

1-1-1991

Obtaining norm-referenced scores from criterion-referenced tests : an analysis of estimation errors.

Charlene Gower Tucker
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

Tucker, Charlene Gower, "Obtaining norm-referenced scores from criterion-referenced tests : an analysis of estimation errors." (1991). *Doctoral Dissertations 1896 - February 2014*. 4817.
https://scholarworks.umass.edu/dissertations_1/4817

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.



312066013296890

OBTAINING NORM-REFERENCED SCORES FROM CRITERION-REFERENCED TESTS:
AN ANALYSIS OF ESTIMATION ERRORS

A Dissertation Presented

by

CHARLENE GOWER TUCKER

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 1991

School of Education

© Copyright by Charlene Gower Tucker 1991
All Rights Reserved

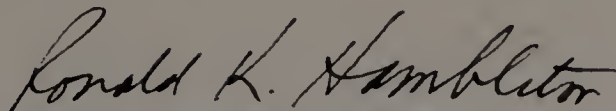
OBTAINING NORM-REFERENCED SCORES FROM CRITERION-REFERENCED TESTS:
AN ANALYSIS OF ESTIMATION ERRORS

A Dissertation Presented

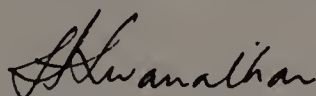
by

CHARLENE GOWER TUCKER

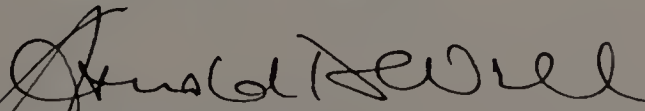
Approved as to style and content by:



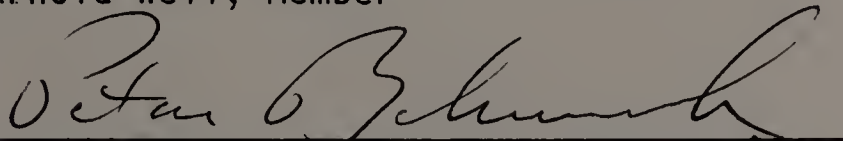
Ronald K. Hambleton, Chair



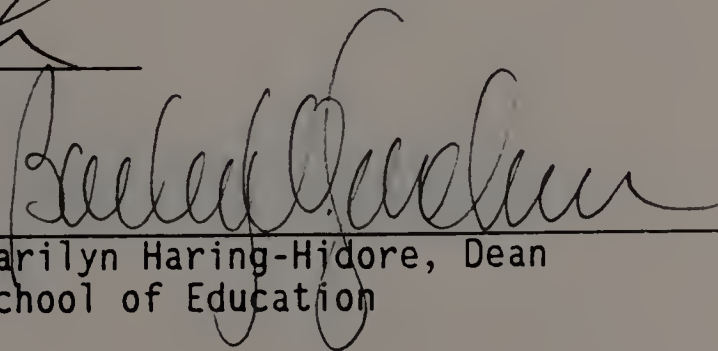
H. Swaminathan, Member



Arnold Well, Member



Peter Behuniak, Consulting Member



Marilyn Haring-Hidore, Dean
School of Education

To my mother, Lois Kimball Fournier

ACKNOWLEDGEMENTS

First, I must acknowledge Ronald Hambleton who served as my academic advisor, instructor, mentor, and as the chairperson of my dissertation committee. I worked closely with Ron throughout my six years in the graduate program. It was his expertise and wisdom that ensured that I had the necessary knowledge, experiences, and support to successfully complete the graduate program and to function as a productive member of the professional community.

I would also like to thank H. Swaminathan and all other faculty at the University of Massachusetts who offered me instruction and support in my graduate work.

The Connecticut State Department of Education provided the necessary data, computer facilities, and technical support for this research. In particular, the ongoing support of Peter Behuniak was critical and is greatly appreciated.

A special thank-you to my husband, Steve, who was my lifeline through the dissertation process, taking responsibility for my personal well-being and the maintenance of our home and family.

To my sons, Andrew and Benjamin, thank you for adjusting to Mom's need for separate time and space, for surviving and continuing to grow in my absence, and for welcoming me back into the family when it was over. I hope I can do the same for them someday when they need it.

To my secretary, Marge Brousseau, thanks for the many hours of expertise and patience which were devoted to the preparation of this document.

Finally, I want to extend gratitude to all the unmentioned friends, colleagues and family members who sustained me and my effort with their interest and encouragement.

ABSTRACT

MAKING NORM-REFERENCED INFERENCES FROM CRITERION-REFERENCED TESTS:
AN ANALYSIS OF ESTIMATION ERRORS

MAY 1991

CHARLENE GOWER TUCKER B.S., UNIVERSITY OF MAINE

M.Ed., UNIVERSITY OF MASSACHUSETTS

Ed.D., UNIVERSITY OF MASSACHUSETTS

Directed by: Professor Ronald K. Hambleton

One customized testing model equates a criterion-referenced test (CRT) to a norm-referenced test (NRT) so that performance on the CRT can produce an estimate of performance on the NRT. The error associated with these estimated norms is not well understood.

The purpose of this study was to examine the extent and nature of error present in these normative scores. In two subject areas and at three grade levels, actual NRT scores were compared to NRT scores which were estimated from a CRT. The estimation error was analyzed for individual scores and for group means at different parts of the score distribution.

For individuals, the mean absolute difference between the actual NRT scores and the estimated NRT scores was approximately five raw score points on a 60-item reading subtest and approximately two points on a 30-item mathematics subtest. A comparison of the standard errors of substitution showed that individual differences were similar whether a parallel form or a CRT estimate was substituted for the NRT score.

The bias present in the estimation of NRT scores from a CRT for groups of examinees is shown by the mean difference between the estimated and actual NRT scores. For all subtests, mean differences were less than one score point, indicating that group data can be accurately obtained through the use of this model.

To examine the accuracy of estimation at different parts of the score distribution, the data was divided into three score groups (low, middle, and high) and, subsequently, into deciles. After correcting for a regression effect, mean group differences between actual NRT scores and those estimated from a CRT were fairly consistent for groups at different parts of the distribution. Individual scores, however, were most accurate at the upper end of the score distribution with a decline in accuracy as the score level decreased.

In conclusion, this study offers evidence that NRT scores can be estimated from performance on a CRT with reasonable accuracy. However, generalizability of these results to other sets of tests or other populations is unknown. It is recommended that similar research be pursued under varying conditions.

TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGEMENTS	v
ABSTRACTvii
LIST OF TABLES	xi
Chapter	
1. INTRODUCTION	1
1.1 Background	1
1.2 Statement of the Problems	4
1.2.1 Need for Content Similarity	4
1.2.2 Overestimation	5
1.2.3 Time and Population Dependence	5
1.2.4 Variation Across the Score Distribution	5
1.2.5 Group Size	6
1.2.6 Customized Testing Model	6
1.3 Purposes	6
1.3.1 Research Area #1: Extent of Error/Bias for Groups and Individuals	7
1.3.2 Research Area #2: Variation in Error Across the Score Distribution	7
1.4 Educational Importance of the Study	7
1.5 Outline of the Dissertation	8
2. REVIEW OF LITERATURE	10
2.1 Introduction	10
2.2 An Historical Perspective	10
2.3 Current Models for Customized Testing	16
2.3.1 Shortened Version of NRT	18
2.3.2 NRT Intact with CRT Inferences	20
2.3.3 Locally Calibrated Items (Replacement or Addition)	22
2.3.4 Customized NRT from an Item Bank	25
2.3.5 CRT-Only	27
2.4 Summary	30
3. RESEARCH METHODOLOGY	32
3.1 Introduction	32
3.2 Description of the Sample	32
3.3 Instrumentation	34

Chapter	<u>Page</u>
3.3.1 Metropolitan Achievement Test, Sixth Edition (MAT6)	34
3.3.2 Connecticut Mastery Test (CMT)	36
3.3.3 CMT/MAT6 Link	39
3.4 Research Design	49
3.4.1 Research Area #1: Extent of Error/Bias for Groups and Individuals	51
3.4.2 Research Area #2: Variation in Error Across the Score Distribution	51
3.5 Data Collection/Editing	51
4. RESULTS	53
4.1 Introduction	53
4.2 Research Area #1: Extent of Error/Bias for Groups and Individuals	53
4.3 Research Area #2: Variation in Error Across the Score Distribution	60
5. CONCLUSIONS	67
5.1 Discussion of Results	67
5.1.1 Research Area #1: Extent of Error/Bias for Groups and Individuals	67
5.1.2 Research Area #2: Variation in Error Across the Score Distribution	71
5.1.3 Exceptional Dataset: Grade 6 Reading Comprehension	74
5.1.4 Exceptional Dataset: Grade 6 Mathematics Problem Solving	75
5.2 Implications of the Study	76
5.3 Limitations of the Study	78
5.4 Recommendations for Further Research	79
BIBLIOGRAPHY	81

LIST OF TABLES

Table		<u>Page</u>
3.1	Numbers of Subjects in Each Sample	34
3.2	KR-20 Reliability Coefficients for the MAT6 Survey Battery, Form L	37
3.3	Correlations of Alternate Forms of MAT6 Survey Battery, Forms L and M	37
3.4	Content Comparison of CMT and MAT6, Grade 4 Reading Comprehension	41
3.5	Content Comparison of CMT and MAT6, Grade 6 Reading Comprehension	42
3.6	Content Comparison of CMT and MAT6, Grade 8 Reading Comprehension	43
3.7	Content Comparison of CMT and MAT6, Grade 4 Mathematics: Problem Solving	44
3.8	Content Comparison of CMT and MAT6, Grade 6 Mathematics: Problem Solving	45
3.9	Content Comparison of CMT and MAT6, Grade 8 Mathematics: Problem Solving	46
3.10	Summary of Score Distribution Indicators Connecticut Mastery Test (CMT), Form C, Metropolitan Achievement Test, Sixth Edition (MAT6), Form L	47
3.11	Correlations Between MAT6 and CMT Subtests	48
4.1	Differences Between MAT6 Raw Scores as Estimated from CMT Performance and Actual MAT6 Raw Scores .	54
4.2	Comparison of Standard Error, Using the Same Form, Parallel Forms, and CRT Estimate, Reading Comprehension	56
4.3	Comparison of Standard Error, Using the Same Form, Parallel Forms, and CRT Estimate, Mathematics: Problem Solving	56
4.4	Distribution of Differences in Raw Scores on Two Test Administrations, 68% Confidence Intervals . .	59

