years between the SAT-V and proper course placement. (226)

On the other hand, UB English faculty had grown skeptical about the link between performance on the SAT-V and writing skills. In the words of the former UB Director of Composition, who was also the chair for the state ad hoc committee,

> It was obvious to me that a low score on the SAT-V was not an adequate measure of writing skills because the test did not measure sentence-level skills that are crucial to separate out people who needed more work before they could enter a college writing course. (205)

Dissatisfied with the SAT-V, a search began at UB for a better placement tool. Serendipitously, a test publisher had contacted the UB registrar in hopes of securing UB as a norming site for a newly-developed test, the National Written English Test (NWET). The registrar realized that by permitting the norming study to be conducted at UB, he could

examine the NWET's utility as a placement tool in the composition courses. After contacting the Director of Composition and discussing the situation, the Registrar agreed to allow the test publisher to conduct the norming study at UB.

In Winter, 1975, the Registrar began a review of UB placement in the two freshman composition courses, Wr 111 and Wr 110. After sampling and testing students in the two courses, NWET scores were compared with class placement and instructor judgment. Tables 7.1-7.3 appeared in the Registrar's report. Table 7.1 shows the distribution of NWET scores by course grades of students enrolled in Wr 111. Table 7.2 displays the distribution of NWET scores by course grades (pass/fail) of students enrolled in Wr 110. Table 7.3 gives the distribution of NWET scores of students who were judged to have been misplaced in either of the two composition courses.

Table 7.1

NWET Scores and Wr 111 Grades

NWET Score	A	B	C	P	N	I	U	#
60	80%	20%	-	-	-	-	-	5
55	33%	38%	10%	10%	5%	5%	-	21
50	22%	46%	19%	3%	3%	3%	5%	37
45	16%	60%	8%	4%	-	8%	4%	25
40	22%	52%	17%	-	4%	4%	-	23
35	13%	38%	17%	21%	-	8%	4%	24
30	8%	42%	17%	-	8%	25%	-	12
25	-	20%	50%	-	10%	10%	10%	10
20	-	-	67%	33%	-	-	-	3
#	32	69	28	10	4	11	6	160
Mean NWET	48	44	38	40	36	38	43	43

Note. Columns P, N, I and U stand for pass, fail, incomplete and unknown, respectively.

Table 7.2

NWET Scores and Wr 110 Grades

NWET score	P	N	#
60	100%	-	1
55	-	-	0
50	100%	-	2
45	83%	17%	6
40	100%	-	11
35	100%	-	2
30	60%	40%	5
25	50%	50%	8
#	34	7	41
Mean NWET	36	25	34

Note. Columns P and N stand for pass and fail, respectively.

Table 7.3

NWET Scores and Misassigned Students

Category	#	Mean NWET	Distribution of scores 20-29	30-39	40+
Should be in Wr 111	6	45.5	0	2	4
Should be in 110	12	36.7	2	7	3
Should not have passed Wr 110	3	32.3	2	0	1

From these and similar tables, the Registrar concluded that students scoring below 30 on the NWET could expect to have problems completing the remedial course (Wr 110) successfully. In light of this and related conclusions, the Registrar made the following placement recommendations:

1. Place students scoring between 20 and 29 on the NWET into a pre-remedial program.

2. Students who score between 30 and 39 should be placed in Wr 110.

3. Students who score between 40 and 54 should be placed in Wr 111.

4. Students who score between 55 and 60 should be placed in Wr 111, with an opportunity to take a supplementary examination for course waiver. (2B)

With the Registrar's study and recommendations in hand, the Director of Composition secured departmental and university approval to use the NWET as a placement device for the UB composition courses. Students with scores below 30 would be placed in the newly-created WR 50. Students scoring below 38 would be placed in WR 110. Writing 111 would be recommended for students scoring at or above 38 on the NWET. The cut-score for Wr 111 placement was two points less than that recommended by the Registrar. When asked why the standard was lowered, the Registrar explained:

> That was the usual compromise. It depended on the number of students who were going to be in Wr 110 and the ability of the English department to handle them. Also, the cut-score related to how many students were being placed in Wr 110 traditionally. I wanted to have more students placed in Wr 110, but they didn't want to handle that many more classes. (222)

Optimistic about the utility of the NWET at UB, the chairman of the ad hoc committee brought the Registrar's study to the attention of the members. A number of members expressed interest in the NWET as a placement tool: two community colleges and one state college undertook studies to examine the test's utility for their placement decisions.

When the ad hoc committee members had completed their charge, they produced a final report which contained thirteen recommendations, two of which merit attention here. The

committee called for the establishment of basic prerequisites for admission to Wr 111 in recommendation #4:

> The Committee recommends that a list of four basic skills be adopted statewide as requirements for graduation from high school and admission into the standard Wr 120 writing course:
> 1. Write complete, correct sentences;
> 2. Use punctuation correctly;
> 3. Follow the generally accepted conventions of standard English usage; and
> 4. Spell correctly, and know the meanings of, the words commonly used in one's own writing. (2C)

The institutional investigations of the NWET must have been impressive; the committee made special note of the test in its fifth recommendation.

> The Committee recommends statewide adoption of the National Written English Test as a diagnostic device for measuring the basic skills as prerequisite for admission to Wr 111. The standardized test may be supplemented by locally-developed instruments and/or writing samples. (2C)

No recommendation was made for the use of a particular score or standard. During his interview, the committee chairman speculated that similar to UB, the choice of cut-scores by the various institutions would be dictated largely by economics.

> When other state institutions adopted the NWET, they used varying cut-scores for placement decisions. I think, in many cases, that this had to do with how many students could be handled in remedial courses. In other words, how expensive would it be to remediate a third or 50% of your incoming student body? So economics really played a big part in this. (205)

Since the committee's recommendations were purely advisory and not binding, only a few of the institutions elected to use the NWET. Southern Brittany State College used a two-tiered system like UB's: setting two cut-scores for

placement into the pre-remedial, remedial and college-level composition courses. Satisfied with the SAT-V, BSU did not choose to adopt the NWET for placement purposes. Officials from the University of Dover (DU) had serious reservations about the NWET, claiming that minority students in their urban institution would be adversely affected. These concerns kept DU from immediately implementing the NWET for placement purposes. By 1978, however, the University of Dover started to use the NWET for making placement decisions in the WR 110 class created a couple of years earlier. Students scoring 35 or below would be placed in WR 110.

The influence and use of the NWET seems to have been greatest at the University of Brittany. Indeed, a new course, Wr 50, was created to accommodate students scoring below 30 on the competency test. Wr 50 would now be called the remedial or developmental composition course, and Wr 110, the former remedial course, was referred to as a corrective composition class. A study was conducted after the first group of freshmen completed the newly-created Wr 50 course. Of these 89 students who originally scored below 30 on the NWET, 79% scored at or above 30 on a second administration of the test. In addition, background data on the 89 students indicated that about 60% were caucasians and about that same percent were from families with annual incomes above $9000. From these data, UB concluded that neither minority nor low-income students were placed disproportionately into this remedial composition course. (2F)

As we stated from the outset, the story of the NWET can be written in two parts, the first of which has dealt with its use as a placement tool. What has been described provides much of the context for understanding what will be described next: the selection of the NWET as the means for tightening Brittany's admission requirements in her two flagship institutions. The story of the mandated use of the NWET for entrance is connected closely to that of the first: actions taken by administrators and policy-makers were based partly on decisions made earlier by practitioners and some of the same managers. Specifically, the score chosen for the admissions standard was the one which identified the hard-core remedial writing students. Students in moderate need of remediation are not denied admission on the basis of the NWET standard, even though this larger group of students is nearly as academically vulnerable as the hard-core remedial group.

Tightening Admission Requirements: A Different Way

Generally, the Board of Higher Education examines admission requirements each year and gradual increases in the high school grade point average (GPA) have commonly followed such evaluations. In 1978, however, the concern for enrollment control ran high. The state's economy had suffered several debilitating blows and the legislature made it clear to all public institutions that funding cuts loomed on the horizon. Post-secondary education would not remain untouched and enrollment control was viewed by the Board of Higher

Education as one means of coping with the impending budget squeeze. The Board of Higher Education instructed its staff to undertake a more in-depth study of post-secondary admission requirements. According to a report on this study prepared by the Board of Higher Education staff,

> Following consultation between and among the Board's staff and institutional representatives, it was concluded that a longitudinal study of a sampling of resident freshmen who entered state system institutions fall term 1976 would provide essential basic information concerning the relationships between and among (1) the predictors: high school preparation, measures of high school academic achievement, and scholastic aptitude and (2) the criterion: academic achievement in college. (2I)

The study was a massive one. Twenty percent of the resident students who entered as freshmen in 1976 were sampled. Researchers collected students' high school transcripts, as well as relevant test scores and various GPAs (cumulative, yearly, academic, and the like). These data were analyzed in relation to academic performance (and persistence) during the students' first two years of college. A Board of Higher Education staff member described one of the conclusions drawn from this study:

> The conclusion by the Vice Chancellor for Academic Affairs was there was no significant reason to change from the current criterion of a high school GPA of 2.5 to coursework patterns. (221)

In other words, the number of high school college prep units was not a good predictor of college academic achievement but the quality of work done in these courses was at least as good a predictor as SAT-V or SAT-M scores. (2I) From these results alone, it appears as if the Board of Higher Education would have no choice but to increase the high school GPA

requirement one more time or institute an aptitude test requirement. Indeed, either of those would have been the only reasonable recourse if one other study had not been conducted. On his own time, the recently retired UB Registrar began to study the association between the NWET and academic performance in college. He described the impetus for this investigation, which he conducted during the months of the Board's admissions study:

> I decided that we needed to do something to tighten up on English, because people were talking about not teaching Wr 50 at the university. This was due to anticipated budget cuts and the philosophical idea that if students couldn't handle English, then why are we admitting them? That gave me an idea to figure out who can't pass the NWET and not admit them. (222)

Specifically, the former Registrar's study looked at the performance of freshmen entering UB in Fall, 1978. He found that students placed in Wr 50 and 110 accounted for 5.5% and 11% of the freshman class, respectively. (2G) Students in these two classes were distributed along the range of high school GPAs and so the former Registrar concluded that high school GPA was not useful in identifying students in need of English remediation. He also examined students grouped by NWET scores and compared them according to their academic standings as of the end of Spring, 1979. Table 7.4 displays information contained in the study's report (2G).

Table 7.4

NWET Scores and Academic Standing

NWET Scores	No. of students	Probationary Status: No probation	Probation[a]	Unkown
20-29	111	61 (55%)	48 (43%)	2 (2%)
30-39	411	254 (62%)	157 (38%)	-
40-49	745	560 (75%)	185 (25%)	-
50-60	745	632 (85%)	113 (15%)	-
Total	2,012	1,507 (75%)	503 (25%)	2 (-%)

[a] Probation includes academic warning, 1st term probation, 2nd term probation, attrition, and subject to disqualification.

Of students scoring 20 to 29 on the NWET, nearly 45% had been subject to some kind of probationary action during their first year. Of students in the 30-39 range, 38% were subject to some kind of probationary action. These two rates were considerably higher than the entire freshman class rate of 25%. Other findings were reported in this study:

> Students with deficient written English skills as identified by the NWET show marginal performance in college (more than .5 below the average GPA). Their high school GPA is only .20-.25 below the average high school

> GPA. Nearly twice as many of this group will fail to compete 12 or more credits during the freshman year. (2G)

The former UB Registrar brought his results and conclusions to the staff conducting the Board's admission study, including the Vice-Chancellor of Academic Affairs. When the report of the Board's admissions study was prepared, a recommendation for a writing requirement was included:

> Students who meet the admission requirements expressed in terms of high school GPA (2.50) or who have a predicted GPA of 2.00 (based on high school GPA and SAT scores) may be admitted to the University of Brittany and Brittany State University, provided they also can demonstrate a competence level in English composition equivalent to that represented by a score of 30 on the Na tional Written English Test. (2I)

Raising the high school GPA requirement was thought to eliminate more students who would succeed academically than who would not, neither would it significantly improve student retention rates. The NWET, however, would prevent those students deemed academically vulnerable from entering the two universities as well as eliminate the need to provide the costly remedial writing course (Wr 50).

The decision for the particular NWET standard was essentially made by the then Vice Chancellor of Academic Affairs for the entire Brittany system of higher education. The Vice Chancellor received input from various university officials and Board of Higher Education staff members, as well as from the former UB Registrar. After reviewing the former Registrar's study the Vice Chancellor determined that a NWET score of 30 would be required for admission.

Data were available to help project the consequences of applying this standard of 30. Based on enrollment figures from Fall, 1978, the University of Brittany and Brittany State University would stand to lose 6% and 8% of their freshman classes, respectively. (2E) For UB and BSU combined, it was estimated that fewer than 500 students would be excluded annually. (2F) According to the UB Director of Admissions, these students were considered lost for the first year only as they would have likely dropped out by their second year. (203) As indicated by the former UB registrar's study, nearly 45% of these students would be subject to some kind of academic probation during the first two years of college. (2G) The former Registrar elaborated:

> We were not so much concerned with keeping students out as eliminating those who had a high potential for failure. It is not a very good student personnel policy to admit students who don't have much chance of making it. (221)

Officials' concerns about admitting students who were potential dropouts did not seem to be the only consideration in setting the NWET standard. Like those scoring below 30, students scoring below 38 on the NWET were academically vulnerable. According to the former UB registrar's study, 38% of this group were likely to be placed on academic probation during their first two years of college-- a rate higher than that for the entire class (25%). However, setting the NWET standard at 38 would have affected a greater proportion of the freshman class in both UB and BSU. If the score used to place students in Wr 110 (38) had been selected as the admissions

standard, an additional 11% of the freshman class would have been inadmissible. In sum, nearly 17% of the freshman class would have been affected by the higher standard.

In October, 1979, the Board of Higher Education incorporated the NWET into the admission criteria for UB and BSU. Beginning Fall, 1981, applicants were required to have a high school GPA of 2.5 and a score of 30 or greater on the NWET. An alternate test and standard were approved if students did not take the NWET. These students would be allowed to submit their scores on the English section of the ACT as demonstration of basic English skills. If using the ACT-E in lieu of the NWET, students were to score at least 12 to be admissible. At this same meeting, the Board of Higher Education raised the quota of specially-admitted freshmen. Historically, the institutions were allowed to admit 3% of the entering freshmen who did not meet the general admission requirements; now 5% would be admissible. (222, 2F) Before we discuss consequences directly related to NWET admissions requirement, we will describe a number of events related to Brittany's efforts to raise standards.