Table	<u>Page</u>
4.5	Differences Between Estimated and Actual MAT6 Raw Scores By Level of MAT6 Score (X) With and Without Correction for Regression Effect, Reading Comprehension 62
4.6	Differences Between Estimated and Actual MAT6 Raw Scores By Level of MAT6 Score (X) With and Without Correction for Regression Effect, Mathematics: Problem Solving 63
4.7	Differences Between Estimated and Actual MAT6 Raw Scores By National Percentile Rank (P) of MAT6 Scores With and Without Correction for Regression Effect, Reading Comprehension, Grade 4 64
4.8	Differences Between Estimated and Actual MAT6 Raw Scores By National Percentile Rank (P) of MAT6 Scores With and Without Correction for Regression Effect, Reading Comprehension, Grade 6 64
4.9	Differences Between Estimated and Actual MAT6 Raw Scores By National Percentile Rank (P) of MAT6 Scores With and Without Correction for Regression Effect, Reading Comprehension, Grade 8 65
4.10	Differences Between Estimated and Actual MAT6 Raw Scores By National Percentile Rank (P) of MAT6 Scores With and Without Correction for Regression Effect, Mathematics: Problem Solving, Grade 4 65
4.11	Differences Between Estimated and Actual MAT6 Raw Scores By National Percentile Rank (P) of MAT6 Scores With and Without Correction for Regression Effect, Mathematics: Problem Solving, Grade 6 66
4.12	Differences Between Estimated and Actual MAT6 Raw Scores By National Percentile Rank (P) of MAT6 Scores With and Without Correction for Regression Effect, Mathematics: Problem Solving, Grade 8 66
5.1	Relationship Between Correlation Coefficients and Standard Error of Substitution 68

Table		<u>Page</u>
5.2	Mean p-Values and Skewness Indicators for Tests Used in Connecticut and Tests Used by Schattgen and Osterlind (1989)	73
5.3	Mean Difference Between Estimated MAT6 Scores and Actual MAT6 Scores ($\hat{X} - X$), Raw Scores and Normal Curve Equivalents (NCE)	74

CHAPTER 1

INTRODUCTION

1.1 Background

Standardized norm-referenced achievement tests are designed to assess an examinee's level of competence on a set of general goals which represent the basic school curricula across the nation. Examinee performance is interpreted through comparison with the performance of a representative national sample.

Norm-referenced tests (NRTs) have been criticized because they do not closely match the curricula of the school districts where they are used (Good & Salvia, 1988; Jolly & Gramenz, 1984; Schmidt, 1983; Wilson & Hiscox, 1984). This match becomes particularly important when the test is used, not only for comparing general student performance to a national group, but also to assess the success of students or programs in relation to a specified set of objectives. Among the implications of this mismatch between the curriculum and the test content are that (1) school districts do not receive information on all the content areas of interest, and (2) administrator and teacher attitudes toward NRTs are often negative because the tests are perceived as lacking educational relevance and possibly as being unfair (see, for example, Jolly & Gramenz, 1984).

Criterion-referenced tests (CRTs), on the other hand, are designed to assess an examinee's level of competence on a set of objectives which are clearly specified for a given state or local curriculum. The examinee's performance is interpreted as the degree to which the specified content domain has been mastered.

CRTs clearly address the issue of match between test content and the curriculum, but they have their own limitations. CRTs are very costly to develop and local school districts often do not have the resources to assure a degree of test quality which can match the standards of a nationally developed NRT. But, perhaps, the most serious limitation of a CRT program is the absence of an independent criterion against which to compare the performance of a particular group or program.

Since both NRTs and CRTs offer a limited testing program, and the implementation of both NRT and CRT programs is often prohibited by factors of time and expense, efforts are being made to create a single test which serves both purposes. Customized tests, tests which can simultaneously provide information regarding an examinee's mastery of particular content, a criterion-referenced inference, and the examinee's standing in relation to the national population, a norm-referenced inference, are being sought.

This psychometric feat is being approached in a number of ways. Keene and Holmes (1987) described four categories of customized testing models which are being explored:

1. NRT-Only Model. Both NRT and CRT inferences are made from the administration of an NRT. CRT inferences may be based on only those items which are appropriate for a given curriculum.
2. NRT-Based Model. The content of an NRT is modified to facilitate better CRT inferences. Items may be added, removed, or replaced.
3. CRT-Based Model. A criterion-referenced test is modified, usually through the addition of some NRT items.

Either the NRT items alone or a combination of NRT and CRT items are used to estimate norms.

4. CRT-Only Model.¹ A CRT alone is used to provide both CRT and NRT information. Normative inferences are made possible through the equating of the CRT to a nationally-normed test.

Within these four models, there is a great deal of variability in the approaches which are being utilized. Each approach is complete with its own set of outstanding questions and concerns.

This study examined one example of a CRT-only model, that which is currently being used in the state of Connecticut. The Connecticut State Department of Education (CSDE) administers a statewide CRT, the Connecticut Mastery Test (CMT), to all public school students in Grades 4, 6, and 8 each fall. The CMT does not provide normative interpretations of examinee performance, but rather determines the mastery status of examinees on clearly specified educational objectives.

Many of Connecticut's students are involved in special programs (e.g., federal ESEA Chapter 1 compensatory education program) which require NRT data for program evaluation. For reasons of educational relevance and test economy, there is a great deal of interest in a testing design which allows the CMT to produce the normative data necessary for this evaluation.

¹ Keene and Holmes (1987) use the term Objective-Referenced Test (ORT); Linn and Hambleton (1990) use the term Curriculum-Specific Test (CST); this paper uses the term Criterion-Referenced Test (CRT). These terms can be used interchangeably.

In order to obtain norm-referenced information from the CMT, a large-scale equating study was carried out. The Connecticut Mastery Test was equated to the Metropolitan Achievement Test, Sixth Edition (MAT6). This equating study provides the mechanism by which MAT6 scores can be estimated from CMT scores; since MAT6 is norm-referenced, estimated norms can then be reported. This design allows a single test, the CMT, to provide both criterion-referenced information and estimated norm-referenced information.

1.2 Statement of the Problems

Any model which uses the norms from a nationally normed NRT to represent performance on a test other than the original NRT raises concern about the accuracy of the estimated (equated) norms. This is true whether a customized version of the original NRT is being used or, as in Connecticut's case, a different test altogether is being administered.

The extent and the nature of error associated with these estimated norms is not well understood at this time. Neither is there an understanding of the effect that the various approaches to test customization and test equating have on the accuracy of the estimated norms. The current literature in this area raises some specific concerns.

1.2.1 Need for Content Similarity

The need for the content of the local test to be similar to that of the normed test is essential according to Yen, Green and Burket (1987); in fact, it is essential to the definition of equated tests.

A study by Way, Forsyth, and Ansley (1989) shows that NRTs which were shortened to focus on the strengths of particular schools yielded higher ability estimates than did the full test. If one's goal is to create a test that is better aligned with a local curriculum than the nationally normed test, the issue of content similarity is a serious consideration.

1.2.2 Overestimation

This concern is that error associated with the estimation of norms on a local test may be systematic error rather than random error. Higher norms may be estimated based on the local test than would actually be achieved if the original NRT had been administered. This may occur if local instruction has a greater impact on examinees' performance on the local test than on the original NRT. A gain on the local test may estimate a larger gain on the NRT than would be obtained in practice.

1.2.3 Time and Population Dependence

If two tests are equated based on the score distributions of a particular group at a particular time, there is question regarding the equivalence of the two tests for other populations or at other times (Yen, Green and Burket, 1987).

1.2.4 Variation Across the Score Distribution

A study by Schattgen and Osterlind (1989) indicates that the accuracy of estimated norms may vary across different parts of the score distribution. In a design where a local CRT was equated to an NRT, they found higher agreement between scores on the two tests at

the lower end of the score distribution than at the upper end of the distribution.

1.2.5 Group Size

There is discussion in the literature regarding the minimal group size for assuring confidence in the estimated NRT scores. Since Chapter 1 evaluation requires norms for individual students, there is great interest in using a customized testing design to produce estimated NRT scores at an individual level. The accuracy of such a design is yet to be determined.

1.2.6 Customized Testing Model

There are many methods being used to create customized tests. Of the models in which two tests are equated, many different equating methods are being used. The relative effectiveness of the different methods, their advantages and disadvantages are not well understood.

If the education community is to continue its pursuit of an all-purpose (CRT and NRT) test, research is needed into these outstanding issues. The integrity of the various models which are being employed needs to be examined in a practical, as well as theoretical, context.

1.3 Purposes

In Connecticut's model, performance on the statewide criterion-referenced test (CRT), the Connecticut Mastery Test (CMT), is used to estimate performance on the norm-referenced Metropolitan Achievement Test, Sixth Edition (MAT6). The purpose of this study was to examine the extent and nature of any error present in Connecticut's

estimated norms. By comparing actual MAT6 scores to MAT6 scores which were estimated from CMT performance, the following research questions were addressed:

1.3.1 Research Area #1: Extent of Error/Bias for Groups and Individuals

- a. What is the extent of the error present in individual MAT6 scores as estimated from CMT performance?
- b. What is the extent and direction of the error present in group MAT6 scores as estimated from CMT performance?

1.3.2 Research Area #2: Variation in Error Across the Score Distribution

- a. Are there differences in the degree of error present in individual MAT6 scores as estimated from CMT performance among three ability groups: low, middle, high?
- b. Are there differences in the degree or direction of error present in group MAT6 scores as estimated from CMT performance among three ability groups: low, middle, high?
- c. What error patterns can be observed across the score distribution?

1.4 Educational Importance of the Study

Given the current emphasis on accountability in education, criterion-referenced tests (CRTs) are playing an increasingly significant role in assessment. However, they have not eliminated the need for norm-referenced information. The education community is seeking an efficient and coherent way to meet these multiple testing needs.

The challenge of designing a customized testing program which can meet both criterion-referenced and norm-referenced testing needs from one test is being approached from many angles, often without the wisdom of experience or the support of research. The literature which does exist in this area is somewhat contradictory, varying in degrees of enthusiasm and skepticism. Furthermore, existing research tends to be highly theoretical with little focus on the integrity of the various testing designs in actual practice.

This study has examined one model which is currently in place in Connecticut. Some of the concerns which have been raised in the literature were confronted head-on as they showed themselves in practice. The results of this study will provide needed guidance to the state of Connecticut and other pioneers in this area as they further explore and refine these testing methodologies. If this study shows that Connecticut's design is working accurately, it will provide a model for the national education community. If problems are revealed, Connecticut and the rest of the nation will be in a position to proceed more wisely.

1.5 Outline of the Dissertation

There are four additional chapters in this paper. Chapter 2 presents a two phase literature review. First, historical background information is presented; next, a summary of different models for creating a customized test to meet both NRT and CRT needs is presented. Chapter 3 presents the methods that were used to investigate Connecticut's model in terms of the research questions stated in Section 1.3. The results of the study are summarized in

Chapter 4; and, finally, the results of the study are discussed in Chapter 5 in terms of conclusions and future implications.

C H A P T E R 2

REVIEW OF LITERATURE

2.1 Introduction

In order to gain direction for the exploration of the testing model being used in Connecticut, and in order that knowledge gained from this study may be appropriately interpreted in relation to the more general field of educational testing, an extensive literature review was conducted.

The literature which was reviewed is presented in the remaining sections of this chapter. First, in Section 2.2, an historical perspective is offered beginning with norm-referenced testing, progressing to the introduction of criterion-referenced testing, and on to the concept of a customized dual-purpose test. Secondly, in Section 2.3, attention is focused on the various models which are used to create customized tests. Finally, in Section 2.4, the information which emerged from the literature review as most relevant to this study is summarized.