Remedial Courses and State Funding

As anticipated, post-secondary budgets came under legislative review in the following months. Admission requirements and remedial education were targeted in a number of documents produced during this review. In June, 1980, a state administrative agency prepared a planning paper on

remedial education in Brittany's post-secondary institutions. Two major observations were made by the agency, the first of which was that only 3% of the applicants to Brittany's universities (UB, BSU and DU) were denied admission. The agency's second observation was that math, and not writing, accounted for the the greatest amount of remediation dollars spent by institutions statewide (2M). The authors of the report suggested following one of two alternatives. The first alternative included the establishment of more selective admission requirements. The second alternative called for revisions in remedial education; the agency recommended that 1) credit in remedial courses not count toward graduation and 2) state funding for remedial courses be removed from institutions located within community college districts.

The acting Vice Chancellor for Academic Affairs responded to the agency's planning paper. She noted the agency's omission of the Board of Higher Education's adoption of the NWET requirement:

> The paper on Remedial Education tends, we think, to overstate the problem. The remedies you suggest might be appropriate for remedial writing courses and, indeed, the University of Brittany and Brittany State University have imposed a writing competency requirement for admission of new freshmen effective Fall, 1981, an action you will probably wish to mention in the document. (2M)

The acting Vice Chancellor also addressed remedial mathematics. She argued that remedial mathematics courses, particularly Math 90, was not remedial in the same sense as the writing courses.

> Mathematics is a different situation. Most of the enrollment in mathematics instruction reported in our

> tables is Intermediate Algebra (Math 90). Any student coming to college who wishes to major in a science or science-based professional field or business administration must take mathematics. Students who have two years of high school algebra start with College Algebra. Students with less than two years of high school algebra are placed in Math 90. For most of the students in Math 90, they are not remediating something learned poorly, but rather covering material not previously studied. (2M)

The lines were drawn. The Board of Higher Education and its staff indicated that standards had been raised in writing. The state administrative agency felt that higher standards could have been imposed. According to the deputy director of this agency:

> Somebody might look back and say, 'Yes, we raised standards to improve the quality of the students being admitted'. If that was so, why didn't they raise the NWET cut-score to 38, the score necessary for placement in college-level composition? Well, they didn't do that because that would cut out a lot of students. (204)

In addition, the administrative agency disagreed with the Board's contention that Math 90 was not a remedial course in the same sense as Wr 50. Indeed, the agency viewed Wr 50, Wr 110 and Math 90 all as remedial courses and equally qualifying for budget cuts.

In Spring, 1981, this same agency prepared an analysis of possible post-secondary budget reductions. The agency made a number of recommendations, including one about state reimbursement for remedial courses. The Brittany legislature eventually approved their recommendation: general fund support for remedial courses (those numbered below 100) would be eliminated for the 1982-83 academic year. (20)

As a result of this legislative action, Wr 50 would no longer receive state funding. Since the implementation of the NWET requirement, this result should have had few consequences as students eligible for Wr 50 were said to be inadmissible. In some institutions, Wr 110 was renumbered below 100 while in others it was not. Some institutions were able to find other funding sources for these two courses while others required their students to take the courses through continuing education or nearby community colleges. Unlike the remedial writing courses, the Math 90 course was spared the state funding cuts. The Board of Higher Education approved institutional requests to renumber to Math 101, relieving the course of its remedial connotations.

While these and other changes were being made in the post-secondary system, the NWET standard was nevertheless applied to students seeking admission to UB and BSU. In the next section, we will describe some of the consequences and reactions related to the NWET.

Consequences of the NWET

How many students have been affected by this requirement and standard? We were able to obtain some information from the BSU Director of Admissions. Roughly 2,000 applicants were admitted by BSU for Fall, 1985-- we did not get a figure of the total number of applicants. If admission rates were similar to those in 1980 where only 3% of the applicants were denied admission, we might infer that at least 2,060 applied

to BSU for Fall, 1985. Of these 2000+ applicants, 93 were inadmissible on the basis of the NWET alone, 75% of whom were state residents. BSU did not deny admission to all of these 93 applicants-- 39 (42%) were eventually admitted for the fall semester. According to the BSU Director of Admissions, these 39 students became eligible for admission by one of three ways: 1) retaking the NWET and scoring at or above 30; 2) taking and scoring at least 12 on the ACT-E; or 3) enrolling in a special summer term program. Ultimately, 54 students, 67% of whom were state residents, were denied admission to BSU. We do not know if the numbers at UB compare in any way with those cited from BSU.

What has happened to students denied admissions to either UB or BSU on the basis of the NWET? Some may have been admitted under the 5% exception rule. According to the UB Director of Admissions, only a few students have been admitted by exception:

> We've admitted a few, but if we're going to have the requirement, we're going to have the requirement. My judgment is that a number of them chose to go to the community colleges. Maybe they went to some other kind of training program. That is likely. (203)

Satisfaction with the NWET requirement is not the same between UB and BSU. The Director of Admissions at UB indicated that attrition rates have improved for entering freshmen, dropping from 20% to 10%. He attributed this improvement to the NWET requirement. In fact, he expressed a desire to raise the NWET standard to 38, the score used to place students in the regular composition class.

Approximately 200 students in the 1985-86 freshman class scored between 30 and 37, but the UB Director of Admissions estimated that only 100 students would be affected by a higher standard. He anticipated that 50% of the 200 students would earn a score of 38 on a second administration of the NWET.

The Director of Admissions at BSU held a different opinion about the utility of the NWET requirement and standard. In fact, he has supported efforts to dismantle the NWET requirement as part of the admissions criteria. Three reasons were cited in defense of his position: 1) the dissimilarity of the two institutions; 2) the image of the NWET standard; and 3) the addition of new admission requirements. According to the BSU Director of Admissions:

> UB and BSU are very dissimilar institutions. The University of Brittany is a liberal arts related institution that also has science and some professional programs. Brittany State University is a science-related institution that also has the strong profession programs and liberal arts. The emphasis is almost reversed. (226)

He believed that the focus on the writing skills and the approval of the NWET admissions standard were more a reflection of UB's desires that those of BSU. The second reason for his opposition to the NWET involved the image of the NWET standard:

> The standard sounds ridiculous. The kind of students that we talk to, both within and outside of the state, are surprised to learn about this requirement. The typical reaction that I've observed is that the requirement is so low-- half of 60 points. A score of half the points doesn't sound good at all. (226)

The BSU Director of Admission also questioned the utility of the NWET requirement since new admission criteria had been

adopted. When a new Chancellor for the Brittany system of higher education was hired in 1982, he worked toward the adoption of required high school courses. Effective for freshmen entering Fall, 1985, the coursework requirement was seen as fulfilling the intent of the NWET:

> Very honestly, I think that now we've included the course requirements-- four years of English composition for every entering freshman-- I personally don't see any justification for the NWET requirement at all. (226)

Given the different positions held by the two university admissions directors, there has been some discussion of modifying the NWET requirement. BSU requested that the Board of Higher Education approve the Test of English as a Foreign Language (TOEFL) as an alternate for the NWET. The Board denied BSU's request in Spring, 1985. Other informal discussions continue about waiving the NWET requirement for BSU and raising it at UB, but no formal action has been taken as yet. According to a Board of Higher Education staff member, such action may hinge on the results of an anticipated study of NWET scores and freshman GPA vis a vis the implementation of course requirements.

CHAPTER VIII

TESTING FOR ADMISSION TO TEACHER PREPARATION PROGRAMS IN GRANADA

A Glimpse of the Present

The state of Granada has been testing elementary and secondary teachers since the early 1980's. At present, there are two separate basic skills testing programs operating in the state for students seeking admission to teacher preparation programs. One test, the Pre-Professional Skills Test (PPST), is administered by the authority of the Granada Board of Regents. The PPST, developed by ETS, is given to all students applying for admission to any state university or community college teacher preparation program. The second test, the Granada Test of Teacher Proficiency (GTTP), is administered by the Granada State Department of Education, under the auspices of the State Board of Education. Developed in-house, the GTTP has two components: basic skills and pedagogical knowledge. The GTTP basic skills component (GTTP-BS) is administered to students applying for admission to teacher preparation programs in private colleges. The second component of the GTTP is designed to assess pedagogical knowledge (GTTP-PK) and must be passed by all individuals seeking certification.

Teacher testing practices appear to be redundant in Granada. At the point of entry into teacher preparation programs, two basic skills testing programs operate

concurrently while at the point of certification only one of the two programs operates. Why are there two separate tests, administered by two separate agencies, both of which are designed to measure essentially the same thing (basic skills) for the same purpose (entry into a teacher preparation program)? What happened in Granada to produce this apparent duplication? In the following case study we describe how this current state of affairs came about and examine the means by which standards were set on each of the basic skills tests. The description will be a chronology of events from 1980 to 1986 and from this chronology, it will be seen that the two agencies, the State Board of Education and Granada Board of Regents, came into conflict over testing and standards and this conflict led to the creation of two parallel testing programs.

1980 to 1983: The Granada Test of Teacher Proficiency

The 1980 Granada legislature amended a bill dealing with tax reform and school finance to include a requirement that prior to certification, all new teachers must pass a proficiency examination in reading, grammar, and mathematics. This move occurred during the final days of the legislative session and without the benefit of public hearings. According to a university professor reflecting on the motives guiding such a requirement:

> I think that the program came as a result of legislators being unhappy with teachers not being able to write home communications that had decent spelling and

> grammar . . . Parents reported to legislators and legislators got upset about it and so they wanted to be sure that the people who became teachers had a background in grammar as well as some basic skills in mathematics and reading. (101)

A staff member of the Granada Department of Education explained his understanding of the legislature's intent:

> . . . the legislature pointed out when they passed the original legislation, that this did not in itself either make or break a good teacher. This is no indication of your ability to teach. It is merely an indication that you have the very basic of skills in order to act as any kind of a role model for students. (104)

The State Board of Education was given the tasks of selecting the test, administering it and determining the standard by which students would pass or fail. The program was to be in place by October 1, 1980. Since the legislative mandate came so late in the session, the State Board had only three months to implement the program. To comply with the mandate, a State Board subcommittee adopted a nationally-normed test on an emergency basis in September, 1980. Concerns arose over the appropriateness of the test as a measure of basic skills and as a result, a task force advised the State Board to develop its own test. Test items in the three content areas were evaluated from an existing test bank developed and maintained by the Segovia County School Superintendent's office in a nearby state. All items in this bank had been subjected to technical analysis and screened for bias. Fifty questions for each subject area were selected, all of which were written on a grade 14 level. These 150 items were assembled to form the initial GTTP basic skills

test, which was approved by the State Board of Education in mid-October, 1980. In describing the origin and development of the GTTP-BS, a document produced by the Granada Department of Education (GDE) echoed the legislative intent of the testing requirement:

> The GTTP is neither designed nor intended to measure or predict teaching ability. It is a minimal competency exam in those basic skill areas needed by all teachers. It is, in essence, an attempt to determine teacher proficiency in those same basic skills we expect them to teach our children. (1B)

The State Board announced a field test period for the GTTP-BS starting in November, 1980, through June of the following year. A score of 75 (50% of the items) was selected as the standard which students must earn to pass the test. It did not matter whether a student passed all 50 questions in reading and only 25 questions between the mathematics and grammar subtests or passed 25 questions on each of the three subtests; 75 questions correct, no matter where they came from, was a "pass". According to the previously cited GDE document:

> This arbitrary score was set to comply with the law, and also to avoid placing the State Board in legal jeopardy from unsuccessful applicants during the field test period. (1B)

The State Board met to review the available field test data at the end of April, 1981. A decision was made by the Board to adopt a two-stage increase in the standard for the GTTP-BS. The standard would be raised to 75% in July, 1981, and it would be raised one more time to 80% in January, 1982.

The details of the decision-making process are elusive. In September, 1981, the Granada Department of Education hired an individual to oversee the entire teacher certification program, including the GTTP. To get a better understanding of events preceding his arrival, he traced the process which produced the two-stage increase in standards. After interviewing different Board of Education members and collecting in-house memos, the staff member concluded that the field test normative data served as the basis for determining the GTTP-BS standards. Psychometric methods designed to produce standards (e.g. Nedelsky, Angoff, and Ebel) were not used. By this staff member's account:

> And instead of going through the normal kinds of standard-setting approaches, where you review the item to judge whether it's necessary and all- they did a little of that- but, what they did was display how much damage will the standard do and at what level? (105)

To lend support to his account, the staff member produced a document entitled Granada Test of Teacher Proficiency Field Test Summary. In this document, the field test data were arrayed to show passing rates on each section of the GTTP-BS as well as by battery and conjunctive standards. The battery standard reflected the overall percent of correct items needed to pass while the conjunctive standard was based on attaining a specified percent on each of the three sections. During the field test period, 3,921 persons took the GTTP-BS and 3,841 (98%) passed using the 50% battery standard. Passing rates for each subtest were well above 90% using a 50% standard. Even when a conjunctive standard of 50% for each subtest was

applied, the overall passing rate remained high: 90.3%. According to this staff member, the State Board was able to inspect the passing rates for an additional seven standards ranging from 55% to 85%. For example, by applying a battery standard of 60% it could be seen that 92% would pass the GTTP-BS whereas a battery standard of 70% would produce a passing rate of only 79%. In addition to looking at overall passing rates, the staff member explained that the State Board looked at passing rates by ethnicity. Based on the staff member's understanding of the process, he typified the Board's concerns:

> . . . how many blacks, Hispanics and Native Americans, as well as overall, are we going to exclude by having a particular cut-off? It was purely what seems to be reasonable and limit the damage. When they looked at the passing rates, some wanted to go as high as an 85% cut, with others saying, 'No, if we do that, we'll exclude eight out of every ten Native Americans, and we will truly have an impact that the courts have said we can't have. They thought that a two-phase approach would give people notice, which was a concern at the time. (105)

This account was corroborated in an earlier interview with the current GDE coordinator of the GTTP. When asked whether the State Board had an idea of what percent would fail with an 80% standard, the GTTP coordinator replied:

> Oh yes. They had a good history of what had happened in the past, and by looking at the 80%, you could tell how many people would fail, as opposed to putting the cut score at 70% or 50%. They were aware of what the casualties would be. (104)

In addition to concerns over minority impact and possible legal entanglements, beliefs about the standard's image played

a role in the decision-making process according to this former staff member:

> And then there was the belief that the higher the standard, the greater the expectation, the greater the performance. When I asked what the 80% meant in terms of correlations with the California Achievement Test or something else, the only response I could get was from the deputy superintendent: 'In most people's minds, 80% reflects a B, which is above average'. That is what they wanted to establish- a standard which would be above average. (105)

The current GTTP coordinator voiced a similar understanding of the Board's beliefs about the 80% standard:

> They felt that there is a certain spot that you can't retreat from as far as passing this basic skills test, and they thought that 80% was it. (104)

Given the State Board's decision to implement the 80% standard in two stages, the anticipated fail rates can be traced back to the field test summary document (1I, 1L). Approximately 66% of the examinees would pass the GTTP-BS when applying a standard of 75%. With a cut-off score of 80%, the passing rate would drop to nearly one-half of the examinees. Did the State Board of Education actually anticipate and approve a 50% failure rate among certificate applicants? It seems unlikely. According to the former GDE staff member, the State Board estimate that the actual passing rate would be greater than 50% based on the premise that examinees would treat the exam seriously, knowing that they must pass to qualify for a teaching certificate.