2.2 An Historical Perspective

It seems useful and appropriate to understand any new idea in terms of both its current context and its place in history. The idea of getting both criterion-referenced and norm-referenced information from one test is a relatively new development, but one which has a logical place in the history of educational testing.

A norm-referenced test (NRT) as defined by Yen, Green, and Burket (1987) is "a test for which national norms have been obtained by

administering that test to a representative national sample of examinees and producing score distributions" (pp. 7-8). Standardized achievement tests are NRTs designed to measure a set of general goals which represent the basic school curricula across the nation. An extensive analysis of the various curricula and textbooks being used, as well as input from curriculum content experts, provide the data from which a common core of general goals is identified (Diamond, 1984; Mehrens & Phillips, 1986).

The test items which measure these general goals must also satisfy several psychometric criteria if they are to function properly as NRT items. Difficulty and discrimination indices play an important role; items of moderate difficulty and high discrimination are preferred since they make the greatest contribution to test score variance and, ultimately, test score reliability and validity (Hambleton, 1985). Well-behaved NRT items also show continuous growth from grade to grade; despite an inclination to include less well-behaved items for reasons of content coverage, it is understood that, in order to maximize the accuracy of the derived scores, NRT items should exhibit monotonic growth patterns (Green & Yen, 1984; Diamond, 1984).

Norm-referenced achievement tests were used to measure student achievement for more than 40 years with generally successful results. Through the 1970s and 1980s, however, the growing concern for the quality of the nation's educational system, the increased state involvement in education, and the introduction of federally mandated program evaluation have all increased the use and, unfortunately, the misuse, too, of NRT information (Keene & Holmes, 1987; Schmidt, 1983; Jolly & Gramenz, 1984). The range of applications and interpretations

of the scores derived from NRTs has been expanded to include the assessment of curricula and programs, the assessment of teachers, and also to look diagnostically at student performance in relation to a set of desired competencies (Wilson & Hiscox, 1984; Good & Salvia, 1988; Schmidt, 1983; Jolly & Gramenz, 1984).

As the importance placed on NRT scores increased, and as the range of inferences made from the test scores widened, attention was directed to the content of the tests (Goldsby, 1988; Mehrens, 1984). Is it fair to assess a student, teacher, or curriculum on the basis of a test which appears to measure different content from the content which was taught?

Norm-referenced achievement tests are designed to measure student performance on a common set of general educational goals, and to compare that student's performance to a representative sample of students across the nation. It is not likely that any district's curriculum is perfectly matched to the content of a particular norm-referenced achievement test, or that any NRT is perfectly matched to a particular district's curriculum. There is always content tested but not taught and content taught but not tested (Mehrens, 1984; Kean, 1986; Good & Salvia, 1988). Furthermore, the degree of match between test content and content taught is, in general, different for each test-curriculum combination. This differential test-curriculum match has been shown to have an effect on test scores; students systematically achieve higher NRT scores on tests which exhibit a stronger match with their curriculum (Good & Salvia, 1988; Yen, Green, & Burket, 1987).

With the onset of objective-based instructional programs in the late 1960s and the minimum competency movement of the mid-1970s, the measurement of student performance on a specific set of objectives or competencies became important (Popham, 1978). With interest centered on the assessment of competencies, the match between test content and content taught became critical, and NRT characteristics such as differentiation among students and monotonic growth curves for items became secondary. The stage was set in the late 1960s for a new wave of testing: criterion-referenced testing (see, for example, Jaeger & Tittle, 1980).

Criterion-referenced tests (CRTs) are designed to measure a very specific set of objectives or competencies. CRTs can be developed to assess student achievement in relation to a state or local curriculum, and/or to evaluate the effectiveness of an instructional program in meeting its particular goals. CRTs are particularly useful for diagnosing student, or program, strengths and weaknesses (Hambleton 1985; Popham, 1978).

Of prime importance in developing a CRT is the definition of the specific domain of content. The level of content specificity required for the development of a CRT is much more detailed than for an NRT (Popham, 1978). Item statistics are less important in CRT development than they are in NRT development. The critical characteristic of CRT items is their adherence to content specifications (Hambleton, 1985).

CRT results can be reported as a description of examinee performance or as a classification of the examinee as a master or nonmaster of a particular competency (Hambleton, 1985). The test results can easily be interpreted and applied in the context of a

program/curriculum. Test results are more readily accepted by school personnel due to their obvious validity for their purposes. The direct relationship between the test content and the curriculum can encourage instruction to the desired curriculum which, in turn, can influence test results.

CRTs do not fill all achievement testing needs, however. There is often a need to compare a particular school, program, or child to a national norm group. There is still a need for external criteria for judging curricular effectiveness; a program must not only meet its specified goals, but also maintain a favorable standing in relation to other instructional programs. These are the functions of norm-referenced testing.

Throughout the 1970s, debate abounded between proponents of norm-referenced testing and proponents of criterion-referenced testing. NRT advocates argued that their tests could provide information on the mastery of objectives, as well as normative data. CRT advocates argued that the CRT data, adequately provided only by CRTs, was preferred to the data provided by NRTs (Hambleton, 1985). Since that time, an understanding seems to have been reached that CRTs and NRTs are two different types of tests with different characteristics. CRTs are valid for some purposes, and NRTs are valid for other purposes (Mehrens, 1984; Hambleton & Rogers, 1989).

Since the education community generally recognizes the value of both criterion-referenced and norm-referenced measurement, and the limitation of resources allocated for testing (e.g., money, time) often prohibits the coexistence of two separate testing programs, experimentation is taking place with customized testing programs which

can provide information specific to a given curriculum along with national norms. Is it possible to create a single test with the content coverage necessary for valid criterion-referenced measurement and the psychometric properties necessary for valid norm-referenced measurement?

New possibilities have been created by the development of item response theory (IRT) in the 1970s and 1980s (Hambleton & Swaminathan, 1985). In item response theory, examinees are considered to have a particular quantity of ability on a latent trait. A mathematical relationship is established for each test item between examinee ability level on the latent trait and the probability that the item will be correctly answered. From an examinee's performance on a set of calibrated test items, that examinee's ability on the latent trait can be estimated.

IRT has definite advantages over traditional test theory models for the customization of tests. Since an examinee's ability is estimated from information provided by individual test items rather than by a test in its entirety, there is room for more flexibility in terms of the items which compose the test. In an IRT model, an examinee's ability can be estimated regardless of the subset of items to which the examinee responds. Furthermore, when equating two tests using traditional methods there is concern that the tests be of similar difficulty and that examinee groups be similar; using an IRT model, issues of group similarity and test difficulty are less critical. As long as the underlying assumptions of the IRT model are met, an examinee's ability estimate will be the same, apart from

measurement error, regardless of the choice of items in the test. (Cook & Eignor, 1983; Cook & Eignor, 1989).

For purposes of this study, some aspects of the IRT model must be considered. One assumption of relevance is the assumption of unidimensionality. The assumption of unidimensionality requires that only one underlying trait or ability accounts for performance on the test. Although this requirement cannot be strictly met, it is expected that one trait be clearly a "dominant" factor (Hambleton & Swaminathan, 1985).

Another aspect of IRT which should be considered is the range of available IRT models. A commonly used model, the Rasch model, is a one-parameter logistic model which uses only a difficulty parameter to determine the item response functions. Other models have additional item parameters. In a two-parameter model, an item's discriminatory power is also considered. In a three-parameter model, a third characteristic related to the influence of guessing is considered. The choice of IRT model can be based on resources, preference, and/or the degree to which various models "fit" the particular set of data (Hambleton & Murray, 1983).

In the next section, Section 2.3, several models for creating a customized test are described. Some of the models are possible due to the advent of item response theory.

2.3 Current Models for Customized Testing

There are currently a wide range of models which the education community is using to derive both criterion-referenced and norm-referenced information from the same test. In an important

review by Keene & Holmes (1987), models were described which ranged from using only an NRT to using only a CRT along with other models striking some sort of compromise between the two.

There is currently no model which offers both ideal criterion-referenced inferences and ideal norm-referenced inferences. One dimension tends to be compromised for the other. Careful study of the various models is needed to inform those who are searching for that appropriate balance.

In reviewing the relevant literature, this author found five different models for creating a customized dual-purpose test:

1. Shortened Version of NRT;
2. NRT Intact with CRT Inferences;
3. Locally Calibrated Items (Replacement or Addition);
4. Customized NRT from an Item Bank; and
5. CRT only (Equated to NRT).

The first three models offer methods of modifying the content of an NRT so that it can provide a closer content match and, therefore, enhance the resulting criterion-referenced inferences. The fourth model uses the concept of an item bank from which a customized test can be created which provides both NRT and CRT information. The fifth model uses a CRT which has been equated to an NRT, providing CRT information and estimated (equated) NRT information.

Each of these models has its own set of advantages and disadvantages. In the remainder of Section 2.3, each of these five models will be described in terms of its procedures and outstanding concerns.

2.3.1 Shortened Version of NRT

One method for customizing an NRT so that it better matches a curriculum of interest is to remove those items which represent content not included in the local curriculum from the norm calculations. These recalculated norms have been termed "curriculum-referenced norms." This type of test customization deals with the issue of content tested but not taught. However, it does not address the issue of content which is taught but not tested (Keene & Holmes, 1987). Three studies described below look at the effect of recalculating norms after removing test items from a norm-referenced achievement test on the basis of content.

In the first study, Allen, Ansley, and Forsyth (1987) created three shortened versions of the Quantitative Thinking Subtest of the Iowa Tests of Basic Skills. Each shortened version was customized along different content lines. Sample schools were selected for analysis whose percent correct scores were higher for the content selected for the customized test than their scores on the content which was eliminated. This sample selection criteria simulates the realistic setting where school personnel would select items which correspond to their curricular emphases. For most of the schools in the study, the customized tests overestimated abilities as compared to the full test. In conclusion, the researchers recommended caution in using a shortened version to predict performance on a full-length, standardized achievement test.

In the second study, Way, Forsyth, and Ansley (1989) created two shortened versions of four subtests from the sixth grade Iowa Tests of Basic Skills. One version was representative customized (RC); that

is, the content of the RC version was representative of the content on the normed test. The other version was content-customized (CC); that is, clusters of content objectives were selected for inclusion, as a school district concerned with curricular match might have done.

Two stages of analysis were carried out. First, the customized tests were compared with the full test in the national standardization sample. Secondly, the three tests (full, RC, and CC) were compared for four schools, which were selected because they performed better on the CC version than on the full test.

There was no evidence to show that different abilities were estimated in the national sample among the three tests: full, RC, and CC. However, in the selected schools, three of four CC subtests yielded higher ability estimates than the full test. For unknown reasons, the two RC subtests yielded ability estimates which were lower than those derived from the full test. The authors concluded as follows:

...for certain populations, scores on customized versions of standardized achievement tests cannot be expected to be equivalent to scores based on the full-length test (p. 35).