Before discussing consequences related to the implementation of the GTTP-BS, the major policies and practices will be outlined (1D, 103, 104):

1. Examinees needed to apply only two weeks prior to the test's administration, with no waiting period necessary between subsequent retakes.

2. A charge of $10 was levied for the initial administration, as well as each additional retake.

3. Two hours and ten minutes were allowed for each test administration.

4. The tests were administered twice a month in the cities where the state's major universities were located.

5. The preceding practice contributed to the development of another in which examinees could take the test as early as their junior year in college.

6. Each examinee was sent a report detailing performance on the three subtests. Additionally, each was given a list of the individual skills appearing on each subtest which they could use for personal remediation. No formal means of remediation existed for those failing the exam.

7. Examinees who failed the exam were allowed an unlimited number of retakes.

How did the actual passing rate on the GTTP-BS compare to the field test passing rate of 50%? It depends on which document one consults. The earliest source detailing pass-fail data was produced by the GDE sometime after June of 1982 (1H). No breakdowns by demographic features were included in this document. The data were based on the performance of 1,631 examinees during the period from January 1, 1982,

through June 30, 1982. Table 8.1 displays the fate of these 1,631 examinees during this six-month period.

Table 8.1

GTTP-BS Pass Rates from 1/82 through 6/82

Test administration	Number taking exam	Number passing	Cumulative % passing
First	1,631	1,218	75%
Second	236	136	83%
Third	54	19	84%
Fourth	22	13	85%
Fifth +	12	5	85%

Approximately 75% (n=1,218) passed on the first administration of the GTTP-BS. Among the 413 who failed the first round, 236 elected to take the exam a second time. After the second round, the cumulative passing rate increased to 83%, leveling off to 85% after the fourth retake. The 177 examinees who failed the first round and who did not attempt a retake during this period constituted about 11% of the original group. The cumulative dropout rate increased to 15% after the fourth retake and remained there throughout the six-month period. Among the examinees who did not dropout by the third retake (n=1,408), only 2% failed the GTTP-BS.

Another document prepared and presented at a professional conference by the former GTTP coordinator (1L) contained some information on GTTP-BS performance patterns over a 30-month period (1/82-7/84):

> Roughly 67% of the total exams administered for certification purposes have been passed. While females tend to pass at a slightly higher rate than males, Granada minorities display significantly lower pass rates than Anglos. Indeed, the performance pattern for ethnic groups indicate the following pass rates: Anglos = 73%; Hispanics = 42%; Blacks = 25%; and Native Americans = 23%. (1L)

Imprecise language and insufficient information in this document make for ambiguous interpretations of the passing rates. By reading this paper, one cannot conclude whether the rates reported reflect retakes or only first attempts. The current GTTP coordinator indicated that, indeed, 67% of the examinees passed on their first administration. (104) The percent of examinees who eventually passed after retakes was not provided in any of the collected documents. It is curious that such relevant information was absent in these published documents. Nevertheless, based on passing rate patterns in the six-month period, one may conclude that the overall passing rate was considerably higher than 67%.

Representatives of various minority groups expressed concern over the disproportionate numbers of their groups who failed the basic skills exam. They argued that the test was unfair. On the other hand, some of the majority complained that the test was too easy, i.e., too many people were passing who should not. The most organized and visible response to the GTTP-BS came from the Granada Association for Tax Research

(GATR), a group that could not be classified simply as siding with criticisms by either the majority or minority. While the underwriting organization did not have an obvious connection to testing of prospective teachers, the motives behind the effort appear relatively straightforward. Specifically, the author explicitly stated early in the report that the GATR:

> . . . did not intend any criticism of using a valid test for screening teaching applicants, but the decision to deny or award certification should be based on an accurate determination of whether the person has an adequate knowledge of basic skills. (1M)

The GATR study examined three technical issues related to the GTTP-BS. The first issue dealt with invalid test questions. This group procured item level data from the Department of Education on nearly 2,500 examinees who took the exam during a nine-month period. Basing their opinion on the correlations between total subtest scores and item performance, the researchers claimed that over one-third of the GTTP-BS items were invalid. The concern for bias against minorities, the second issue, was addressed by comparing these same correlation coefficients between one ethnic group and the majority. The authors concluded that the evidence was not compelling enough to assert that cultural bias invalidated the suspect items. The third issue receiving attention in the GATR report concerned the cut-off score of 80% for the entire battery:

> . . . the test is scored for overall knowledge rather than knowledge on the individual basic skills. . . . . The level of proficiency needed to pass the test does not seem to be adequately defined. The cut-off score of 80% does not appear to have any relation to any determined level of proficiency. The cut-off appears to have been

> selected primarily upon the political effect different cut-off scores would have considering the effect on minorities. There is no evidence any attempt was made to find out what minimum level of proficiency was needed, nor to determine whether the test validly measured this proficiency. (1M)

The criticisms leveled by the GATR and other individuals would eventually serve as the weapon in the coming conflict between the Granada Board of Regents and State Board of Education. Before we discuss events related to this conflict and the ultimate creation of two testing programs, we will describe the second component of the Granada Test of Teacher Proficiency: the pedagogical knowledge subtest (GTTP-PK). In December, 1980, the Granada State Board of Education determined that the basic skills measure should be accompanied by a pre-service exam of pedagogical understanding: knowledge of the essential skills which should be learned by a student in a two-year teacher education program. (1B, 1L) As a result, the State Board approved the requirement that prospective teachers take and pass a pedagogical knowledge test, yet to be developed and incorporated into the GTTP.

Over the next nine months, five teams participated in the development of the GTTP-PK. The first team, composed of university faculty, documented the essential specific skills which should be gained from the State Board-approved courses. (1B, 1L) A second team reviewed the approved teacher education programs for purpose and content; the makeup of this team is not clear. A third team of 267 persons (90% of whom were classroom teachers and administrators) independently listed essential skills that should be learned in a two-year

teacher preparation program. The tasks of analyzing the products of the first three teams and developing a list of skills to be assessed fell to the fourth team. Finally, the fifth team met in August, 1981, to review and approve specific test items for appropriateness in a assessing the identified skills. (1B, 1L) Having passed the inspection of this team- membership unknown- the items were delivered to the Board for final approval. Field testing of the GTTP-PK began shortly after the State Board's approval, using the same standard applied to the GTTP-BS during its field test period- 50% of items correct. The GTTP-PK field test period continues to present time, a duration of nearly five years.

When queried about passing rates on the GTTP-PK, the administrative assistant for the GDE program indicated that anywhere from 97-99% passed at any given time. His estimate is supported in a separate account by the former GTTP coordinator:

> In terms of actual performance, it should be noted that approximately 98% of the exams administered [n=2500+] have been passed at the 50% correct responses criterion. This pass rate holds across virtually all content areas and across demographic variables such as gender, ethnicity, and college or university teacher training program. (1L)

The assistant GTTP coordinator described the importance of this high passing rate in the following interview excerpt:

> 103: It looks good on paper. You can say 'Well, look. All of our people are passing it'.
> MCE: And that's important to the State Board and teachers'...
> 103: Yes. In this state, the chief state school officer is elected to office.
> MCE: I see. So it doesn't look very good if people were to pass at a rate of 50% or 20%.

103: We've got some figures that if they were to raise the score say up to 80%, then we'd drop to something like a 60% fail rate.
MCE: And that would be unacceptable not only to the chief state school officer, but to other people involved as well?
103: Yes. (103) , cp. 167-168).

It appears, then, that the high passing rate on the GTTP-PK satisfied political concerns. A lower passing rate would likely have been unacceptable to the Chief State School Officer.

By the end of 1983, the Granada Test of Teacher Proficiency was being administered to thousands of prospective teachers across the state. Although the GTTP (both components) was originally intended for individuals seeking certification having completed their professional training, many were taking the GTTP-BS during their college years. Two conditions may have contributed to this emerging practice. First, it was convenient for college students to take the exam since the GTTP was administered in cities where the major universities were located. Second, and possibly most influential, was the fact that examinees had unlimited opportunities to take the GTTP. By taking the exam during their junior or senior year, students could get an idea of what was expected without suffering any delay in eventual certification. According to the current coordinator of the GTTP testing program,

> So if you failed it, by gosh, you have a chance to study for it, and I am sure a number of them just took the thing to see how it was, and what they needed to do if they didn't pass it. (105)

This coordinator expressed his conviction that the high initial failure rate was influenced largely by students taking the exam in a situation in which passing or failing was irrelevant. The college students taking the GTTP-BS may or may not have contributed to the GTTP-BS failure rate, but it is clear that their growing numbers were the subject of attention by the Department of Education, as well as the Board of Regents and eventually, the state legislature. The former GTTP coordinator reflected on how the focus on the college students sharpened, leading to key 1984 legislation:

> Understand that when I was in the state department, it was very clear. 'We are going to publish these results in terms of which institutions are producing what people. We are going to make it their problem.' The institutions are saying 'Hey, these basic skills are taught long before they come into our programs'. About the time the papers started publishing the passing rates by each university, the institutions said 'This is crazy. We are going to start demanding GTTP-BS passage before we let them into our program.' (105)

In the following section, we will describe the legislation that mandated testing prior to entry into Granada teacher preparation programs. We will show how events following this legislation led to conflict between the two boards resulting in the creation of two testing programs.

1984: Legislated Cooperation that Led to Conflict

In Spring, 1984, the Granada legislature enacted a bill which required all students to pass a basic skills exam before entering any of the state teacher education programs. The Associate Director of Academic Programs for the Board of Regents characterized the legislature's position in this way:

> Well, we don't want students going through a teacher education program getting up to certification and finding out they don't have the basic skills, wasting the taxpayers' dollars and the students' time. So test basic skills before entering the teacher education program. (102)

The legislature did not adopt a particular examination, but mandated that the exam be composed of three subtests: reading, writing and mathematics. In addition, the legislature stipulated that the test standard must not be based simply on an overall percent of items correct. The standard had to be a conjunctive standard in which applicants must pass each of the three subtests. According to the legislation, the Board of Regents and State Board of Education were to agree on the particular exam to be used for the teacher education applicants. If a test other than the GTTP-BS was selected, the State Board of Education must accept it in lieu of the GTTP-BS. In other words, the legislature did not wish to have students take a basic skills test twice: before and after their teacher education.

Confident that the GTTP-BS would be adopted, the State Board of Education moved to comply with the legislative stipulation for a conjunctive performance standard. Some members indicated a desire to keep the same standard of 80% items correct and apply it to subtest rather than total test performance. Others members expressed a desire to reconceptualize the standard-setting process after listening to the following suggestion made by the former GTTP coordinator:

> Let's set the cut score in a reasonably technical fashion so that we can have a legal basis if challenged. Let's do this rather than sit around, flip a coin, and decide on that basis what the cutscore will be. (105)

The former coordinator had already initiated some work on this plan:

> I wanted to take into account standard error of measurement and other things which would result in alternative cuts because of the difference in discriminative values and the difficulty level of the items. I had alternate raw score cuts for each area. (105)

The State Board of Education was not willing to follow the suggestions for a more technical standard-setting approach. Rather, the Board examined projected passing rates by various standards, arrayed by ethnic group (1K). Eventually, they selected a conjunctive standard of 80%. Table 8.2 shows the anticipated pass rates by ethnic group for this new standard.

Table 8.2

Anticipated GTTP-BS Pass Rates by Ethnic Group

Ethnic group	Number tested	Percent passing
White	2,752	44.6%
Black	86	10.5%
Hispanic	223	16.6%
Native American	164	8.5%
Asian	20	45.0%
Total	3,341	39.6%

The overall passing rate of nearly 40% of the examinees was considerably lower than what had been observed with a battery standard of 80% of the items. Disparate passing rates among ethnic groups would continue to be observed with the use of the conjunctive standard. Nevertheless, the State Board decided to adopt the 80% standard for each of the three subtests on the GTTP-BS, effective August, 1984. The anticipated passing rates may not have been as alarming to the State Board given the experience with the field test of the GTTP-BS where actual passing rates were higher than the anticipated rates.

In July, 1984, the GTTP coordinator resigned from the Department of Education. He felt that technical considerations were being compromised for political concerns:

> . . . in part, one reason why I left the state department was because my recommendations were not being considered or followed through. I said I couldn't be a part of something that I had problems with. (105)

As the conjunctive standard was being set by the State Board of Education, the Board of Regents began deliberations about the utility of the GTTP-BS. According to a former staff member for the Board of Regents, the executive director wanted to use the GTTP-BS:

> The Regents' executive director initially intended to employ the state department test, or a somewhat altered version, for this purpose. (126)

To this end, the Board of Regents created a panel of individuals from the universities and community colleges to review the GTTP-BS. Their task was to determine whether the GTTP-BS would be a reasonable way to screen people for admission to teacher education programs. The former GTTP coordinator had assumed a position at one of Granada's universities and was asked to participate in the panel's activities. He readily agreed, saying:

> It would give me a chance to raise some issues that I tried to raise before and never had a group of individuals that would either understand or take to heart. So when we got together, I was asked many times to kind of bring the group up to date, and I would resort back to some of the information that I had kept. Here were the problems, here were the recommended actions, and here is what was done. The panel finally agreed at one point that there were enough serious questions about the GTTP and we needed to look at some alternative. (105)

What problems were identified with the GTTP-BS? Minority issues were cited as a major problem. Committee members indicated that sensitivity panels had not been convened in Granada to review items for possible bias. Minorities were failing at a higher rate than Anglos and charges of bias were coming from minority groups. The state department's failure to conduct an in-state validation of the GTTP-BS was a second problem cited by the review committee. The non-technical nature of the standard-setting process was a third problem identified by the committee.

It would appear that the GTTP-BS had serious problems and the committee's decision to look at other alternatives was well-grounded. However, there are a number of considerations that would have weakened charges against the GTTP-BS. The GDE did not use sensitivity panels from Granada to review test items, but the items had been purchased from an established item bank in a state with similar minority demographics. Minorities were failing at a higher rate than whites-- on the initial test administration. Data on ultimate passing rates were essentially unavailable and so conclusions about disparate passing rates remained unsubstantiated. Conclusions drawn in the GATR report indicated minimal bias after examining item correlation coefficients between ethnic groups. Finally, while charges of bias were asserted, no legal action had been taken against the GDE in the three years since the inception of the GTTP. Consideration of these points should have softened the committee's attacks on the GTTP-BS. The

nature of the problems with the GTTP-BS was such that efforts could have been made to correct them. Sensitivity panels could have been convened. An in-state validation process could have been conducted. Alternate standards could have been set. None of these steps was taken.

Conflict began to surface between the State Board of Education and the Board of Regents. The Board of Regents perceived the problems identified by the committee to be serious and so the Board sought legal counsel. According to a former Regents' staff member:

> Legal advice pointed out to the Regents the high degree of liability, perhaps even personal liability, if they excluded minorities through the use of a test that could not be demonstrated in a court of law to be the best available, unbiased instrument. The fear of liability was a major element in all of the Regents' deliberations. The executive director became convinced that the state department test would be 'duck soup' for any civil rights lawyer. (126)

The former GTTP coordinator indicated that the State Board of Education was unwilling to cooperate with the Board of Regents in making changes to improve the GTTP-BS:

> There was very little willingness to move or bend with regard to the test. It was 'Take it as is, don't tear it apart and try to reassemble it to look different. We are going to keep what we've got'. If the State Board had allowed the test to be changed, it would be an admittance that there were problems that they hadn't dealt with before, and that could have a negative reflection on the superintendent who is a politician (105, cp. 59).

However, the assistant GTTP coordinator indicated that the struggle did not stem from the State Board of Education's unwillingness to changes the GTTP:

> But we thought the Board of Regents' people would look at our exam, make any necessary changes or revisions

> so that it could stand the scrutiny of academia and still be acceptable to the public. We had a nice, acceptable exam for everybody, but somehow the two boards didn't get together (103, cp. 41).

Laying blame on one or both boards is secondary to the point that conflict indeed existed and was being played out. The Board of Regents argued that the GTTP-BS was technically faulty. The State Board of Education disputed the Regents' claims. Political and legal pressures added fuel to the sparks and the hope of using the GTTP-BS turned to ashes. The Regents' committee would start to consider alternatives to the GTTP-BS. A university professor who would play a major role in later events described the committee's further activities:

> And at that meeting, we began asking the question, 'What do you want people to be when they're admitted to the College of Education?'. We started out by saying what you first do is identify the skills that you want people to have when they go into the program. Once you have those prerequisite skills, then we can find out whether this test is an appropriate test for it. And if it's not, we'll build a test or select another. We collected skills from all over the place. We suggested another panel be set up to review these skills and decide. All of a sudden, we were told 'Thank you for your time' by the former executive director of the Board of Regents. (101)

Reflecting on why the panel was dismissed so suddenly, this professor remarked:

> My personal impression is that identifying the skills seemed to be such an onerous and complex task that the Regents didn't want the panel to continue with it. It wasn't a nice, quick solution. (101)

The identification of objectives, test development, piloting and refinement would consume too much time; the test had to be adopted and implemented by August, 1984. The Board of Regents needed a quick solution and a safe one. An

alternate test had not been identified by the deadline and on an interim basis, the Regents' agreed to use the GTTP-BS for admission to teacher education programs. However, if they failed the exam, they would still be admitted provisionally. This practice would continue until an alternate test was selected and approved for use. In the following months, the Board of Regents approved a test developed by Educational Testing Service: the Pre-Professional Skills Test (PPST). According to the Associate Director for Academic Programs for the Board of Regents, the PPST had two advantages over the GTTP-BS: a writing component and the expertise of ETS.

> We knew that ETS had a major, qualified staff to monitor the test, to provide us with the statistical analyses we would need to make sure the test was reliable. In other words, we had the expertise for testing. While there were some people with expertise in the state department, they didn't have the substantial expertise in testing and measurement that ETS had. (102)

This explanation was echoed by the Associate Director's predecessor:

> The Board of Regents chose to go with the PPST and align itself with ETS because of their experience and national reputation. The Regents met with the ETS lawyers who assured that they [the lawyers] would go to court with the Board. (126)

And so 1984 drew to a close. The State Board of Education had lost the fight to use the GTTP-BS as the entrance exam for state-supported teacher education programs. By rejecting the GTTP-BS, the Board of Regents effectively kept the State Board of Education from trespassing on its terrain. In the months to come, the PPST would be implemented

by the Board of Regents and the state of Granada would have two basic skills testing programs.

1985 to Present: The Pre-Professional Skills Tests

The PPST has 125 multiple-choice questions in reading, mathematics and writing, as well as a writing performance component. The test was developed by ETS as part of its National Teacher Examination programs. The PPST had been validated previously by six states and Granada would become the seventh to conduct a content validation and standard-setting study (1A). The university professor mentioned earlier was asked by the Associate Director of Academic Programs to consult and assist with the Granada standard-setting study. Using what other states had done as a guide, these two individuals prepared to carry out their task.

The first step in the study was to select a representative panel of Granada teachers and teacher educators. To this end, each superintendent and university president in the state was sent a letter asking for nominations of possible panelists. A list was generated of the nominees and their characteristics (e.g., sex, ethnicity, position and subject matter specialty) which was used as the sampling frame. According to a report published by the Board of Regents:

> The final determination of which nominees to invite to participate was made after carefully considering the results of the 1980 census which described the Granada population by ethnic group distribution. (1A)

The Board of Regents' representative expanded on the selection process:

> We selected randomly a group of 60 people: teachers nominated from all over the state and from a variety of disciplines; teacher educators from the three universities; and school administrators. We had this elaborate random selection process that went through several iterations to achieve the proper balance of ethnicity, sex, subject matter distribution, and geographic distribution. (102)

A slightly different account came from the university professor, an expert in educational measurement. While all panelists were selected with an eye towards representing all groups, the selection process was not quite random:

> Driving to a Board of Regents meeting to let them know what we were planning to do, I read the list of people and he [the Associate Director of Academic Programs] just sort of said, 'Well, we'll go with that school, and that school, and that school. So it was really a judgment sample, but a judgment sample that ensured that we had a certain number of people from all interested groups. (102)

Once selected, the nominees were sent invitations to participate in a one-day meeting in April, 1985. When a nominee did not accept the invitation, a replacement was invited whose characteristics matched the original nominee's as closely as possible (1A, p. 3).

Prior to the April meeting, plans were laid for the panel's activities. The Regents' staffer relayed suggestions and examples from ETS to the consulting professor for determining standards on the PPST. In choosing a particular cut-score method, the professor explained:

> I guess I felt constrained to do something very similar to what had been done by the other states. It was a matter of choosing say between the Ebel and Angoff methods, as both of those had been used. Of the two, I

> thought that Angoff was much more feasible as Ebel would be too difficult for a lay panel to do. I guess I really didn't feel there would be that much difference in the results. (101)

Specific instructions for the panelists were prepared for using the Angoff method and the other tasks to be performed. Each panelist would be asked to make three separate judgments for each item in the PPST:

1. The extent to which the knowledge or academic skill tested in the question is relevant to competent performance as a teacher in Granada.

2. Whether a typical applicant who graduated from a Granada high school and/or who has met the course prerequisites for admission to a teacher education program had an opportunity to acquire the knowledge or skill which is required to answer the test question.

3. The proportion of marginally qualified applicants for a teacher education program who you would expect to answer the test questions correctly. (1A)

In addition, instructions for identifying biased questions were developed. Panelists were to identify any questions that would result in a bias against a member of an ethnic minority. They were to write detailed explanation of the bias problem for any question identified.

When the panel was convened, the 56 members were introduced to the day's tasks by an ETS staff member. Following an explanation of the procedures by the Granada professor, panelists were given an opportunity to practice on a reading test item similar to those found on the PPST:

> . . . then we discussed their judgments to be sure they all were in the same frame of mind. Once the dummy item was judged, they were told to do the PPST items in the same way, on their own. (102)

Discussion of the standard-setting procedure centered on the question, 'How many, out of 100 marginally qualified applicants for a teacher education program, would be able to answer the item correctly?'. (1A). Specifically, panelists were concerned about basing their judgments on what would happen or what should happen. The ETS consultant explained to the group that:

> . . . a panelist's judgment about what an idealized conceptual group of marginally qualified individuals would do and a panelist's judgment about what a group of marginally qualified real individuals should do amounted to the same thing. (1A)

With that pronouncement, panel members made their individual judgments about each multiple choice item on the PPST. These judgments were analyzed after the panel was dismissed, summaries of which were available in a Board of Regents' publication (1A). With respect to judgments of relevance and opportunity to learn the skills required in the PPST items:

> It is clear that 68% or more of the items on every test were individually judged by 75% or more of the panelists as very relevant or moderately relevant, and every item on every test was seen by 75% or more of the panelists as at least slightly relevant. It can also be seen that every item on every test was seen by at least a majority of the panelists as requiring knowledge or skills which applicants would have had a chance to learn. Indeed, no fewer than 87% of the items on any test were seen this way by 75% or more of the reviewers. (1A)

Concerns for possible bias were investigated. Thirteen panelists made 28 comments about 25 of the PPST items. Only

two items received more than one comment. (1A) Those items had been removed from all currently used forms, except for the one used by the panelists.

The authors reached the following conclusions based on the preceding results:

> There is considerable evidence that a substantial majority of all panelists and of most of the relevant subgrouping of panelists judged the individual items on all the tests as relevant to the performance of students in educational programs and of teachers in Granada. Thus, it can be said that the PPST has considerable content validity for use as an entrance requirement to any of the state universities. It is unlikely that any other currently available test would show greater content validity than the PPST. Since no potential bias was identified by any panelist in 80% of the items and only two items were called into question by more than one reviewer, it would seem reasonable to conclude that the tests are likely to be as free from bias as current test construction technology would allow. (1A)

The passing scores on each test in the PPST were calculated for each panel member. For each test, scores were averaged after determining that the distributions were fairly normal. Table 8.3 displays the passing scores determined for each of the three multiple choice tests and the essay writing exam. Passing scores were translated into scaled scores which ranged from 150 to 190. The translation produced three scores- the multiple choice writing score and essay score were weighted (60-40) and combined into one scaled score:

Table 8.3

PPST Standards Derived by Angoff Cut-score Method

Subtest	Number of items	Raw score	Percent correct	Scaled score
Reading	40	26	65%	176
Mathematics	40	25	63%	175
Writing				177[a]
multiple choice	45	27	60%	
essay	12	9	75%	

[a] Writing scaled scored derived from combination of multiple choice and essay cutscores.

When the passing scores were to be presented to a Board of Regents sub-committee, estimates were provided of failure rates based on administrations in other states during two periods of time. In addition, alternate passing scores were available; the original passing scores less by one and two standard errors of measurement. Estimated failure rates were provided with these scores as well. The impetus for including scores minus one and two standard errors came from an earlier meeting with the Board of Regents sub-committee. When approving the plans for the validation study, questions were

raised about errors of measurement. The university professor reflected on their concerns during his interview:

> And at that time, the Board of Regents started asking questions about error of measurement. They weren't sophisticated questions, they just recognized that someone could have had a bad day when taking the exam--that sort of thing. That led me to say if you want to be sure that a large portion doesn't fail just because of error of measurement, we can take a cutting score and go below by one or two standard errors. They seemed intrigued with that idea. (101)

The Board of Regents sub-committee surveyed the standards and failure rate information in a form similar to that shown in Table 8.4.

Table 8.4

Estimated PPST Fail Rates for Candidate Standards

Possible passing scores	Test		
	Reading	Math	Writing
Panel judgment	176	175	177
Fail rates			
1983	52%	45%	55%
1984	55%	49%	59%
Cut-score minus one SEM*	173	172	174
Fail rates			
1983	39%	34%	35%
1984	40%	36%	34%
Cut-score minus two SEM	170	169	171
Fail rates			
1983	27%	25%	24%
1984	25%	21%	17%

[a] SEM indicates standard error of measurment.

The Regents decided to use the scores that were one error unit below the standards produced by the panel. When asked why the Board elected to lower the panel standard by one error of measurement, the Regents' staff member explained:

> We wanted to avoid making a Type II error where someone is incorrectly labeled incompetent. So, we subtracted one standard error from the panel standards. We felt that subtracting two standard errors would result in a much higher passing rate, making the exam a little too easy. (102)

As can be seen in Table 8.4, a standard of 173 on the PPST reading test would have failed 39-40% of the examinees in the 1983-84 data. A standard of 172 on the mathematics test would have failed 34-36% and 34-35% of the examinees would have failed the writing test with a standard of 174. Interestingly, no information was available on conjunctive failure rates even though examinees would be expected to pass each of the three exams. Before we discuss consequences that followed the setting of standards on the PPST, we will describe a number of policies and practices related to the test's use.

1. The PPST is administered six times during the academic year at state university and college testing centers across the state.

2. A charge of $30 is levied for the initial administration. Charges for retakes are pro-rated: $20 for one test; $25 for two; and $30 for retakes on all three tests.

3. Two hours and thirty minutes are allowed for administration of the PPST.

4. Optional remedial courses, offered by the community colleges on university campuses, are available for those failing the PPST.

5. Examinees who failed the exam are allowed unlimited retakes.

Data on examinee performance were collected between September, 1985, and February, 1986. These data included only the scores of students who authorized the release of their scores to the Granada institutions. Initial passing rates on the individual tests were slightly higher than those anticipated from the 1983-84 data. The anticipated and actual passing rates for each of the three tests are shown in Table 8.5.

Table 8.5

Anticipated and Initial PPST Pass Rates

Test	Anticipated pass rate	Actual pass rate
Reading	60-61%	73%
Mathematics	64-66%	74%
Writing	65-66%	73%

The overall passing rate on the initial administration of the PPST to 1073 college students (university and community college) was 56%. (10, 1P) Of the 477 who failed the PPST,

221 (46%) failed only one of the three exams and 119 (25%) failed all three tests.

Disparate passing rates were observed among ethnic groups on the initial administration of the PPST. First time passing rates for the various ethnic groups are shown in Table 8.6.