In the third study by Harris (1987), customized versions of a 40-item mathematics test were created by omitting selected subtests, resulting in tests with differential content. IRT ability estimates were then derived for examinees based on the total test and based on the customized tests. Clear differences were found in the ability estimates when a customized version was used.

This model, shortening an NRT to exclude content which is irrelevant to the local curriculum, perhaps, enhances the face validity of the tests. However, there seems to be reason for concern

over the integrity of those recalculated norms, particularly where content is disproportionately affected by the customization. The overestimation of norms which seems to occur with this type of test customization may be the desired outcome of some school districts. They may perceive this procedure as correcting a previously unfair arrangement where their students were tested on content that was not taught, thus correcting for previously underestimated achievement.

This situation can best be understood by considering the norming sample. The normed test was not perfectly matched to the curriculum of those students either; there is likely to have been some content tested but not taught for many of those examinees. If a school district chooses its NRT partly on the basis of content match, an advantage is already present over that of the norming sample. With additional customization, perhaps the appropriate interpretation of a fiftieth percentile would be that an examinee performed better on content that was taught in his/her school district than 50% of the students in the norming sample who, on the average, did not receive instruction in as large a proportion of the test content. It just seems to be an unfair comparison which should be considered with great caution.

2.3.2 NRT Intact with CRT Inferences

This second method for customizing an NRT keeps the NRT and its norms intact, as they were designed to be used. In addition, CRT score interpretations are made, sometimes exclusively on the basis of NRT items, and sometimes on the basis of both NRT items and

supplemental items which address the content taught but not measured in the NRT (Keene & Holmes, 1987).

Example 1: Wilson and Hiscox (1984) administered a complete norm-referenced achievement test, and used its associated norms as provided by the publisher. Then, they reanalyzed only the items that matched their learning objectives and reported percent correct criterion-referenced scores for those objectives which were adequately assessed in the NRT. The validity of the norms was not threatened, and some additional information, however limited, was gathered in relation to their learning objectives.

Example 2: Jolly and Gramenz (1984), of Palm Beach County School System, developed a system which used a combination of NRT items and supplementary items for criterion-referenced assessment of their local objectives. Again, the NRT was used in its entirety for norm-referenced measurement. Administered in conjunction with the NRT were the supplementary items necessary for assessment of the local objectives not adequately assessed by the NRT. Each local objective was measured on the basis of four items; those items may have been exclusively NRT items, exclusively supplementary items, or some combination.

In this approach, the face validity of the test was enhanced, criterion-referenced data were reported on all local objectives, and the integrity of normative inferences was maintained. The expense of this comprehensive testing program was in the development of supplementary test items and increased testing time.

2.3.3 Locally Calibrated Items (Replacement or Addition)

This model for customizing an NRT deals with the issue of content which is part of a local curriculum but is not on the normed test. In this model, test items designed to measure local objectives are locally developed and calibrated. These items can be used in addition to or in place of some NRT items, and, once calibrated, they are used to contribute to the NRT scores.

It is possible, through IRT, for both the original norm-referenced items and the locally developed items to contribute to the NRT scores. The local calibration of the new test items requires the use of an NRT (or some part) as an anchor. Once a sample of examinees has taken both sets of items, local performance on the locally developed items can be meaningfully compared to local performance on the nationally normed items. This allows both sets of items to be placed on a common IRT scale. Then, an ability estimate derived from the customized test can be used to estimate performance on the original NRT.

Example: New York City recently developed a customized version of the Metropolitan Achievement Test (MAT). Some items which tested content not taught in the New York City Schools (or were viewed as unimportant at the grade levels where they were assessed) were deleted from the MAT, and new items were developed to measure the content taught but not measured in the tests. Great care was taken to control both the content and the psychometric properties of items. Old items were replaced by new items exhibiting the same difficulty and discrimination indices but better matched to the New York City curriculum. Using the original MAT as an anchor, the new items were

calibrated, and the customized test was shown to be psychometrically equivalent to the original NRT. The new test yields criterion-referenced information, has necessary face and content validity, and produces the same score distribution as the original NRT (Taleporos, Canner, Strum, & Faulkner, 1988).

This model places the locally-developed items on the same IRT scale with the nationally normed test items. If one could assume that these items all measure one predominant latent trait, item response theory would assure comparability between ability estimates from the customized test and ability estimates from the original norm-referenced test. However, since achievement tests are generally not unidimensional, the two tests must be matched for their multidimensionality (Yen, Green & Burket, 1987). This means that there should be a close content match between the two tests. If a close content match cannot be established, as may be the case given the purposeful content changes in New York's design, there is reason for caution. The concern is that local instruction may have a greater impact on performance on the locally-developed items than on the NRT items, threatening the validity of norm-referenced interpretations. A hypothetical situation created by Yen, Green, and Burket (1978) illustrates this threat:

A school district created a customized test consisting of items written locally to reflect the special goals of a new instructional program. At the beginning of the new program these items were locally calibrated to a scale defined by a nationally normed test that contained a broader sampling of content than was in the customized test. When the customized test was given again near the end of the program, all but two schools showed gains of 15 to 20 points on the national percentile norms. Investigation showed that those two schools had not really implemented the new program. The principals in these schools insisted that nevertheless they had done a good job in teaching that subject and

asked that the full standardized test be given to students in all the schools. This was done and these two schools showed about as much growth as the other schools. Clearly the customized test reflected student learning of the materials in the new program but overestimated growth on the nationally defined scale. In other words, growth on the special local material did not lead to corresponding growth on the more broadly defined national scale (p. 12).

Through the equating process a relationship can be established between the original NRT and the customized test for the local population at the time of the equating. In predicting local performance on the original NRT from subsequent administrations of the customized test, it must be assumed that the relationship established in the equating process is stable over time, instruction, and local population variance.

Three studies have examined the effect of adding locally calibrated test items to a norm-referenced test. In a study by Dungan (1988), hypothetical customized versions of the MAT6 mathematics test were created which were purposely more difficult than the shelf test. Shortened versions were created by deleting the twenty easiest test items; then, customized versions of the original length were created by adding twenty new test items to the shortened versions. Using item response theory, the customized tests were linked to the MAT6 scale. Dungan found that the differences in MAT6 scaled scores for groups of examinees between the shelf test and the customized tests were not substantial. Individual score differences were more significant. Substituting new items had a greater effect than just removing items. Furthermore, a strong relationship was detected between the degree of change in scaled scores and the change in the difficulty of the tests.

In another study, Green (1987) looked specifically at the effect of time on the relationship between the nationally and the locally calibrated test items. In Philadelphia, local test items were calibrated and added to a national test in 1984. The test was used again in 1985 and in 1986. Performance on the nationally calibrated items was compared to performance on the local items at the three points in time to determine whether the local curriculum had more impact over time on the locally calibrated items. He found some effect of the local curriculum, but it was fairly small.

A study by Qualls-Payne, Raju, and Groth (1989) looked at the accuracy of estimated national p-values for locally calibrated test items in a model which uses a core set of nationally normed test items to calibrate the local items. The effect on the accuracy of the national p-values of three variables was investigated: the number of items in the core set, the IRT model used for the calibration, and the calibration sample size. Core sets were chosen to be of comparable content and difficulty to the national test. Qualls-Payne, Raju, and Groth found that national p-values were quite accurately estimated for local items. The length of the core set had little effect on the accuracy. Increasing the calibration sample size did strengthen the estimation. Also, the one-parameter Rasch model produced more accurate estimates of p-values than the three-parameter IRT model.

2.3.4 Customized NRT from an Item Bank

This approach for creating a customized norm-referenced achievement test requires a large pool of nationally calibrated achievement test items. A test user can then provide specifications

for the test items which will compose the desired customized test. When the items in a bank measure the same trait and are referenced to a common scale, performance on one set of items should be able to predict performance on another set of items (i.e., the set of items which were nationally normed) (Yen, Green, & Burket, 1987; Hambleton & Martois, 1983). A study by Hambleton and Martois (1983) looked at the accuracy of normed test score predictions from different sets of items in the same item bank.

Four 50-item achievement tests were created in three subject areas (reading, language arts, math) at grades 2 and 5: normed, easy, medium, and hard. The "normed" test was composed of items selected for being most representative of national curricula. The items for the easy, medium, and hard tests were selected to cover similar content as that in the normed test but at varying difficulty levels. A representative national sample of approximately 2,500 students in each subject area completed the normed test and one other test which was selected at random from the remaining three (easy, medium, or hard).

The analysis was centered on the comparison between the actual norm-referenced test scores in each subject area and the predicted test scores obtained from one of the other three forms (easy, medium, difficult) drawn from the item bank, using both the one- and three-parameter logistic test models. Results of this study were promising. Predictions from both the one- and three-parameter models showed almost no bias. Differences in the difficulty level of the tests seemed to adversely affect prediction accuracy, but not to an

alarming degree. Overall, errors were not much larger than the standard errors of measurement for the tests.

Yen, Green, and Burket (1987) supported this testing design as one that produces norm-valid scores as long as item statistics are up to standards and the content covered is proportional to the content covered in the normed test. If the test users specify content composition which is different from that represented in the normed test, they can "jeopardize the goal of obtaining a norm-valid customized test" (p. 12).

Since the study by Hambleton and Martois (1983) maintained relatively consistent content composition across the tests, the issue of content match did not come up, and was not examined. Test customization for purposes of test-curricular content match, on the other hand, would definitely bring this issue to the forefront.

2.3.5 CRT-Only

In the CRT-only model, a CRT is used to obtain both norm-referenced and criterion-referenced information. This design is made possible through the equating of the NRT and the CRT. Two different equating methods, equipercentile and IRT techniques, have been used; a study by the Texas Education Agency (1986) showed that results were identical with the two methods. Once the tests are equated, performance on the CRT can be used to estimate performance on the NRT, and the estimate of NRT performance can be expressed in normative terms.

According to the Standards for Educational and Psychological Testing (APA, 1985), this procedure probably yields comparable rather

than equated test scores since the two tests, a CRT and an NRT, are not likely to measure identical content or to have identical psychometric properties. Comparable scores cannot generally be used as substitutes for NRT scores, but they can be shown to be valid substitutions for certain purposes (Schattgen & Osterlind, 1988).

Example: In Missouri, statewide CRTs, the Missouri Mastery and Achievement Tests (MMAT) (Osterlind, 1987) have been equated to NRTs in order that they may produce normative information that is required for Chapter 1 evaluation. Using equipercentile methodology, the MMAT was equated to the norm-referenced Iowa Test of Basic Skills (ITBS) (Hieronymous & Hoover, 1986) in grades two through eight, and to the Test of Achievement and Proficiency (TAP) (Scannell, 1986) in grades nine and ten. This equating is redone each year rather than only once. Estimated scores are produced for individual students for use in Chapter 1 evaluation (Schattgen & Osterlind, 1988).