Table 8.6

Initial PPST Pass Rates by Ethnic Group

Ethnic group	#	Pass rate	
		Number	Percent
Black	5	0	0%
Hispanic	93	32	34%
Native American	29	1	3%
Asian	8	5	63%
White	927	550	59%
Other	11	8	73%
Total	1,073	596	56%

As with the GTTP-BS, cumulative passing rates on the PPST were nebulous. Retake information was limited to those students who authorized release of their scores to Granada institutions, as well as those who retook the test within the

six-month data collection period. Because of these limitations, conclusions could not be drawn about 274 (27%) of the original sample of 996 university students. Of the remaining 722, 626 had passed after one or two attempts. Only 96 had failed after two attempts. No information was available regarding further retakes among examinees who failed on their second attempt. Table 8.7 summarizes this information in tabular form.

Table 8.7

PPST Cumulative Performance during a Six-month Period

Test administration	Cumulative pass rate	Cumulative fail rate	No available information
First	56%	44%	0%
Second	63%	10%	27%

Cumulative passing rates among ethnic groups were even more difficult to discern. Information comparing Hispanic and white populations were available in a report to the Granada Board of Regents (1P). Table 8.8 shows the performance of examinees on the second attempt, classified by ethnicity.

Table 8.8

PPST Second Attempt Performance by Ethnic Group

Ethnic group	# who failed first try	# of second attempts	Second attempt: Passed #	Passed %	Failed #	Failed %
Hispanic						
Math	44	10	3	30	7	70
Reading	39	5	2	40	3	60
Writing	40	8	2	25	6	75
White						
Math	203	67	34	51	33	49
Reading	217	77	43	56	34	44
Writing	218	80	38	48	42	53

Data for each test include the number who failed on the first try; the number who authorized release of their second attempt scores; and the numbers passing and failing. When comparing Hispanics and whites, it appears that Hispanics had lower proportions of retake attempts, as well as lower passing rates on the three tests. This observation cannot be extended to the Hispanic retake population due to the selection phenomenon operating in the score release process. Table 8.8 does not include combined performance, that is, we cannot

discern duplications in the numbers reported to know how many passed the entire PPST.

Before examinee performance on the PPST had been analyzed, it was clear that the Board of Regents would not be the only agency interested in the results. The Granada Department of Education and the State Board of Education were eager to know what happened. According to the assistant GTTP coordinator:

> Well, it's just natural. Like I said, we were hoping that they'd take our exam and make it better- acceptable to everybody. But they didn't. So naturally, there's going to have to be some comparison. It would just be nice to say 'Well, look at our test. If it's so bad, why are more people passing it?' I don't know how you would do it, but somehow, you could come up with a percentage for each one of the cut scores on the PPST. And if our people are passing our exam at 80% and they're failing more of them with lesser percentages on the PPST, you could say that and let people infer what they want. (103)

Traces of agency conflict remained. Perhaps the GDE and Board of Education looked for some kind vindication after losing out to the Regents' technically-sophisticated, legally-defensible, and no doubt, costly test. According to this GDE employee, higher fail rates on the PPST would reflect poorly on the Board of Regents. The Board of Regents would get its turn on the political hot seat.

The Granada Board of Regents received reports on PPST performance in May, 1986. The university professor responsible for the analyses and report cautioned the Board about the inconclusive nature of the findings, especially those related to minority performance. According to his report:

> The number of students includes only those who authorized the release of their scores. The number of students in most of the ethnic categories is too small to justify the drawing of any conclusions. (1P)

Further research would be necessary before drawing firm conclusions about the impact on minorities as well as the general examinee population. At the time, the Board of Regents did not authorize any further study of the PPST. When questioned in September, 1986, the university professor confirmed that indeed, followup studies had yet to be commissioned by the Regents. Two months later, however, a student newspaper from one of the three universities reported that the Board of Regents commissioned a study to analyze the high minority failure rate on the PPST (1Q). The Board's action followed its November meeting in which criticisms were leveled against the PPST and its impact. Minority enrollments were reported to have varied since the PPST became a requirement in September, 1985. Other criticisms came from people attending the November meeting. According to this newspaper's account:

> At the November 7 Board of Regents meeting, a minority students with a 3.4 grade point average told the Regents she has taken the test four times unsuccessfully. An English instructor at the same university, told the Regents he took the test and had trouble with it. 'How can you expect some freshman to pass it when three English professors had trouble?' he asked. He said the passing score on the test was determined arbitrarily and suggested lowering the passing score. The Director of Minority Affairs at another university said she was at the Regents' meeting and was aware of the concern but has heard no complaints from students regarding the test. (1Q)

Although the Regents' test was promoted as technically superior to the GTTP-BS, it was not spared from attack.

Indeed, the complaints registered against the PPST were similar to those against the GTTP-BS. We do not know how influential these criticisms were on the Board's decision to conduct further studies of the PPST. However, the timing of the Board's action would suggest that the complaints made some contribution to its decision.

Results of the evaluation were presented three months later to the Board of Regents (1R). By ethnic group, success rates after four attempts are shown in Table 8.9.

Table 8.9

Cumulative PPST Pass Rates after Four Retakes

Ethnic group	Cumulative passing rate
Black	10%
Hispanic	49%
Native American	9%
Asian	73%
White	72%
Overall	67%

Board members expressed disappointment and concern over the low and disparate PPST passing rates. Three suggestions were made by the President of one of Granada's universities

and the Board of Regents agreed to study the suggestions for possible future action:

1. Base admission on passage of two of the three PPST subtests.

2. Lower the passing scores on the PPST.

3. Allow students to substitute the GTTP-BS in lieu of the PPST.

While studies of PPST alternatives continue, students are still required to take the examination before entering any of the state's teacher preparation programs. The other Granada test, the GTTP-BS, continues to be used for applicants to private teacher education programs as well as out-of-state teachers and home instructors seeking certification.

CHAPTER IX

INTEGRATION AND CONCLUSIONS

Answers to the primary and supporting research questions are woven into each of the five site narratives. Each includes descriptions of the origins of the testing program, its intended purposes, the manner by which standards were determined, the decision-makers involved in the standard-setting process, and to the extent possible, the consequences that stemmed from the tests and standards. This chapter will not reiterate answers to these concerns. Rather, sections in this chapter are devoted to three issues: the resolution of the working hypotheses, the advancement of cross-site propositions, and a discussion of implications for policy, practice and research.

Resolution of the Working Hypotheses

As indicated in Chapter 3, the utility of the working hypotheses was deemed limited for a number of reasons. Nevertheless, this work would not be complete without a concluding discussion of the notions that helped to shape the initial intent of the study. The eight hypotheses, evenly divided among the two posited theories, addressed the manners by which standards may be determined. The rational decision-making hypotheses posited a unitary decision-maker (whether individual or group) who sets standards after careful examination of goals, alternatives and their efficiency,

potential use of technology and expected outcomes. The decision-making process was hypothesized as organized, consensual, and satisfactory to all involved. On the other hand, the bureaucratic politics hypotheses characterized the decision-making process as one of political compromise among decision-makers who held varied goals, stakes, stands and power in the situation. These disparities were hypothesized to foster dissension and conflict during the standard-setting process. Ultimately, the decision or standard would be determined largely by those who wielded power or bargained most effectively.

If one were to imagine a continuum of decision-making theories, these two sets of theories and hypotheses could very well represent its opposing ends. Given the contrasts, it seemed reasonable to expect individual sites and their standard-setting processes to conform to one or the other theory. Although the sites did not mirror the theories and hypotheses as closely as initially expected, enough evidence was uncovered to make some general observations about the standard-setting processes in these five sites and their resemblance to bureaucratic decision-making theories.

Using the aforementioned continuum as a heuristic, one might attempt to locate the sites in relation to the two sets of hypotheses. Among the five sites, Victoria most closely approximated the rational decision-making model. Technical Advisory Committee (TAC) members unanimously indicated that the decision-making process involved a unitary goal, search

for alternatives, application of technical methods, and consensus. Their unitary goal was to produce a standard using the best technical methods available without the interference of political considerations. With this goal in mind, they evaluated the various cut-score methods, narrowed the list of possible alternatives, and decided to employ a number of them. Averaging the results of the various cut-score methods was expected to produce a more reliable estimator of the "true" cut-score. All TAC members emphasized the collegial and professional atmosphere surrounding the decision-making process and no one indicated that members were drawn into conflict because of differing goals, stakes, stands or power.

Identifying the site among the five that most closely approximated the bureaucratic politics hypotheses was a more difficult task. If one were to view pervasive and recurring conflict as a hallmark of this model, then perhaps Granada is the site of choice. Little can be said about the initial standard-setting process on the GTTP-BS as only second-hand accounts could be elicited. When the time came for revised GTTP-BS standards, glimpses of conflict were evident when the GTTP Coordinator argued for the use of cut-score methods and the State Board of Education disregarded his recommendations and set the cut-score according to anticipated fail rates. Nevertheless, among the Board of Education members, both of the standard-setting processes could have been marked by a rational evaluation of outcomes and consensus. The standard-setting process for the PPST was equally difficult to discern

as individuals involved in the application of the Angoff method were not accessible. However, from other informed accounts it appears that the individuals had little opportunity to chart their own course and interact with one another. Rather, they were asked to operate according to specifications laid down by the technical representative of the Board of Regents. Little evidence was generated about the deliberations of the Board of Regents when the cut-score was presented to them. The ultimate modification of the standard by the Board of Regents may have been the result of rational evaluation of potential consequences or the product of political maneuvering-- there simply was no evidence one way or the other. In sum, the evidence is equivocal and does not seem to support Granada's location on the continuum. However, if the narrow focus on the standard-setting process is broadened to include the activities and interactions between the two Granada boards, the bureaucratic politics hypotheses become relevant. The 1984 legislative mandate forced these two agencies to work together to produce a competency test for applicants to the state teacher education programs. Board of Regents (BOR) members anticipated costly losses if they were to adopt the State Board of Education (SBE) test. Fears of litigation were cited by informants as uppermost in the minds of BOR members. SBE informants also indicated that political and territorial concerns may have increased resistance to the GTTP. Whatever the reasons, the BOR rejected the SBE test by citing technical inadequacies and consequently championed the

selection of the PPST, a technically superior test. The SBE and Department of Education members were not able to combat the bad press given to the GTTP in part because they lacked the technical prowess that ETS and BOR representatives claimed and used so effectively. As a consequence, the BOR came out the victor and the SBE the loser in the battle over tests and standards in Granada.

The remaining three sites can be placed between Victoria and Granada on the theoretical continuum. The lack of sufficient evidence in Lucerne with regard to either theory leaves little choice but to place it in the center of the continuum or leave it off completely. Finally, the arrangement of Waterford and Brittany is open to question. Indeed, glimpses of rational decision-making and bureaucratic politics were evident in both Waterford and Brittany. Waterford informants frequently pointed to group efforts to produce policies and practices related to the promotional gates tests and standards. Many indicated that the selection among alternatives was consensual and followed only after careful considerations of efficiency, use of technical methods, and expected outcomes. However, a number of informants provided evidence of disagreement and conflict. In drafting the test and standards policies, a number of committee members argued against the use of tests for retention decisions. Further, during standard-setting sessions, some committee members were reported to be concerned about the implications of the tests and standards.

Specifically, some were concerned about remediation, others were concerned about logistics, and still others were concerned about adhering to the techncial method. Despite the disparate stakes, informants were reluctant to discuss specific conflicts and all indicated that efforts were made to achieve consensus on the standards produced from the deliberations.

Brittany's standard-setting process was difficult to reconstruct completely due to the lack of informants who were directly involved. However, the central figure in the standard-setting process, the former UB Registrar, provided substantial evidence that was corroborated by other primary and secondary informants. The standards on the NWET were determined after careful consideration of the relation between the test and performance in writing courses. The standard was determined in a hierarchical manner as the UB Registrar submitted his recommendation to the Vice Chancellor for Academic Affairs who made the ultimate decision. Although elements of rational decision-making were evident, so too were indications of bureaucratic politics. Specifically, BSU representatives spoke against the use of the NWET for admission, denigrated the appearance of the standard, and eventually lobbied for the removal of the requirement. To date, their efforts have been in vain.

Working Hypotheses: Summary and Conclusions

The resolution of the working hypotheses proceeded by placing the five sites along a continuum of decision-making theories. As was seen, the ends of the continuum were identified most readily and distinctions among the "middle" sites were more difficult. The utility of the working hypotheses in this study was constrained when examining the decision-making process in these relatively less extreme sites. Other mitigating factors described in Chapter 3 included the role of the investigator, availability of informants, and the historical nature of the study. These theories and hypotheses might be employed more effectively in a contemporary study of the standard-setting process (e.g., in progress) drawing on methods of observation (participant or otherwise) as well as in-depth interviews. Nevertheless, these theories and hypotheses did not accommodate or fit the data generated in this study and so they were set aside for a more grounded analysis of competency testing and standard-setting in the five sites. The next section describes the fruits of the cross-site analyses conducted apart from the working hypotheses.

Cross-site Propositions

As explained in Chapter 3, the cross-site analyses were initially based on observations and themes drawn from the individual site analyses. After careful comparative examination, these and other themes were developed into seven

cross-site propositions. Following the list of these propositions, each is discussed using site-specific instances to support, illustrate, and qualify.

1. Standards are influenced by normative considerations when agencies must deal first-hand with testing consequences.

2. As standards are erected, safety nets are strung up to catch those who fall.

3. In setting test standards, the more technical looking approach is ascribed greater credibility. The extent to which technical plans are actually implemented is mitigated by political or practical considerations.

4. Agency concern about minorities and competency testing is reflected in efforts to build unbiased and fair tests. The attention paid to unbiased test development contrasts sharply with that given to minority performance and impact.

5. Organizational efforts are most visible and detailed during earlier phases of competency testing reforms than later phases.

6. Schools and educational agencies honor court rulings that mitigate the potentially severe consequences of competency testing. Fair warning entered the classroom even earlier than it entered the workplace.

7. Competency tests and standards function primarily as symbolic gestures rather than instrumental reforms.

Proposition 1

Standards are influenced by normative considerations when agencies must deal first-hand with testing consequences. With the exception of Victoria, test standards were largely determined by normative considerations of how many should pass and how many should not. These considerations varied in specificity, ranging from intuitive estimates to empirically-based projections of pass-fail rates. Intuitive estimates were heeded in Lucerne where educators wanted to provide a "fail-safe" mechanism for the criterion of 100% objectives correct. This mechanism would protect the unknown but estimated large number of students destined to fail by the original criterion-based standards. Empirically-based projections of fail rates were available and examined by standard-setters in Brittany, Waterford and Granada (both the GTTP and PPST). Standards in these sites were determined after careful examination of possible fail rates. Even those sites that used established cut-score methods, Waterford and Granada (PPST), consulted normative data. In Waterford, normative considerations were integral to the Contrasting Groups method. However, such considerations were not a formal part of the Angoff method used with the PPST in Granada. Instead, normative issues were addressed by the use of the standard error of measurement (SEM) following concerns expressed by the Board of Regents that too many might fail the test with the standards produced by the Angoff method. The eventual suggestion to lower PPST standards to increase the

initial passing rate adds further evidence to the mitigating role of normative considerations.

Victoria stands as the exception to the conclusion about normative standards. Decision-makers employed cut-score methods and only after setting test standards did they examine potential fail rates. Furthermore, no modifications were made in light of these anticipated fail rates. In the site narrative, we discussed earlier events in Victoria that probably influenced the standard-setting environment, but another related explanation may help to account for the absence of normative concerns with the VEHSG standards.

Unlike consequences related to teacher testing and certification, the Victoria State Board of Education and its agency did not have to deal first-hand with consequences stemming from the high school graduation test. The State Board and Department of Education were once-removed from actual instruction; local school districts were the buffers between these state-level agencies and students. Specifically, local school districts had to bear the responsibility for remediating students who failed the VEHSG (at the district's own cost). The regulatory nature of the state-level agencies would permit the setting of non-normative standards as any blame for failures would revert, ultimately, to individual school districts. The point is not that the state-agencies held a laissez-faire or antagonistic attitude toward fail-rates and their impact on local districts-- they did not, as evidenced by efforts to prepare districts for the

test and assist with plans for remediation. Instead, the absence of direct responsibility or accountability by the state agencies allowed the derivation and application of criterion-based, non-normative standards.

Each of the other four sites had to deal first-hand with testing consequences. Waterford and Lucerne had to find ways to remediate and retrain the students who failed the exams. Similarly, Brittany and Granada (PPST) had to consider potential threats to enrollments that might result from competency testing. The Granada State Board of Education would have had to account for a shortage of certified teachers if the GTTP standards had been set too high. Indeed, such was the case in Victoria where the State Board of Education confronted teacher shortages and a lawsuit filed by outraged minorities as a result of the teacher test.

Proposition 2

As standards are erected, safety nets are strung up to catch those who fall. Safety nets take the form of repeated test-taking opportunities, modifications of standards, alternatives to the tests or standards, provisions for overruling test results, and examinee exemptions. Whether or not safety nets were acknowledged as such at the sites, they were apparent and commonplace. Multiple attempts were available at all of the five sites, but most common in Lucerne, Victoria, and Granada. Examinees in Victoria had at least four retake opportunities before they graduated; Lucerne examinees had as many as eleven. Prospective teachers in

Granada could take either the GTTP or PPST as many times as they wish or could afford (the charge for retakes on the PPST is two to three times the cost of taking the GTTP). Brittany and Waterford both allowed for multiple attempts, but retakes were less frequent. Brittany students could retake the NWET, but other options or safety nets were provided. Waterford elementary students could retake a test after special permission had been granted. Like Brittany, Waterford erected other safety nets.

Modifications of standards serve as another type of safety net. In two sites, Granada (PPST) and Waterford, the standard error of measurement (SEM) was used to modify standards determined from cut-score methods. In Waterford, two SEMs were subtracted from the passing scores to arrive at the gray area scores. In Granada, the standards derived by the Angoff method were decreased by one SEM to arrive at the final PPST standards.

Alternatives to the standards or tests is another kind of safety net used in Lucerne and Brittany. If examinees failed to meet the Lucerne standard of 100% objectives correct, they could still pass the test if they answered 70% items correct. Brittany allowed students to submit ACT-English scores in lieu of taking the NWET.

A fourth kind of safety net is the ability to overrule or disregard test results. A Waterford policy empowered a teacher or a committee of educators to review performance and rule against retention on a case by case basis. An informal

Waterford policy permitted promotion of students who fail a promotional gate test if they had been retained in a preceding non-promotional gate grade. Brittany post-secondary institutions may admit up to 5% of their freshman classes with students who fail to meet one or more of the admission requirements (including the NWET).

A final safety net, exemption, was strung up in two sites for special educations students. In Waterford, the Individual Education Program (IEP) lists the specific competency tests, if any, that a special education student must take. In Lucerne, the IEP may exempt a special education student from the high school graduation test. However, this safety net is not extended for high school students in Victoria. All special education students who desire a regular high school diploma must pass the VEHSG.

The rigging of safety nets around tests and standards serves numerous functions, the most obvious of which is to keep examinees from harm. Without these safety nets, examinees would have failed in staggering numbers in each of our sites. The absence of safety nets surely affects minorities. But given the paucity of data, we can only speculate as to the extent of these effects. In the following way, safety nets also serve the organizations and professionals within these agencies. Organizations would be hard-pressed to accommodate students who failed to meet the mark. For districts, the need for classroom space would increase exponentially and already strained budgets would

collapse under the weight of new personnel hired to serve these students. Safety nets allow organizations to point to tough standards while, maintaining business as usual. For professionals, the use of safety nets reclaims and fortifies professional discretion, an integral element of schooling threatened by a centralized testing and decision-making policy. Indeed, the use of these safety nets is an eminently reasonable response to the havoc certain to be created by thoughtless implementation of tests and standards.

Proposition 3

In setting test standards, the more technical looking approach is ascribed greater credibility. The extent to which technical plans are actually implemented is mitigated by political or practical considerations. The most compelling examples of this proposition were found in Granada. The selection of the PPST was predicated on allegations that the GTTP-BS was technically inadequate. The Board of Regents attacked the validity of the GTTP-BS as well as the test's arbitrary standards. In contrast, the PPST could be subjected to an in-state validation under the guidance of ETS. Similarly, standards could be set using a technical approach espoused by ETS. In light of these technical possibilities, the PPST was touted as superior to the questionable GTTP-BS and ultimately selected for use. Unfortunately, the technical superiority of the PPST did not shield the Board of Regents from problems. The low minority pass rates and retake attempts prompted them to reconsider current practices. Some

of the alternatives included lowering the technically derived standards and permitting students to substitute the alleged technically weaker GTTP-BS for the PPST.

Other incidents in Granada illustrate how technological efforts were constrained by political or practical concerns. To comply with the 1984 legislative mandate, the State Board of Education met to determine conjunctive standards for the three subtests of the GTTP-BS. The former GTTP coordinator urged the Board to use technical cut-score methods. The State Board of Education members acted against the coordinator's advice and set the conjunctive standards as they had done earlier-- normatively, i.e., by assessing potential fail rates. Another incident involved the Board of Regents and the committee it created to examine the utility of the GTTP-BS. After the committee had rendered its verdict on the GTTP-BS, members began to speak about basic test development (e.g., identification and validation of objectives, item writing, and the like). Such activities were clearly time-consuming and hence, unacceptable to Board of Regents members who were under the gun to comply by the legislated deadline. As a result the committee was dismissed abruptly and Board of Regents selected the PPST.

The use of technical methods was not restrained in Victoria. Rather, their use was congruent with political and practical considerations. The problems brought by the teacher test were said to be due, in part, to the lack or inadequacy of technical efforts. Ensuring the technical adequacy of the

VEHSG would be one major step toward avoiding a repeat of similar problems. The more scientific and technical the methods, the better. Hence, politicians and managers would allow technicians to ply their trade without interference.

Proposition 4

Agency concern about minorities and competency testing is reflected in efforts to build unbiased and fair tests. The attention paid to unbiased test development contrasts sharply with that given to minority performance and impact. The debate surrounding competency tests and adverse impact on minorities will not be settled by findings from this study. What little evidence was generated is far from unequivocal; while differential passing rates were observed, issues of impact beyond the tests themselves were left unexamined in the sites. Indeed, the point at which minority issues received the most careful attention was during test construction and validation. Schools and agencies expended minimal effort in analyzing minority impact unless forced to do so.

In all of the study sites, organizational attention was directed to concerns of test bias. During and following test development, items were subjected to judgmental and technical reviews in Waterford, Lucerne and Victoria. Judgmental reviews usually entailed committee or group reviews of items. These panels were composed of members from various ethnic groups and organizations, representing both educators and lay people. Generally, their task was to examine items for misleading language, stereotyping, and ambiguous

interpretations. Technical reviews involved a comparison of item level statistics (e.g., difficulty and correlation with overall test performance) by ethnic group to determine whether items functioned similarly or disparately. Tests purchased by Brittany and Granada (PPST) had been subjected to review by the test publisher, but additional bias reviews were conducted at the sites as well. The Granada GTTP-BS test items were purchased from an established item bank in which items had been screened for potential bias. In sum, every effort had been made by the sites to eliminate as much bias as possible from the competency tests.

Aside from issues surrounding test bias, the ability to examine impact on minorities was limited. Published data were available from Waterford and Granada, but not from the other three sites. Consequently, informants at all sites were asked to relate what they knew. Hence, the conclusions made in this section are drawn from both informants and documents but constrained by the two as well.

Documentation from Lucerne, Victoria, and Brittany contained no minority passing rates. Brittany did not provide minority passing rate data and Lucerne informants indicated that such data were not kept. Victoria informants indicated that minority passing rates were lower, but they did not say by how much. In addition, when the Victoria School Journal published VEHSG passing rates, no mention was made of minority performance, a curious omission for a district and a state with a large minority subpopulations.

Waterford and Granada provided glimpses of minority test performance. Waterford routinely publishes a summary report of promotional gate testing results and a copy of the 1985 edition was obtained during data collection (3F). Passing rates were detailed for Native American, Black, Asian, Hispanic and White subpopulations for each grade. Additionally, the percent of students in each scoring group (e.g., fail, gray area, and pass) were reported for each promotional gate subtest. Thus, one could examine and compare subtest performance across ethnicity. Two of the four promotional gates (grades K and 2) had standards in place by 1985. On these tests, passing rates for Native Americans and Blacks were consistently lower, ranging from nine to twenty-three percentage points lower than white passing rates. These two groups account for the largest proportion of minorities in Waterford schools. Asians and Hispanics fared better. Asians performed as well as or better than whites, while Hispanics passed at a lower rate on the Kindergarten test and higher on the grade two tests. Since standards had not been set on the grade five and seven promotion gates, Waterford published performance according to quartiles. On the fifth grade tests, a smaller proportion of Native Americans, blacks and Hispanics scored in the upper two quartiles and a much greater proportion of students from these groups scored in the lowest (first) quartile. The same pattern held true for the seventh grade tests.

Minority passing rates for the two Granada tests were available for examination. Minority passing rates on initial administrations of the GTTP-BS were much lower than that of whites, ranging from thirty-one to fifty percentage points lower. No data were available to illuminate any changes on cumulative administrations of the test. A comparable situation was evident with the initial administration of the PPST. With the exception of Asians, minorities passed at much lower rates, ranging from twenty-five to fifty-nine percentage points lower than the white passing rate. The latest Board of Regents study, prompted by minority criticism, revealed disparate success rates after four attempts between minorities and whites. Asian students stood as an exception; they passed at a rate comparable to whites.

The picture of minority impact becomes even less clear when one tries to examine test-related decisions. With the exception of the Waterford kindergarten test, there was no information about how students were slotted into the pass, retain, and review categories by the district decision rule. Moreover, the kindergarten data were not published in the report by ethnic group. Hence there was no way of determining whether minorities were recommended disproportionately for review or retention. Some data generated in the Waterford evaluation of retention services indicated that compared to test performance, a disproportionate number of minorities (and males) were retained in kindergarten. However, any conclusions drawn from these observations must remain

speculative. In Granada, the comparison between Hispanic and Anglo performance indicated that Hispanics were less likely to retake the PPST after failing it on the first try. A Granada university official stated that Native American retake behavior followed a similar pattern.

Inconclusive and relatively unexamined data constrain any attempts to understand thoroughly the impact of competency testing on minorities. Indeed, the ambiguity is disconcerting, but not unusual when one considers competency testing reforms in general. As with the earlier phases of reform, organizational attention to minority concerns like test bias was structured, detailed, and frequently technical. After test administration, the focus blurred and one saw little of the same structure and detail used to examine minority test performance, subsequent decisions based on this performance and other concerns of impact.

What might account for this contrast? One explanation focuses on the different meanings attached to the term "bias". One definition existed during judgmental reviews of tests and their items. Offensive, derogatory, or misleading words were deemed evidence of bias and the few culprit items were struck from the tests. Another definition was basic to the technical reviews for test bias. Essentially, psychometricians defined biased items as those that "behaved" differently than the majority of the test items. When the item-level statistics indicated a mismatch the items were purged and the tests were heralded as bias-free. Unwittingly, technicians' definitions

of bias and non-bias quickly found favor with politicians and managers. Using these narrow conceptions of bias, agencies could announce that tests had been scientifically examined and pronounced clean. Unfortunately, such pronouncements satisfied minorities only until they realized that their members were passing at alarmingly low rates. To minorities, disparate group differences on the tests were evidence of bias and unfairness; a situation left unexamined by technicians and conveniently unaddressed by politicians and managers.

Proposition 5

Organizational efforts are most visible and detailed during earlier phases of competency testing reforms than later phases. In all five sites, organizational activity was heaviest during the early stages of the competency testing reforms. Committees of educators were impaneled at every site; in addition, citizens were engaged to participate in committees at two of the sites. Depending on the urgency of the task, these committees met anywhere from a few months to years to discuss, recommend, plan, initiate or implement testing programs. Public and internal hearings were held to inform interested parties of committee progress. Agency approval was given to committee recommendations. Documents were produced in which committee decisions were detailed. Efforts were made to implement committee plans and frequently, committees oversaw the starting of programs. Although these groups varied in composition, life span, and charge, each

played a visible role in the realization of competency testing reforms.

Their contributions were generally technical. Objectives were identified, negotiated and approved. Test items were bought or written, reviewed for bias, piloted, modified and validated. Policies were created for test administration, examinee eligibility, and use of results. Standards were set after careful application of technical methods or consideration of potential consequences. Tests were administered with care to maintain standardization and security. Report forms for test results were designed painstakingly and distributed quickly to examinees. Remediation and retake options were made available for those who failed the exam. Initial passing rates were figured and published.

Attention to structure, detail, and technical issues was far-reaching in the early stages of these competency testing reforms. However, in all five sites, such activity waned as time passed. Indeed, so much time and energy had been invested in the start-up of the testing machine that there seemed to be little left over to gauge and evaluate the effects of what it accomplished. Or perhaps, its actual effects were incidental.