Two papers (Schattgen & Osterlind, 1988; Schattgen & Osterlind, 1989) investigate the effectiveness of this model as it is used in Missouri. These studies found that the equated tests are similar in terms of content and statistical properties. A validation study used chi-square procedures to examine the decision accuracy when selecting students who scored below the 45th percentile for Chapter 1 services and when selecting students who scored above the 90th percentile for gifted education services. A strong relationship was observed for Chapter 1 eligibility based on actual and estimated percentiles. At the upper end of the distribution, a much weaker relationship was observed, probably due to ceiling effects in the criterion-referenced tests.

Schattgen and Osterlind (1988) recommended further investigation into several aspects of the CRT-only model:

the worth of the CRT-ONLY MODEL relative to the other three models,

the appropriateness of equipercentile equating for obtaining comparable scores,

the effects of content and test level on the equating results,

the accuracy of comparable scores at the individual student level,

the accuracy of student level comparable scores in the low, middle, and high ranges of the distribution,

the validity of specific uses of comparable scores,

the effects of annual recalibration on the accuracy of comparable scores, and

the effects of instruction and, as a result, increasingly skewed CRT data, on the accuracy of comparable scores (p. 15).

The CRT-only model has definite strengths, but also has some nagging outstanding questions. It is a very desirable model from a CRT point of view since its base is a CRT designed to measure the given curriculum. However, its integrity as a norm-referenced design is less apparent. Schattgen and Osterlind (1989) found more accuracy at the 45th percentile than at the 90th percentile. Roudabush (1975) found overestimation at the low end and underestimation at the high end of the distribution. Hirsch and Keene (1989) found that a CRT with a dimensional structure (i.e., content specifications) which is more similar to the NRT produced less biased estimates. Questions are also unanswered regarding the use of aggregate level vs. individual level estimated norms and regarding the stability of the link over time and across different populations. There appears to be a need for validity evidence to support particular applications of this model.

A recent review of customized testing methodology by Linn and Hambleton (1990) concludes that "customized tests and customized norms can yield valid information about performance both in relation to specific curriculum objectives and in relation to national norms," but recommends "cautious application with frequent checks on the validity of the norm referenced inferences" (p. 27). Some of their recommendations which are applicable to the CRT-only model are:

1. The content of the customized test should be closely matched to the content of the norm-referenced test.
2. Additional content areas, which are not included in the norm-referenced test should not be part of the calculation of norm-referenced scores.
3. The test length and test difficulty of the customized test should be similar to that of the norm-referenced test.
4. Equating results should be investigated periodically.

2.4 Summary

Historically, the time is right for the emergence of customized dual-purpose testing. Both norm-referenced testing and criterion-referenced testing are valued by the education and psychometric communities. With limits on resources available for testing programs, a customized test which can provide both CRT and NRT information is desirable. With the development of item response theory in the 1970's, new options for test customization are available.

Five models for customized dual-purpose testing have been presented. In each model, both criterion-referenced and norm-referenced information is provided, sometimes from an NRT,

sometimes from a CRT, and sometimes from a clever compromise. Each model has its own strengths as well as its own limitations.

The model which is investigated in this study is a variation of the CRT-only model. The literature review offers some preliminary evidence in Missouri's case that the CRT-only model can be used to produce valid estimated norms, at least in some applications. The literature review also raises many questions regarding the use of this model which are as yet unanswered.

C H A P T E R 3

RESEARCH METHODOLOGY

3.1 Introduction

In Connecticut's model, performance on the statewide criterion-referenced test (CRT), the Connecticut Mastery Test (CMT), is used to estimate performance on the norm-referenced Metropolitan Achievement Test, Sixth Edition (MAT6) so that MAT6 norms may be used to describe CMT performance. This study examined the extent and nature of the error associated with this estimation of MAT6 scores from CMT scores. In two subject areas (reading and mathematics) and at three grade levels (4, 6, and 8), the sample under study had both MAT6 scores as estimated from the CMT and actual MAT6 scores.

The research methods which were used in this study are presented in this chapter. The sample is described in Section 3.2 along with the methodology by which it was selected. In Section 3.3, characteristics of the two instruments are presented in three separate subsections: MAT6, CMT, and MAT6/CMT link. In Section 3.4, the details of the research design are explained, and in Section 3.5, the logistics of data collection are described.

3.2 Description of the Sample

The population of interest to this research is the group of fourth, sixth, and eighth grade public school students who took the CMT in the fall of 1989.

The data for this study was derived from the 1989 CMT/MAT6 equating study. As will be described in Section 3.3.3, the CMT and

MAT6 are equated each year to annually update the CMT/MAT6 link. For equating purposes, each year, all fourth, sixth, and eighth grade public school students taking the CMT, also take one subtest of the Metropolitan Achievement Test.

In the fall of 1989, five different MAT6 subtests were distributed among the CMT examinees at each grade level: Reading Comprehension, Mathematics Concepts, Mathematics Problem Solving, Mathematics Computation, and Language. These MAT6 subtests were distributed through a systematic sampling of the public schools in Connecticut. For each grade level (4, 6, and 8), all public schools which contained that particular grade were ordered alphabetically by town/district and within each town/district alphabetically by the name of the school. The MAT6 subtests were then systematically distributed down the list so that one school in each sequence of five schools took each subtest. The result is that a representative sample of approximately 3,000 - 6,000 students took each subtest of the MAT6 along with the CMT.

The students which comprise the sample used in this study were among those in schools which were selected to take either the MAT6 Reading Comprehension Subtest or the MAT6 Mathematics: Problem Solving Subtest. That is, in the fall of 1989, these students took the entire CMT and one subtest of the MAT6, either Reading Comprehension or Mathematics: Problem Solving. Of the groups who took each of these two subtests, one half, approximately 1,500 - 3,000, in each subject area at each grade level were systematically selected for inclusion in this study.

This sample contains a very large number of students who are representative of Connecticut's fourth, sixth, and eighth grade students. Important to this study, the students in the sample also represent the full ability range of CMT examinees. The subjects which were used in this study are summarized in Table 3.1.

Table 3.1
Numbers of Subjects in Each Sample

	Reading Comprehension	Mathematics: Problem Solving
Grade 4	3,202	2,871
Grade 6	2,028	2,300
Grade 8	1,589	1,912

3.3 Instrumentation

Two different instruments were used in this research study: the MAT6 and the CMT. In Section 3.3, each of these instruments is described in terms of its purpose, content, and psychometric characteristics. A subsection is also included which summarizes the relationship between the two tests.

3.3.1 Metropolitan Achievement Test, Sixth Edition (MAT6)

The MAT6 is a comprehensive norm-referenced achievement test battery which was published by The Psychological Corporation in 1985. It was designed to provide norm-referenced information in a full range of subject areas for students in kindergarten through grade twelve.

In the fall of 1984 and the spring of 1985, large-scale standardization studies were conducted. For each study, more than

200,000 students were selected to be representative of the nation's students on the following variables: geographic region, school system enrollment, socioeconomic status, and public vs. nonpublic schools. Through comparison to the performance of the students in these norming samples, examinees' performance on the MAT6 can be reported in terms of national percentile ranks, stanines, grade equivalents, and normal curve equivalents (NCEs).

The MAT6 Survey Battery is composed of ten subtests, three in the reading area, three in mathematics, two in language arts, and one each in science and social studies. The structure of the battery is outlined below:

Reading

- Vocabulary
- Word Recognition Skills
- Reading Comprehension

Mathematics

- Mathematics: Concepts
- Mathematics: Computation
- Mathematics: Problem Solving

Language Arts

- Spelling
- Language

Science

Social Studies

For this study, two subtests (Reading Comprehension and Mathematics Problem Solving) are of particular interest. Since these are the areas of primary focus for Chapter 1 Compensatory Education

Programs, these are the subtests which have been equated to the CMT for purposes of Chapter 1 evaluation. Corresponding to CMT administration years, three levels of the MAT6 were used in this study: Elementary, Intermediate, and Advanced 1. The Form L version was used. Table 3.2 summarizes internal reliability indices (KR-20) for the subtests which were studied. Table 3.3 summarizes the correlations between Form L and an alternate test, Form M, for the subtests under study. Reference to these correlations will assist in interpreting the data from this study.

3.3.2 Connecticut Mastery Test (CMT)

In 1984, the General Assembly of the State of Connecticut passed Education Evaluation and Remedial Assistance (EERA) legislation. EERA requires that school districts regularly assess the progress of their students, identify those in need of remedial assistance, provide the needed remedial assistance, and evaluate the effectiveness of their instructional programs. As part of the EERA legislation, the creation of mastery tests in the basic skill areas of mathematics and language arts was authorized.

The resulting Connecticut Mastery Test is a criterion-referenced test which is administered each fall to fourth, sixth, and eighth grade public school students in Connecticut. Test results are used along with other data to monitor the effectiveness of programs, to provide objective-based assessment for individual students, and to identify students in need of remediation.

The CMT was not designed to be a norm-referenced test. It was not administered to a national group of students, and normative inferences

Table 3.2

KR-20 Reliability Coefficients for the MAT6 Survey Battery,
Form L

Subtest	Elementary Level	Intermediate Level	Advanced 1 Level
Reading Comprehension	.95	.93	.94
Mathematics: Problem Solving	.85	.87	.88

(Prescott, Balow, Hogan, & Farr, 1986)

Table 3.3

Correlations of Alternate Forms of MAT6 Survey Battery,
Forms L and M

Subtest	Elementary Level	Intermediate Level	Advanced 1 Level
Reading Comprehension	.87	.86	.85
Mathematics: Problem Solving	.82	.84	.83

(Prescott, Balow, Hogan, & Farr, 1986)

cannot be directly made on the basis of the CMT. The CMT was designed to be a criterion-referenced test. It provides scores for students at an objective level and classifies students as masters or nonmasters of these specific objectives. Additionally, remedial standards which were established for reading, writing, and mathematics, allow identification of students who may be in need of remediation.

The CMT mathematics test is a multiple-choice test with objectives in four domains: Conceptual Understanding, Computational Skills, Problem Solving/Applications, and Measurement/Geometry. The CMT Language Arts Test has two domains: Reading/Listening and Writing/Study Skills. The Reading/Listening Domain has three parts: a multiple-choice reading comprehension subtest, a multiple-choice listening comprehension subtest, and the Degrees of Reading Power (DRP) test. The Writing/Study Skills Domain has three parts: a holistically scored writing sample, a multiple-choice writing mechanics subtest, and a multiple-choice study skills subtest.

For this study, only certain portions of the CMT were of interest. For Connecticut's original criterion-referenced purposes, all parts of the mastery test are utilized and are important. However, for the purpose of providing norms for evaluating compensatory education programs, only those portions of the CMT which can be most appropriately equated to the MAT6 subtests of interest are utilized. In reading, the CMT's multiple-choice reading comprehension subtest was used. In mathematics, the portions of the CMT math test which are most closely related to the MAT6 Problem Solving Subtest were used. These portions of the CMT are used to estimate MAT6

performance, but criterion-referenced inferences continue to be based on the full-scale CMT (Connecticut State Department of Education, 1987).