The deterioration in organizational attention to the testing reforms is most evident right after the assessment of passing rates. Routinely, initial passing rates were figured at all sites. Minority passing rates were addressed by the

agencies as well. However, ultimate passing rates were not determined as easily or uniformly. Only those sites with relatively intact examinee groups (e.g., Class of 1985) collected and reported ultimate passing rates. Other sites had piecemeal data or did not produce such information. Beyond the fairly tangible or technical outcome of passing rates, agency attention to consequences was generally vague, impressionistic and haphazard; a disturbing contrast to earlier efforts. Planned evaluation efforts were scant or focused on mundane, peripheral questions, questions that could be answered using available technical expertise. The more complex and relevant questions of impact and utility were ignored.

Three noteworthy examples will be used to support this conclusion. The first example is drawn from the sites that tested for high school graduation. No investigations were initiated or planned to examine the relationship between test performance and earning of Carnegie units, a pre-existing requirement. Although Lucerne officials stated that all who earned enough credits passed the competency test, information about other aspects of the relationship between requirements was unavailable. Officials in Victoria could not even supply the rudimentary information that Lucerne provided. Individual school districts were said to be responsible for keeping track of such information and it was not available at the state level. The second example deals with the coupling between test performance and decision-making. In Waterford, we were

limited to inferences about test performance and grade promotion. In like manner, college or teacher program admission decisions linked to test performance had not been studied at Brittany and Granada. It appears that no one really knows what happens to examinees as a consequence of the competency tests whether or not the students pass or fail. The final example addresses unstudied effects on examinees. Informants provided vague and impressionistic explanations of shifts in dropout rates, attitudinal changes, and impact of cost and opportunity. However, these issues were not the focus of systematic examination in any of the sites.

The one planned evaluation that addressed matters beyond fail rates was conducted in Waterford. The evaluation was designed to examine the retention services delivered to students retained in grades K and 2. The evaluation focused on the effects of retention services on student achievement and teacher perception of achievement and achievement-related behaviors. However, two compelling questions went unanswered. The link between test performance and promotion decisions was never documented, but simply inferred by the district. Second, the impact on the pupils of retention versus promotion was addressed loosely. Attempts were made to compare retained students with promoted students on later promotional gate performance. However, since the students were not tested on a common metric, conclusions were merely speculative.

In one site, controversy precipitated evaluation efforts that earlier had been abandoned. An initial evaluation of the

PPST passing rates by ethnic group was requested by the Granada Board of Regents. Although adverse impact on minorities was suggested from this evaluation, the Board of Regents initially elected not to support further investigation. Six months later, however, after minorities had leveled several criticisms of bias and adverse impact the Board requested further analyses.

Proposition 6

Schools and educational agencies honor court rulings that mitigate the potentially severe consequences of competency testing. Fair warning entered the classroom even earlier than it entered the workplace. With the exception of Granada, competency tests and standards were announced well before they were applied. Generally, tests and standards were adopted two to five years before students would be held accountable for them. Although the Granada Legislature mandated a competency test with little advanced warning, both the State Board of Education and Board of Regents took steps to cushion any harmful blows. Specifically, the Board of Education adopted a three-phase increase in GTTP-BS standards. Later, the Board of Regents permitted students to take the GTTP-BS while another test was sought and validated. By doing this, the Board of Regents could placate legislative concerns for speedy implementation and, like earlier efforts by the Board of Education, shield themselves from legal jeopardy.

After a general announcement was issued in the sites, additional and routine efforts were expended to notify

specific cohorts of examinees. Entering high school students and their parents in Lucerne and Victoria were apprised in writing of the impending tests and standards. Victoria demands that parents sign these notices which are filed in students' cumulative folders.

Examinees were advised about test performance beyond pass or fail. Students received feedback on performance by objectives or subtests. Further, they were told which subject areas needed remediation. Lucerne and Victoria provided remedial instruction, Lucerne in the form of competency classes and Victoria in various forms (e.g., tutoring and after school classes) depending on individual districts. Documentation of remediation was required by the Victoria SEA and was said to be enforced. Nevertheless, legal responsibilities rested with Victoria districts and not the SEA.

Proposition 7

Competency tests and standards function primarily as symbolic gestures rather than instrumental reforms. Minimum competency testing and standards serve symbolic ends. Five themes were identified that revealed the principally symbolic motives underlying the standards reform movement: the standard's image, the apparent redundancy with pre-existing standards, the coupling of test performance with decisions, the emphasis on initial fail rates, and the contrast between early and late phases of competency testing reforms.

The image of the standard held by the public was acknowledged as important in all of the sites. Informants in Lucerne, Waterford, and Granada (GTTP) indicated that a selected standard must appear rigorous, not too lax, and conform to prevailing notions:

> An 80% represented a B to the Board of Education.
>
> If a cut-score is lower than 60%, it is pretty low for a competency test.
>
> A 70% proficiency level is a widely used and popular ly recognized passing level in American education.

The standard's appearance was the focus of criticism in Brittany. The BSU Admissions Director criticized the NWET as being "Mickey Mouse" and embarrassingly low. Officials in Victoria and Granada (PPST) were also concerned about the image of the standard. Raw cut scores were generally not publicized. Instead, standard scores were concocted and announced. These standard scores would serve the purposes of deflecting attention from actual raw scores needed to pass and would project an aura of authority and scientific rigor.

The considerable redundancy with pre-existing standards is another source of symbolism. Lucerne and Victoria showed instances of such redundancy. In Lucerne, no one has been denied a high school diploma solely on the basis of failing the competency test. In Victoria, 99% of the regular educations students passed the VEHSG. No one knows how many among those who failed the tests statewide would have been prevented from graduating because of the lack of Carnegie credits, the traditional standard.

The coupling of test performance and actual decisions speaks of symbolic motives as well. Waterford did not have unequivocal information about test performance and promotion decisions. Similar data were unavailable at the state level in Brittany and Granada. The numbers of those who fail and who are not admitted may exist within local institutions but as a matter of routine or public information, they were not figured or available.

The fourth symbolic theme is the emphasis on initial rather than ultimate fail rates. For a variety of reasons, some obvious and some not, initial fail rates are much higher than cumulative rates. Much attention, public and internal, was focused on first-time fail rates. They were heralded as upholding rigorous standards or denigrated as biased against minorities. Whether the attention was laudatory or critical, it was certainly greater than what was received by ultimate passing rates or decisions resulting from test performance. When ultimate passing rates were discussed publicly, they were pointed to as evidence of improvement. Officials publicly attributed the higher passing rates to increased competence or effective remediation. Only privately, and less frequently, would informants acknowledge the role of statistical artifacts (e.g., testing effects, regression and unreliability) in accounting for higher cumulative passing rates.

Finally, symbolism is evident in the contrast between early and late phases of competency testing reforms. The focus on structure, detail, precision and control during the

initiation and implementation of competency testing programs blurred during later stages. Painstaking attention was given to test development, standard-setting, and general implementation. All of this attention was characterized by emphasis on the test as the end and not the means. Serious, focused attention has not been directed to the basic and more compelling questions of impact, utility and value of competency tests and standards. Instead of receiving careful examination, these issues got cursory and impressionistic treatment. Unarguably, these issues are complex and perhaps intractable. Further, they outpace our technical and narrow understanding of testing, and pursual of these issues would likely reveal our inadequacies in dealing with our own creation. However, the lack of attention to impact, utility and value only serves to underscore the symbolic value of competency tests and standards while concealing any instrumental value these reforms may or may not bring. There seems to be little to say about the instrumental value of competency tests and standards-- it simply has not been (nor will it likely be) examined in any of the five sites. Granted, some of the reforms have been in place for a short time and any evaluation of impact, utility and value would take a number of years to complete. However, plans for such activity were absent, or at best, went without mention; a curious, but telling, state of affairs given all the organizational fanfare that welcomed these reforms.

Implications

There are a number of implications for research, practice, and policy that may be drawn from this study. The competing decision-making theories provided a restrictive focus on the standard-setting process and strict adherence to the hypotheses would not have uncovered the more compelling issues. Reanalyses of these data or future studies of standard-setting would benefit from a less deterministic focus. Specifically, conceptual tools employed in the policy sciences could guide and inform studies of standard-setting without imposing an inflexible structure. For example, Lasswell (1956) identified seven categories within the decision-making process: intelligence, recommendation, prescription, invocation, application, appraisal, and termination. Basic to all decision-making processes, these categories or phases could be explored and elaborated within the context of standard-setting without sacrificing attention to unique or interesting phenomena.

As discussed in the final proposition, the data indicated that competency tests and standards function primarily as symbolic gestures rather than instrumental reforms. If sites currently engaged in competency testing and standard-setting wish to make claims of effectiveness, then consequences of such reforms must receive the attention and study received during their implementation. Agency attention must begin with tracking student performance and linking that to subsequent decision-making. Further study by agencies or

educational researchers should attend to changes to achievement, attrition, attitudinal changes, and the like. Of special note, efforts must be expended to understand the more subtle impact of opportunities denied minorities as a result of test performance and subsequent decisions. Until empirical data have been generated about these concerns, policymakers considering such expensive reforms are cautioned against their enactment. The lack of evidence demonstrating positive or instrumental benefits for students should slow the race to implement competency testing programs and the setting of standards. Indeed, we have yet to see if the symbolic nature of such reforms truly benefits schools.

REFERENCES

Allison, G. T. (1971). Essence of decision: Explaining the Cuban missile crisis. Boston: Little, Brown and Company.

Andrew, B. J. & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Beavers, J. L. (1983). A study of the correlation of selected data on high school transcripts, English qualifying exam scores, and subsequent freshman/developmental English grades at Wytheville Community College (Report No. 83-2). Wytheville Community College, VA. (ERIC Document Reproduction Service No. ED 231 488).

Berk, R. A. (Ed.). (1980). Criterion-referenced measurement. Baltimore, MD: Johns Hopkins University Press.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21, 147-152.

Bloom, B. S. (1976). Human characteristics and school learning. New York: McGraw Hill.

Bowman, H. L., Petry, J. R., Rakow, E. A., Bowyer, C. H., & Nothern, E. F. (1985). Validation of the NTE and recommended performance standards for certification in Tennessee. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Brickell, H. M. (1978). Seven key notes on minimal competency testing. Phi Delta Kappan, 59, 589-592.

Cangelosi, J. S. (1984). Another answer to the cut-off score question. Educational Measurement: Issues and Practice, 3(4), 23-25.

Chafin, A. E. (1983). Setting a standard for standard-setting. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Montreal, Quebec.

Chafin, A. E. & Lindheim, E. (1982) Standards for standard setting. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Collins, R. (1979). The credential society: an historical sociology of education and stratification. New York: Academic Press.

Crews, W. B. (1986). Master achievement of mildly handicapped students on Florida's Statewide Student Assessment Test, Part II (SSAT-2). (Doctoral dissertation, University of Florida) Dissertation Abstracts International, 47, 05-A.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.

Cross, L. H., Frary, R. B., Kelly, P. P., Small, R. C. & Impara, J. C. (1985). Establishing minimum standards for essays: Blind versus informed reviews. Journal of Educational Measurement, 22, 137-146.

Evan, M. C. (1985). Is state legislation a catalyst for change? Phi Delta Kappan, 66, 501-503.

Fillbrandt, J. R. & Merz, W. R. (1977). The assessment of competency in reading and mathematics using community based standards. Educational Research Quarterly, 2, 3-11.

Firestone, W. A. & Herriott, R. E. (1982). Multisite qualitative policy research: Some design and implementation issues. American Behavioral Scientist, 26, 63-88.

Garcia-Quintana, R. A. & Mappus, M. L. (1980). Using norm-referenced data to set standards for a minimum competency program in the state of South Carolina. Educational Evaluation and Policy Analysis, 2, 47-52.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Gray, M. C. (1981). The perceived effect of the C-1 mandate and CAPPS on curriculum and instruction in school districts in Indiana. (Doctoral dissertation, Indiana State University, 1980). Dissertation Abstracts International, 41, 11-A.

Greene, D. & David, J. L. (1984). A research design for generalizing from multiple case studies. Evaluation and Program Planning, 7, 73-85.

Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. Journal of Educational Measurement, 22, 262-269.

Hall, J., Griffin, H., Cronin, M., & Thompson, B. (1985). Factors related to competency test performance for high school learning disabled students. Educational Evaluation and Policy Analysis, 7, 151-160.

Halpin, G., Sigmon, G. & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43, 185-196.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.) Criterion-referenced measurement. Baltimore, MD: Johns Hopkins University Press.

Hare, J. D. (1984). Impact of minimum competency testing on the mathematics curriculum in North Carolina (Doctoral dissertation, Vanderbilt University, 1983). Dissertation Abstracts International, 45, 03-A.

Hayes, J. E. (1981). A comparison of minimum competency testing programs in five selected Illinois public school districts (Doctoral dissertation, Southern Illinois University, 1980). Dissertation Abstracts International, 41, 08-A.

Hector, J. H. (1984). Establishing cut-off scores for placement in community college developmental courses. Walters State Community College: Morristown, TN. (ERIC Document Reproduction Service No. ED 246 934).

Huynh, H. (1982). A Bayesian procedure for mastery decision based on multivariate normal test data. Psychometrika, 47, 309-319.

Huynh, H. & Casteel, J. (1985). A comparison of the minimax and Rasch approaches to set simultaneous passing scores for subtests. Journal of Educational Statistics, 10, 334-344.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-475.

Kapes, J. T. & Welch, F. G. (1985). Review of the scoring procedures for the occupational competency assessment program in Pennsylvania. Pennsylvania State University. (ERIC Document Reproduction Service No. ED 268 289).

Karabel, J. & Halsey, A. H. (Eds.) (1977). Power and ideology in education. New York: Oxford University Press.

Kerins, T. C. (1982). Factors underlying local education agency decisions to adopt or reject the use of a minimum competency test as a requirement for high school graudation. Unpublished doctoral dissertation. University of Illinois.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Lasswell, H. D. (1956). The Decision Process: Seven Categories of Functional Analysis. College Park, MD: Bureau of Governmental Research.

Lockwood, R. E. (1985). The Alabama high school graduation examination experience: Technical concerns. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Chicago, IL.

Lathrop, R. L. (1986). Practical strategies for dealing with unreliability in competency assessments. Journal of Educational Research, 79, 234-37.

Madaus, G. F. (1981). NIE clarification hearing: The negative team's case. Phi Delta Kappan, 63, 92-94.

Maher, H. P. and Thomas, H. B. (1982). The effects of the Florida State Student Assessment Test on secondary level vocational enrollments: A time series study. Journal of Vocational Education Research, 7, 49-68.

McKinney, J. D. (1983). Performance of handicapped students on the North Carolina Minimum Competency Test. Exceptional Children, 49, 547-550.