3.3.3 CMT/MAT6 Link

For the primary purpose of satisfying the evaluation requirements for the federally funded Chapter 1 Compensatory Education Program, portions of the CMT were equated to portions of the MAT6. Chapter 1 requires that compensatory education students be pretested and posttested on matched tests which can provide NCE scores. An increase in a student's NCE standing leads to an interpretation that the student made a greater gain than would have been expected in the absence of the compensatory education, and, therefore, the program was successful. In Connecticut, the CMT is given in the fall of fourth, sixth, and eighth grades, but it cannot, on its own, provide norm-referenced information such as NCEs. Through equating the CMT to the MAT6, estimated MAT6 norms can be obtained in fourth, sixth and eighth grades based on the CMT. If the MAT6 is administered in the non-CMT grades, a testing design is created whereby MAT6 norms and estimated MAT6 norms can be compared to obtain pretest-posttest gain. For example, MAT6 NCE standing in the fall of grade three can be compared to estimated MAT6 NCE standing derived from the CMT in the fall of grade four.

In response to the Chapter 1 need for evaluation data in the areas of reading comprehension and mathematics problem solving the CMT/MAT6 equating was focused in those areas. The MAT6 Reading Comprehension Subtest was equated to the multiple-choice reading comprehension

portion of the CMT, but the CMT did not have an intact section which was adequately similar to the MAT6 Problem Solving Subtest.

The CMT mathematics items which were equated to the MAT6 Problem Solving Subtest were selected on the basis of content coverage and their statistical contribution to the estimation of the MAT6 score in a stepwise regression analysis. The selected items are a combination of problem solving items and computation items. It was necessary to include many CMT computation items, since MAT6 problem solving items involve more computation than CMT problem solving items.

A list of MAT6 objectives was obtained from the publisher and used to analyze the content match between corresponding portions of the CMT and MAT6. The MAT6 reading comprehension objectives are listed in Tables 3.4, 3.5, and 3.6. All of the items from the CMT reading comprehension subtest were judged to fit into one of the MAT6 broad categories: Literal Comprehension, Inferential Comprehension, or Critical Analysis. However, some of the items did not fit neatly into the subcategories, particularly in the area of Critical Analysis.

The MAT6 mathematics problem solving objectives are listed in Tables 3.7, 3.8, and 3.9 along with corresponding CMT items. Since the CMT mathematics test is divided into two sessions at grade four and three sessions at grades six and eight, the CMT items are listed in columns labeled by testing session (e.g., Math I, Math II, and Math III). The testing sessions are not strictly associated with content categories.

Summary statistics to describe the score distributions of the CMT and MAT6 subtests are presented in Table 3.10. Correlations between

Table 3.4

Content Comparison of CMT and MAT6,
Grade 4 Reading Comprehension

(Total # CMT Items: 36)

Elementary Level MAT Objectives		Corresponding CMT Items	# CMT Items
C4-01	Literal Comprehension	2, 3, 5, 8, 9, 17, 20, 21, 26, 27, 31, 34	12
C4-011	Detail	2, 5, 8, 9, 20, 21, 27, 31, 34	9
C4-012	Sequence	3, 17, 26	3
C4-02	Inferential Comprehension	1, 4, 7, 10, 11, 12, 14, 15, 18, 24, 28, 29, 32, 33	14
C4-021	Inferred Meaning Figurative Language	10	1
C4-022	Cause and Effect	11, 18, 29	3
C4-023	Main Idea	15, 24, 33	3
C4-024	Character Analysis	1	1
C4-03	Critical Analysis	6, 13, 16, 19, 22, 23, 25, 30, 35, 36	10
C4-031	Drawing Conclusions	13, 22	2

Table 3.5

Content Comparison of CMT and MAT6,
Grade 6 Reading Comprehension

(Total # CMT Items: 36)

Intermediate Level MAT Objectives	Corresponding CMT Items	# CMT Items
C4-01 Literal Comprehension	1, 2, 8, 9, 21, 22, 32,	7
C4-011 Detail	1, 2, 8, 21, 32	5
C4-012 Sequence	9, 22	2
C4-02 Inferential Comprehension	3, 10, 11, 15, 17, 18, 19, 23, 24, 27, 28, 33, 34	13
C4-021 Inferred Meaning Figurative Language	15, 33	2
C4-022 Cause and Effect	17, 27	2
C4-023 Main Idea	3, 18, 24, 34	4
C4-024 Character Analysis	10, 11	2
C4-03 Critical Analysis	4, 5, 6, 12, 13, 14, 16, 20, 25, 26, 29, 30, 31, 35, 36	15
C4-031 Drawing Conclusions	7, 26	2
C4-032 Author's Purpose & Fact or Opinion	4, 5, 12, 16, 20, 35	6

Table 3.6

Content Comparison of CMT and MAT6,
Grade 8 Reading Comprehension

(Total # CMT Items: 36)

Advanced Level I MAT Objectives	Corresponding CMT Items	# CMT Items
C4-01 Literal Comprehension	1, 7, 8, 13, 19, 20,	6
C4-011 Detail	1, 7, 13, 19,	4
C4-012 Sequence	8, 20	2
C4-02 Inferential Comprehension	4, 9, 10, 14, 15, 21, 26, 27, 28, 31, 32, 33, 34	13
C4-021 Inferred Meaning Figurative Language	2, 25, 26, 32	4
C4-022 Cause and Effect	14	1
C4-023 Main Idea	2, 21, 25, 34	4
C4-024 Character Analysis	15	1
C4-03 Critical Analysis	3, 5, 6, 11, 12, 16, 17, 18, 22, 23, 24, 29, 30, 35, 36	15
C4-031 Drawing Conclusions	5, 11, 18, 22	4
C4-032 Author's Purpose & Fact or Opinion	16, 23	2

Table 3.7

Content Comparison of CMT and MAT6,
Grade 4 Mathematics: Problem Solving

(Total # CMT Items: 51)

Elementary Level MAT Objectives	Corresponding CMT Items		# CMT Items
	Math I	Math II	
E1 Problem Solving	32	25, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	15
E1-04 Add/Subtract Beyond Basic Facts No Regrouping		28, 35	2
E1-05 Add/Subtract Beyond Basic Facts With Regrouping	17, 18, 19, 20, 53		5
E1-06 Multiply/Divide Basic Facts	26, 27, 28, 49, 50, 51, 52	32	8
E2 Graphs & Statistics		16, 21, 22	3
E2-01 Charts & Graphs		16, 21, 22	3
Other CMT Items Used in the Equating	2, 10, 12, 22, 23, 24, 38, 39, 42, 44, 46, 47, 48, 54, 55, 56, 57, 58, 59, 60	4	21

Table 3.8

Content Comparison of CMT and MAT6,
Grade 6 Mathematics: Problem Solving

(Total # CMT Items: 55)

Intermediate Level MAT Objectives	Corresponding CMT Items			# CMT Items
	Math I	Math II	Math III	
E1 Problem Solving			9, 10, 12, 13, 16, 17, 18, 19, 20, 21, 24, 26, 28, 29, 32, 33, 34, 39, 41, 43, 44	21
E1-05 Add/Subtract Beyond Basic Facts With Regrouping	10, 12, 37		9, 10, 12, 37	7
E1-06 Multiply/Divide Basic Facts	1			1
E1-07 Multiply/Divide Beyond Basic Facts	15, 32	13, 15	13, 16, 18, 20, 39, 41	10
E1-09 Decimals & Fractions	22, 47, 58 60	26, 27, 28, 34	33, 34	10
E1-11 Multi-Step			19, 29, 32	3
E2 Graphs & Statistics			3, 4, 7,	3
E2-01 Charts & Graphs			3, 4, 7	3
Other CMT Items Used in the Equating	19, 25, 27, 33, 35, 36, 43	3, 5, 7, 17, 22, 24, 37, 39		15

Table 3.9

Content Comparison of CMT and MAT6,
Grade 8 Mathematics: Problem Solving

(Total # CMT Items: 75)

Advanced Level I MAT Objectives	Corresponding CMT Items			# CMT Items
	Math I	Math II	Math III	
E1 Problem Solving	8, 9, 10, 11, 18, 19, 20, 21, 22, 25, 27, 37 38, 39, 41, 44, 49, 50, 51, 52	18, 19, 20		23
E1-08 ASMD Beyond Basic Facts & Unlimited	9, 10, 11, 25, 27, 52, 55, 56	3, 4, 13		11
E1-09 Decimals & Fractions	11, 27, 37, 38 39, 56	15, 16, 25, 26, 27, 28, 32, 34, 35, 37, 38, 39 40	10, 16, 20, 23, 27, 28, 31, 37, 40, 42, 43	29
E1-10 Percents	8	29, 41, 43, 44	33, 34, 36, 41, 44	10
E1-11 Multi-Step	9, 10, 27, 39			4
E2 Graphs & Statistics	6, 7, 8, 21, 22		5	
E2-01 Charts & Graphs	6, 7, 8			3
Other CMT Items Used in the Equating	16, 29, 30, 32, 46, 47, 53	14	3, 4, 5, 8	12

Table 3.10

Summary of Score Distribution Indicators
 Connecticut Mastery Test (CMT), Form C
 Metropolitan Achievement Test, Sixth Edition (MAT6), Form L

Subtest	Number of Items	Mean	Standard Deviation	Mean p-Value	Skewness
Reading Comprehension					
CMT					
Grade 4	36	25.09	6.95	.70	-0.62
Grade 6	36	24.33	6.45	.68	-0.57
Grade 8	36	26.66	5.67	.74	-0.96
MAT6					
Grade 4	60	43.69	12.96	.73	-0.76
Grade 6	60	42.21	11.59	.70	-0.75
Grade 8	60	47.45	10.78	.79	-1.07
Mathematics: Problem Solving					
CMT					
Grade 4	51	38.59	9.03	.76	-0.86
Grade 6	55	40.62	9.55	.74	-0.65
Grade 8	75	53.09	13.50	.71	-0.42
MAT6					
Grade 4	30	20.42	5.11	.68	-0.39
Grade 6	30	23.95	4.55	.80	-1.10
Grade 8	30	22.72	5.20	.76	-0.67

the MAT6 and CMT raw scores and between the MAT6 scores and CMT estimates of MAT6 scores are presented in Table 3.11.

Annually, with each CMT administration, a new equating study is carried out. The link is re-established each year to correct for any drift that may occur over time, even though the link has proven to be quite stable over time. This study was based on 1989 data and uses the link that was established in 1989.

Table 3.11

Correlations Between MAT6 and CMT Subtests

Subtest	MAT6, Elementary CMT, Grade 4	MAT6, Intermediate CMT, Grade 6	MAT6, Advanced I CMT, Grade 8
Raw Score Correlations			
Reading Comprehension	.80	.76	.81
Mathematics: Problem Solving	.80	.85	.85
Correlations Between Actual MAT6 and Estimated MAT6			
Reading Comprehension	.79	.77	.82
Mathematics: Problem Solving	.81	.86	.85

The equating procedures used in the fall of 1989 employed the Rasch model, a one-parameter item response model. At each grade level (4, 6, and 8), a representative sample of approximately 3,000 - 6,000 students took the complete MAT6 Reading Comprehension Subtest along with the CMT. Similar samples of students at each grade level took the complete MAT6 Mathematics: Problem Solving Subtest along with the CMT. The Rasch model was then used to equate the reading comprehension subtests of the CMT and the MAT6 and to equate the problem solving subtest of the MAT6 with the designated set of math

items from the CMT. The steps used in the equating procedures are described below:

1. CMT and MAT6 items were calibrated together, as if they were one test, to obtain Rasch model parameter estimates. This placed all reading comprehension items (CMT and MAT6) on the same scale with an overall mean of zero. Likewise, all mathematics items (CMT and MAT6) were placed on the same scale with mean zero.
2. The data were then linked to the MAT6 scaled score system. This was done by adding an equating constant to the item difficulties derived for the MAT6 items in the calibration. The equating constant was the sum of (a) the additive inverse of the item difficulty mean of the MAT6 items in the equating sample, and (b) a MAT6 constant appropriate for the level of the test.
3. The same equating constant was added to each CMT item difficulty derived from the calibration in Step #1. Once the CMT item difficulties were on the MAT6 scale, they could be used to obtain ability estimates which could be linearly transformed to MAT6 scaled scores (Connecticut State Department of Education, 1987).

3.4 Research Design

The purpose of this study was to explore the extent and nature of the error associated with MAT6 scores as estimated by the CMT for public school students in Connecticut. The data set used in this research design was derived from the 1989 equating study data, and contains approximately 1,500 - 3,000 examinees per subject area and

grade level. In order to avoid a situation where the students in the sample under study are the same students who were in the equating study which established the link, the 1989 equating study data was divided into two equal, systematically-selected groups. Through a recalibration of one half of the 1989 equating study data, approximately 1,500 - 3,000 per subject area and grade level, a link was established. The remaining half of the 1989 equating study sample then became an independent, cross-validation sample for this study, approximately 1,500 - 3,000 in each subject area (reading and mathematics) at each grade level (4, 6, and 8).

Test scores from both the MAT6 and the CMT were obtained for each student in the sample. For students in the reading samples, two scores were of interest: the number of items answered correctly on the MAT6 Reading Comprehension Subtest and the number correct on the multiple-choice reading comprehension subtest of the CMT. For students in the mathematics samples, two scores were of interest: the number of items answered correctly on the MAT6 Mathematics: Problem Solving Subtest and the number correct on a subset of CMT mathematics items which was determined through content and statistical analyses to be the best predictor of the MAT6 Problem Solving Subtest.

For each student, a CMT reading score was used to estimate a MAT6 reading score, or a CMT mathematics score was used to estimate a MAT6 mathematics score. The manner in which these data were used to address the research questions specified in Section 1.3 is described below.

3.4.1 Research Area #1: Extent of Error/Bias for Groups and Individuals

- 1a. What is the extent of the error present in individual MAT6 scores as estimated from CMT performance?
- 1b. What is the extent and direction of the error present in group MAT6 scores as estimated from CMT performance?

In each subject area and at each grade level, estimated MAT6 scores were compared with actual MAT6 scores for the entire sample as groups and for individuals within the samples.

3.4.2 Research Area #2: Variation in Error Across the Score Distribution

- 2a. Are there differences in the degree of error present in individual MAT6 scores as estimated from CMT performance among three ability groups: low, middle, and high?
- 2b. Are there differences in the degree or direction of error present in group MAT6 scores as estimated from CMT performance among three ability groups: low, middle, and high?
- 2c. What error patterns can be observed across the score distribution?

Three groups were formed on the basis of MAT6 scores: high, middle, and low scorers. For each of these groups, individual and group error was analyzed, corresponding to Research Questions 2a and 2b.

Question 2c was addressed by examining MAT6 estimation at 10-point intervals on the MAT6 national percentile scale.

3.5 Data Collection/Editing

The data was derived from the 1989 CMT/MAT6 equating study. Datatapes were available at the Connecticut State Department of Education (CSDE) which contain item responses on both CMT and MAT6

items for the examinees in the equating study. Computer technical assistance was also available through CSDE to help with the retrieval of the relevant data from this database. The Psychological Corporation, a contractor of CSDE, recalibrated the CMT and MAT6 items using a systematically selected half of the students in the equating study, and indicated on the datatape which students were used in the calibration. The remaining half of the students became the sample used in this study.

CHAPTER 4

RESULTS

4.1 Introduction

In order to explore the accuracy of Connecticut's model which uses an equated criterion-referenced test (CRT) to estimate performance on a norm-referenced test (NRT), the difference between actual NRT scores and estimated NRT scores was examined from many angles. The model of parallel norm-referenced tests was selected for use as a comparison model to assist in interpreting the magnitude of the observed differences. All analyses were done for both subject areas (reading comprehension and mathematics: problem solving) and for all three grade levels (4, 6, and 8).

In this chapter, the results of these analyses are presented. Section 4.2 addresses Research Area #1: Extent of Error/Bias for Groups and Individuals, and Section 4.3 addresses Research Area #2: Variation in Error across the Score Distribution.

4.2 Research Area #1: Extent of Error/Bias for Groups and Individuals

For all examinees in all six datasets, three levels of reading and three levels of mathematics, both a MAT6 raw score and a CMT raw score were available. From the CMT raw score, an estimated MAT6 score was derived using equating tables. The difference between the actual MAT6 score (X) and the estimated MAT6 score (\hat{X}) is the value of interest in this study. For each examinee, a difference ($\hat{X} - X$) was computed and an absolute difference $|\hat{X} - X|$ was computed.

The means of these differences are presented in Table 4.1. The mean difference, an indicator of group bias ranged from -0.11 to -0.67 for the 60-item Reading Comprehension Subtest, and ranged from 0.08 to 0.66 for the 30-item Mathematics: Problem Solving Subtest. At all grade levels, the mean difference was negative for reading, indicating an underestimate of MAT6. For mathematics, however, the mean difference was consistently positive, indicating an overestimate. In all cases, the magnitude of the mean differences was no more than .67; that is, the group raw score never differed by more than a fraction of one point between the actual MAT6 score and the MAT6 score which was estimated by the CMT.

Table 4.1
Differences Between MAT6 Raw Scores
as Estimated from CMT Performance
and Actual MAT6 Raw Scores

Subtest	Number of Items	Mean Difference $\hat{X} - X$	Mean Absolute Difference $ \hat{X} - X $
Reading Comprehension			
Grade 4	60	- 0.11	5.72
Grade 6	60	- 0.67	5.23
Grade 8	60	- 0.26	4.59
Mathematics: Problem Solving			
Grade 4	30	0.20	2.58
Grade 6	30	0.66	1.82
Grade 8	30	0.08	2.07

The mean absolute difference is an indicator of individual fluctuation in scores. On the Reading Comprehension Subtest (60 items), the mean absolute difference ranged from 4.59 to 5.72 raw score points. On the Mathematics: Problem Solving Subtest, (30 items), the mean absolute difference ranged from 1.82 to 2.58 points.

There is always error present in test scores. Even if examinees took the very same test twice, one would not expect them to receive the exact same score. In order to interpret the magnitude of the differences reported in Table 4.1, a basis of comparison was necessary. An analysis was carried out to compare the standard error in three situations:

1. the same test is given twice (MAT6, Form L),
2. a score from a parallel form (Form M) is substituted for a MAT6 (Form L) score, and
3. a CMT estimate of MAT6 is substituted for an actual MAT6 score.

These standard errors are compared in Table 4.2 for reading and in Table 4.3 for mathematics.

The standard error of measurement is defined as "the error made in substituting the observed score for the true score" (Gulliksen, 1950, p.43). The standard error of measurement for MAT6, Form L was computed using the equation:

$$S_e = S_1 \sqrt{1 - r_{1m}}$$

where S_1 is the standard deviation from the norming sample for Form L and r_{1m} is the correlation of parallel forms (L and M) (See Table 3.3). The standard error of measurement for MAT6, Form L as reported in Tables 4.2 and 4.3 can be interpreted as the standard deviation of

Table 4.2

Comparison of Standard Error,
Using the Same Form, Parallel Forms, and CRT Estimate,
Reading Comprehension

Error Measurement	Test(s)	Grade 4	Grade 6	Grade 8
Standard Error of Measurement	MAT6, Form L	5.01	4.49	4.84
Standard Error of Substitution	MAT6, Form L MAT6, Form M	6.21	6.04	6.42
Standard Error of Substitution (observed)	MAT6, Form L CMT, Form C	7.92	7.50	6.19

Table 4.3

Comparison of Standard Error,
Using the Same Form, Parallel Forms, and CRT Estimate,
Mathematics: Problem Solving

Error Measurement	Test(s)	Grade 4	Grade 6	Grade 8
Standard Error of Measurement	MAT6, Form L	2.38	2.24	2.47
Standard Error of Substitution	MAT6, Form L MAT6, Form M	3.22	3.20	3.42
Standard Error of Substitution (observed)	MAT6, Form L CMT, Form C	3.31	2.37	2.74

the scores that examinees would receive if they took a particular subtest a large number of times.

The standard error of substitution is defined by Gulliksen (1950) as the "error made in substituting a score on one test for a score on a parallel form" (p. 40). The standard error of substitution for MAT6 parallel forms, Form L and Form M, was computed using the equation:

$$S_d = \sqrt{S_l^2 + S_m^2 - 2r_{lm}S_lS_m}$$

where S_l and S_m are the standard deviations for Forms L and M observed in the norming sample, and r_{lm} is the correlation between the parallel forms (See Table 3.3). The standard error of substitution for Forms L and M of MAT6 can be interpreted as the standard deviation of the differences between observed scores on Form L and observed scores on Form M.

The last row of data in Tables 4.2 and 4.3 is the standard error of substitution for the CMT and MAT6. This reported error was observed in the sample used in this study as the standard deviation of the differences between their actual MAT6 scores and their estimated MAT6 scores.

In all cases, the standard error of measurement using the same test is the smallest. Moving from a single test to either parallel forms or a CMT estimate increases the standard error by approximately 2 score points on the 60-item Reading Comprehension Subtest and approximately 1 score point on the 30-item Mathematics: Problem Solving Subtest. In reading, the parallel forms had a lower standard error of substitution in Grades 4 and 6, but the CMT estimate had a lower standard error of substitution at Grade 8. In mathematics, the

observed standard error of substitution using the CMT estimate was lower at Grades 6 and 8, but higher at grade 4, than what would have been expected if parallel forms were used.

Given the standard errors reported in Tables 4.2 and 4.3, and the mean differences (\bar{D}) reported in Table 4.1, confidence bands can be constructed which define the interval in which the difference between two test scores would be expected to fall approximately 68 percent of the time.

$$CI = \bar{D} \pm S_d, \text{ for 2 different tests}$$

$$CI = 0 \pm S_e, \text{ for the same test}$$

In Table 4.5, 68% confidence intervals are presented for the difference in raw scores which would be expected if two tests were administered: Form L of MAT6 given twice, Form L and Form M of MAT6, or MAT6 and the CMT estimate of MAT6. In calculating the confidence interval for the parallel tests, Form L and Form M, it was assumed that the mean difference would be zero, although that may not necessarily be the case; in calculating the confidence interval for the difference between the actual MAT6 and the CMT estimate of MAT6, the observed mean differences presented in Table 4.1 were used.