Merritt, R. & Coombs, F. (1977). Politics and educational reform. Comparative Education Review, 21, 247-273.

Miles, M. & Huberman, A.M. (1984). Qualitative data analysis: A sourcebook of new methods. Beverly Hills, CA: Sage Publications.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

National Commission on Excellence in Education. (1983). A nation at risk: the imperative for educational reform: a report to the nation and the Secretary of Education. Washington, D.C.: United States Department of Education.

New Jersey State Dept. of Higher Education. (1980). Establishing cut scores on the New Jersey College Basic Skills Placement Test. Trenton, NJ: Author.

Pearse, C. A., Agrella, R. F., & Powers, S. (1982). Identification and placement of college students in developmental education programs. Paper presented at the Annual Meeting of the California Educational Research Association, Sacramento, CA.

Peters, E. (1981). Basic skills improvement policy implementaton guide no. 3: Standards-setting manual Rev. ed. Educational Testing Service: Brookline, MA. (ERIC Document Reproduction Service No. ED 206 696).

Popham, W. J. (1981). The case for minimum competency testing. Phi Delta Kappan, 63, 89-91.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D. & Williams, P. L. (1985). Measurement-driven instruction: It's on the road. Phi Delta Kappan, 66, 628-634.

Popham, W. J. & Yalow, E. S. (1984). Standard-setting options for teacher competency tests. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, LA.

Rosner, F. C. (1982). Testing for teacher certification: State-level examples. Educational Measurement: Issues and Practice, 1(1), 21-25.

Seidel, J. (1983). The Ethnograph [computer program]. Littleton, CO: Qualis Research Associates.

Serow, R. C. (1984). Effects of minimum competency testing for minority students: A review of expectations and outcomes. Urban Review, 16, 67-75.

Serow, R. C. and Davies, J. J. (1982). Resources and outcomes of minimum competency testing as measures of equality of educational opportunity. American Educational Research Journal, 19, 529-539.

Serow, R. C., Davies, J. J., & Parramore, B. M. (1982). Performance gains in a competency test program. Educational Evaluation and Policy Analysis, 4, 535-542.

Smartschan, G. F. (1983). Rx for the credibility gap: District-wide final examinations. Clearing House, 56, 380-382.

Stanard, M. (1985). West Virginia's professional education performance assessments: Quilt or patches? Action in Teacher Education, 7, 31-40.

Tirozzi, G. N., Baron, J., Forgione, P., & Rindone, D. (1985). How testing is changing education in Connecticut. Educational Measurement: Issues and Practice, 4(2), 12-16.

Van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard-setting. Journal of Educational Measurement, 19, 295-308.

Vold, D. J. (1985). The roots of teacher testing in America. Educational Measurement: Issues and Practice, 4(3), 5-7.

Walstad, W. B. (1984). Analyzing minimal competency test performance. Journal of Educational Research, 77, 261-266.

Wise, A. E. (1978). Minimum competency testing: Another case of hyper-rationalization. Phi Delta Kappan, 59, 596-598.

Ziomek, R.L. and Wright, B.D. (1984). A procedure for estimating a criterion-reference standard to identify educationally deprived children for Title I services. Final report. Des Moines Public Schools, IA. (ERIC Document Reproduction Service No. ED 250 342).

APPENDIX

INTERVIEW PROTOCOL

Respondent ID _____ Interviewer_____
Date of interview_____ Time_____

Documents_____
_____
_____

Protocol Outline: Pages:
A. Purposes of the test and standards
B. Position and background of respondent
C. Other decision-makers' position & bkgrnd
D. Expectations of respondent & others
E. Process of setting standards
F. Result of the process: the standard
G. Consequences of the standards

A. Before we talk about the people and processes involved in setting the actual test standards, I'd like you to tell me about the testing program...

A.1. Describe the purpose of the <test> (as communicated to you by <policy, policymakers>).

a. Who is the test designed for?

b. What is the test intended to measure?

c. Is there a specific time when the students are to take the test (i.e. beginning, during, end of program)?

d. What are the test results used for?

A.2. (According to the <policy, policymakers>,) what is the passing score, or standard, intended to reflect (i.e. minimum vs maximum competence in.....)?

A.3. Is this standard a replacement or revision of an older one, or is it new?

[if former] What was wrong with the old one?
[if latter] Why was it felt that a new one was needed?

A.4. What led to the decision to use a test for this purpose(s)?

a. Who was involved?

b. When?

c. What reasons were expressed?

B. I'd like to ask you some questions about your current position and how it relates to your involvement in the standard-setting process...

B.1. What is the formal title of your position in <agency>?

B.2. Is your current position as <title> the same as the one you held when you were involved in determining standards for the <test>?

a. [if yes] How long have you been in this position?

[if no] What was your position during that time?

How long had you been in that position?

b. [both] Would you describe your educational background or training when you started your job at <agency>?

B.3. At the time, what responsibilities were associated with your position?

a. How was your position related to the hierarchical structure of <agency>? (probe for specifics: who person reported to, examples of other positions at comparable level in the structure).

B.4. Describe how you became involved in determining standards for the <test>.

a. When were you first introduced to the task? By whom?

b. Was it considered to be a part of your regular job duties or an added responsibility?

c. What were you responsible for in the standard-setting process? (probe for specific duties like actually setting the standards, overseeing the process, or evaluating recommendations)

d. What amount of discretion did you have in performing your duties? (probe: did respondent have to gain approval for actions; could respondent delegate responsiblities, sub-contract, and the like)

e. Were there any deadlines or constraints that you had to observe? (probe for time pressures, fiscal restrictions, stakeholders to be included in the process)

C. I'm interested in finding out about other people or decision-makers that were involved in the standard-setting process...

C.1. Who were the other participants in the standard-setting process? [if no other participants, skip to D]

a. How did they become involved in the task? (probe: did you bring them in or did they get involved through other means)

b. What responsibilities did they have in the task?

c. Describe briefly how their responsibilities were related to your own. (probe: were you responsible for their actions, did you report to them, or did you have comparable authority and discretion)

d. What do you know of their backgrounds as related to the job of determining standards?

D. I'd like you to reflect on the time right before the standard-setting process was to begin.

D.1. What did you think the testing program would accomplish?

a. weed out incompetent students,
b. motivate students to work harder,
c. improve the image of education?
d. other reasons?

(probe: Can you tell me more about that?)

D.2. With that end(s) in mind, what did you expect the standard-setting process to entail? Was it a matter of:

a. finding the point at which students could be considered competent
b. determining a score that students would be concerned about earning
c. deriving a score that would demonstrate <agency's> concern for raising standards
d. other expectations

(probe for specifics if "other")

D.3. Do you feel that other participants shared the same thoughts about the testing program?

[if yes] What indicated that to be true?
[if no] What led you to think that others had different opinions? (probe: who were these people, what were their perceived opinions, how were they expressed)

D.4. Do you think that other participants shared similar expectations for the standard-setting process?

[if yes] What indicated that to be true?
[if no] What led you to think that others had different intentions? (probe: who were these people, what were their perceived intentions, how were they expressed)

E. Now, I'd like you to describe the process by which the pass-fail standard was determined...

[if by a group of decision-makers, go to E.1.]
[if by an individual, go to E.2.]

E.1. Think back to when the group first began the task of determining standards for the <test>. I'd like you to describe how things got started.

a. How did the group decide to proceed?

b. What did they agree to do?

c. In setting the pass-fail score, do you think people were primarily aware of the desire to achieve a particular pass-rate, or were they thinking primarily about requisite skills and competencies?

[if former] Were there pass rates that were considered undesirable or unacceptable?

For what reasons?

[if latter] Did people speak of how understandings of skills and competencies would be translated into a test score?

Was this seen as a routine matter, one which could draw upon technical or personal knowledge?

d. Did individuals have specific ideas as to what the passing score(s) should be?

   What were the ideas?
   Who was associated with each?
   What reasons were given?
   What did you think of the various ideas?
   How were the ideas received by the group?

e. Did you have any recommendations?

   [if yes] What were they?

   What made the scores or standards appealing to you?

   How did others respond to your ideas?

   [if no] Why not? (you agreed with someone else's suggestion, you had no feeling one way or another)

f. Did anyone assume a leadership role? If so, who and how?

g. How would you characterize the group's interactions at this time? Would you say they were basically: cooperative vs conflicting, purposive vs chaotic, ...)?

   Can you give me some examples?

h. Describe your participation (facilitator, leader, ...) in the group.

i. What may account for the disparate notions among people as to what the standards should have been?

   What could have affected individuals' notions?

   Is it possible that they had something to gain by endorsing a particular standard rather than others?

E.2. What role did cut-off score techniques (e.g., Angoff, Nedelsky, Contrasting Groups) play in determining the pass-fail standard?

a. Were any of these methods considered for use?

[if no, go to E.4.]

[if yes] Which ones were considered?

Why wasn't <method> considered?

b. How were the methods evaluated for their suitability for the situation?

[if by a group] Did the whole group participate in the evaluation of the methods?

[if no] Which participants were involved?

Describe your role during this time.

[either group or individual]

How were the methods judged or evaluated?

What was concluded about each method's utility?

[if by a group] Were there any disagreements over the perceived utility of methods?

If so, between whom?

Over what?

How resolved?

c. Which method(s) was selected to determine the standards? For what reasons?

Was it considered to be the best method for the situation?

[if yes] In what ways?

[if no] Then why was it selected?

[if a group] Was there any individual or group that decidedly influenced the process?

[if yes] Who were they?

How did they influence the process?

How would you characterize your own influence on the process?

What kind of standard was the method(s) expected to produce?

[if group] Did everyone involved seem to agree on the method's utility and ultimate selection?

[if no] Where were sources of disagreement?

Who expressed dissatisfaction?

For what reasons?

How were dissensions handled?

[if yes] No one expressed dissatisfaction with the method?

What did you think of the selected method?

E.3. What happened as the method was implemented?

a. Who was involved?

b. How long did it take?

c. Were there any problems? Like what?

d. How were they resolved? Were changes made in the method?

[go to E.6.]

[from E.2:]

E.4. Why weren't cut-off score methods considered by the group?

a. For what reasons?

b. What did you think about not using these methods? Why?

c. Did individual participants disagree?
[if so] Who?
Why?
How were disagreements handled?

E.5. Since cut-off score methods were not considered, how did the group determine the standards for the <test>?

a. What scores or standards were considered by the group?

b. How did participants' notions of what the standards should be affect the process?

c. Describe how these were dealt with by the decision-making group.

Were there any scores that were considered unacceptable?

[if yes] Which ones?
For what reasons?
Was this unanimous?

[if no] Who dissented?
For what reasons?
How were problems dealt with?

[if no] All scores were considered reasonable and practical?

How did the group distinguish among the [acceptable] scores?

[from E.3. or E.5.]

E.6. Were any arrangements made for pilot studies, validity studies, and the like?

[if yes] What was done?
By whom?
When?
With what results?

[if no] For what reasons?

F. This brings us to the result of the standard-setting process.

[go to F.1. if cut-off method was used, else go to F.2.]

F.1. What did the <method> yield for a standard?

a. Was everybody satisfied with the score?

[if yes] No one expressed dissatisfaction with the standard?
Why do you think that was the case?

[if no] Who was dissatisfied?
For what reasons?
How were dissensions handled?
Was the standard changed?
Under what circumstances?

[go to F.3.]

F.2. As a result of this deliberation process, which score was chosen for the standard?

a. How did the group arrive at that decision?

b. Did everyone agree?

[if no] Who expressed dissension?
For what reasons?
How were dissensions dealt with?

[if yes] No one expressed dissatisfaction with the standard?
Why do you think that was?

c. Was there any individual or group who decidedly influenced the process?

[if yes] Who were they?
How did they influence the process? In what ways?

[if no] Do you mean that everyone shared equal influence on the resulting decision?
No one expressed concern that a particular person or group was calling the shots?

d. How would you characterize your own influence on the process?

F.3. What plans were made to deal with students who failed to meet the standard(s)?

a. Would students have opportunities for re-testing? Under what circumstances?

b. Were any plans made to provide remediation or tutoring?

What were the plans?
How were they to be implemented?
Who was responsible?
When was the program(s) to be implemented?
How were they to be funded?

F.4. Thinking back to right after the standard was determined, what did you think would happen when it was applied for the first time?

a. What did you expect for a passing rate?

b. Did you anticipate any other consequences?

c. What do think other people expected?

G. What can you tell me about the consequences that followed the standard-setting process...

G.1. What happened when policymakers and other people (e.g. teachers, media, parents) were informed about the standard(s)?

a. How did they learn about what the standard was?

b. How did policymakers react?

Were there different reactions?
By whom?
What were their reactions?
How were they expressed?

c. How were their reactions addressed?

Did any of the standard-setters deal with the concerns?
In what ways?

d. Were any changes made in the standards before they were applied for the first time?

G.2. What happened when the test was given the first time?

a. What percent of students failed?

b. Did pass rates vary by ethnicity?
In what ways?
For which groups?

c. What kind of reactions did people express?
decision-makers themselves?
other administrators
policymakers
teachers
parents
media
students
others

d. Was the cut-off score altered?

[if yes] How?
For what reasons?
[go to G.3.]

[if no, go to G.2.e]

e. If too <few, many> students had passed the test on its first administration, would the standard have been changed?

[if yes] How?
What percent would be too <few, many>
Why?
By whom?

[if no] So if 95% of the students had <failed, passed> the test, you don't think the standard would have been altered?

Would you explain why you think that?

G.3. What happened when the test was administered a second time?

a. How did the pass rates compare to the first administration of the test?

In terms of the entire student population?
In terms of ethnic groups?

b. How did people react to these results?
decision-makers
other administrators
policymakers
teachers
parents
media
students
others

c. Were any changes made in the standards at this time?

In what way?
For what reasons?

d. To what did people attribute changes in pass rates?

e. What do you think accounted for the changes?

Practice effects?
Regression?
Gains in achievement?
Successful remediation?

f. Why do you think that <competing explanation> did not account for the changes?

G.4. Were any changes made in <agency>?

a. [if no prior remediation plans] Were any compensatory programs initiated for students who failed?

b. Have drop-out rates changed? How?

c. What about curricular changes? Were there added requirements?

d. Were any changes made in <agency's> positions?

e. How would you say the image of <agency> has been affected by <test>? In what ways?

H. Personal reflections

H.1. Do you think this program is here to stay? For what reasons?

H.2. Do you anticipate any changes in the program or standards in the next year or so?

H.3. Is there anything that we haven't covered that you feel is important to understanding the situation here?

H.4. Is there anything that you would like to ask of me?