The data provided in Table 4.4 allow a comparison of the effects of substituting a parallel form and a CMT estimate for a MAT6, Form L score. According to these data, one would expect that 68% of the time a fourth grader taking the Reading Comprehension Subtest on two parallel forms of MAT6 would obtain a difference in raw scores in the range of -6.21 to +6.21. If a fourth grader took both the MAT6 and the CMT which provided an estimated MAT6 score, it would be expected that 68% of the time the difference between the estimated and actual

Table 4.4

Distribution of Differences in Raw Scores
on Two Test Administrations,
68% Confidence Intervals

Subtest	MAT6, Form L MAT6, Form L	MAT6, Form L MAT6, Form M	MAT6, Form L CMT Estimate $\hat{X} - X$
Reading Comprehension			
Grade 4	-5.01 < D < 5.01	-6.21 < D < 6.21	-8.03 < D < 7.81
Grade 6	-4.49 < D < 4.49	-6.04 < D < 6.04	-8.17 < D < 6.83
Grade 8	-4.84 < D < 4.84	-6.42 < D < 6.42	-6.45 < D < 5.93
Mathematics: Problem Solving			
Grade 4	-2.38 < D < 2.38	-3.22 < D < 3.22	-3.11 < D < 3.51
Grade 6	-2.24 < D < 2.24	-3.20 < D < 3.20	-1.71 < D < 3.03
Grade 8	-2.47 < D < 2.47	-3.42 < D < 3.42	-2.66 < D < 2.82

MAT6 scores ($\hat{X} - X$) would be in the range of -8.03 to +7.81. Thus, in the case of fourth grade reading comprehension, both the upper and lower limit of the 68% confidence interval for the difference is more extreme in the case of the CMT estimate than in the case of parallel forms. At grade 6, the confidence interval is more extreme at both ends for the CMT estimate as well, but at grade 8, both ends of the confidence interval are more extreme in the case of parallel forms.

For the Mathematics: Problem Solving Subtest, different patterns can be observed. At grades 6 and 8, both the upper and lower limits of the confidence intervals are more extreme in the case of parallel forms than in the case of the CMT estimate. At grade four, however, the upper limit is more extreme for the CMT estimate and the lower limit is more extreme for the parallel test; this is possible since different mean differences were used.

4.3 Research Area #2: Variation in Error Across the Score Distribution

This research area was explored in response to related research by Schattgen & Osterlind (1988) in which they found a similar equating model to have a different degree of accuracy at different parts of the score distribution.

The first analysis in this area was done by disaggregating the data into three levels: low, middle, and high. The levels were defined by examining the frequency distribution of actual MAT6 raw scores and determining cutpoints which would most nearly yield three groups of equal size. For each of these groups, the mean difference ($\hat{X} - X$) and the absolute difference $|\hat{X} - X|$ were computed.

Upon review of these computed values, an anticipated regression effect was confirmed. In order to separate the regression effect from actual error in the testing model, a correction was needed. The following equation, derived from the regression of observed scores on true scores (Lord & Novick, 1968), was used to compute a correction factor for each level.

$$\text{Correction Factor} = \left[1 - r_{XX}^{\hat{X}} \frac{\sigma_{\hat{X}}}{\sigma_X} \right] \left[\bar{X}_t - \bar{X}_i \right]$$

where $r_{XX}^{\hat{X}}$ is the correlation between actual MAT6 scores (X) and estimated MAT6 scores (\hat{X}), \bar{X}_t is the mean MAT6 score for the entire population and \bar{X}_i is the mean MAT6 score for the examinees in the subgroup of interest. This correction factor was added to the actual MAT6 score of each examinee before calculating the difference scores.

Both uncorrected and corrected differences are presented in Table 4.5 for Reading Comprehension. After correcting for the regression effect, there was less than a one item point difference between estimated and actual MAT6 raw scores at all levels at all grades with

one exception. At grade six, the lower part of the score distribution exhibited a mean difference with correction for regression of -1.33. The mean absolute difference shows that there was the most accurate individual score estimation at the upper end of the score distribution and the least accurate individual score estimation at the lower end of the score distribution.

Differences for three levels of examinees in mathematics are reported in Table 4.6 with and without correction for regression. With correction, mean differences between estimated and actual MAT6 scores were very small at grades 4 and 8 at all three levels. At grade 6, the mean differences were larger, ranging from 0.53 to 0.70, indicating an overestimation across all levels. As was observed in reading comprehension, the mean absolute differences indicate more accurate individual score estimation at the upper end of the score distribution and less accurate score estimation for examinees at the low end of the distribution.

The second analysis done in this research area is very similar except that it was done at much finer levels. The levels which were used are based on the MAT6 national percentile rank scale. Groups were formed for every decile, ten point percentile interval, the purpose being to observe error patterns across the score distribution. Again, a correction for a regression effect was necessary, and the same differences were calculated: mean difference and mean absolute difference.

The results of the analysis by percentile group level are presented in Tables 4.7 through 4.12. No particular pattern is observable in the corrected mean differences ($\hat{X} - X$) across percentile

rank groups. The one exception is in grade 6 reading comprehension (Table 4.8) where a clear interaction is present, with more group error occurring at the lower percentile groups. The mean differences in grade 6 mathematics are consistently more than grade 4 and grade 8 mathematics, but that greater difference is present across all percentile groups with no apparent interaction.

The mean absolute differences indicate the most accurate individual estimation at the highest percentile groups. As the percentile group gets lower there is a decline in individual estimation accuracy for all grade levels in both subject areas.

Table 4.5

Differences Between Estimated and Actual MAT6 Raw Scores
By Level of MAT6 Score (X)
With and Without Correction for Regression Effect,
Reading Comprehension

Level	N	Mean Difference		Mean Absolute Difference	
		$\hat{X} - X$ Uncorrected	Corrected	$ \hat{X} - X $ Uncorrected	Corrected
Grade 4					
Low (X < 39)	1021	4.91	-0.33	7.71	6.87
Middle	1079	-1.35	-0.51	5.29	5.20
High (X > 52)	1102	-3.56	0.48	4.30	3.54
Grade 6					
Low (X < 38)	628	3.04	-1.33	7.06	7.11
Middle	628	-1.29	-0.96	4.65	4.62
High (X > 48)	732	-3.30	0.15	4.18	3.30
Grade 8					
Low (X < 45)	515	4.17	-0.24	6.35	5.44
Middle	472	-1.10	-0.36	3.73	3.67
High (X > 53)	602	-3.38	-0.18	3.76	2.47

Table 4.6

Differences Between Estimated and Actual MAT6 Raw Scores
 By Level of MAT6 Score (X)
 With and Without Correction for Regression Effect,
 Mathematics: Problem Solving

Level	N	Mean Difference		Mean Absolute Difference	
		$\hat{X} - X$ Uncorrected	Corrected	$ \hat{X} - X $ Uncorrected	Corrected
Grade 4					
Low (X < 18)	826	0.86	0.06	2.96	2.88
Middle	938	0.35	0.31	2.75	2.75
High (X > 22)	1107	-0.43	0.20	2.16	2.20
Grade 6					
Low (X < 22)	549	2.03	0.53	2.86	2.36
Middle	969	0.62	0.69	1.80	1.83
High (X > 26)	782	-0.26	0.70	1.11	1.33
Grade 8					
Low (X < 21)	591	1.54	0.00	2.68	2.38
Middle	628	-0.02	0.06	2.00	2.01
High (X > 25)	693	-1.08	0.17	1.62	1.47

Table 4.7

Differences Between Estimated and Actual MAT6 Raw Scores
By National Percentile Rank (P) of MAT6 Scores
With and Without Correction for Regression Effect,
Reading Comprehension, Grade 4

Percentile Group	N	Mean Difference		Mean Absolute Difference	
		$\hat{X} - X$	$\hat{X} - X$	$ \hat{X} - X $	$ \hat{X} - X $
		Uncorrected	Corrected	Uncorrected	Corrected
0 < P ≤ 10	272	9.68	1.03	10.57	8.55
10 < P ≤ 20	280	4.77	-0.92	7.49	6.58
20 < P ≤ 30	234	2.74	-0.95	6.74	6.14
30 < P ≤ 40	235	1.74	-0.59	5.61	5.54
40 < P ≤ 50	248	1.32	0.32	6.00	5.87
50 < P ≤ 60	272	-0.92	-0.66	5.25	5.25
60 < P ≤ 70	262	-2.24	-0.83	5.13	4.82
70 < P ≤ 80	297	-3.17	-0.79	4.89	4.30
80 < P ≤ 90	442	-3.73	-0.39	4.82	3.87
90 < P < 100	660	-3.45	1.07	3.96	3.47

Table 4.8

Differences Between Estimated and Actual MAT6 Raw Scores
By National Percentile Rank (P) of MAT6 Scores
With and Without Correction for Regression Effect,
Reading Comprehension, Grade 6

Percentile Group	N	Mean Difference		Mean Absolute Difference	
		$\hat{X} - X$	$\hat{X} - X$	$ \hat{X} - X $	$ \hat{X} - X $
		Uncorrected	Corrected	Uncorrected	Corrected
0 < P ≤ 10	113	11.63	3.25	12.81	10.57
10 < P ≤ 20	169	3.44	-1.68	6.21	5.86
20 < P ≤ 30	169	0.33	-3.02	5.81	6.23
30 < P ≤ 40	177	-0.24	-2.32	5.40	5.75
40 < P ≤ 50	212	-0.94	-1.77	5.00	5.14
50 < P ≤ 60	254	-0.97	-0.59	4.40	4.36
60 < P ≤ 70	202	-2.05	-0.57	4.60	4.40
70 < P ≤ 80	230	-2.88	-0.49	4.29	3.77
80 < P ≤ 90	213	-3.04	0.24	4.09	3.40
90 < P < 100	289	-3.82	0.59	4.16	2.85

Table 4.9

Differences Between Estimated and Actual MAT6 Raw Scores
By National Percentile Rank (P) of MAT6 Scores
With and Without Correction for Regression Effect,
Reading Comprehension, Grade 8

Percentile Group	N	Mean Difference		Mean Absolute Difference	
		$\hat{X} - X$ Uncorrected	Corrected	$ \hat{X} - X $ Uncorrected	Corrected
0 < P ≤ 10	66	8.97	-0.67	9.88	7.68
10 < P ≤ 20	115	5.00	-1.16	6.84	5.65
20 < P ≤ 30	117	5.51	1.41	6.74	5.01
30 < P ≤ 40	108	1.78	-0.88	4.69	4.59
40 < P ≤ 50	142	1.16	-0.17	4.81	4.52
50 < P ≤ 60	126	0.59	0.47	3.86	3.84
60 < P ≤ 70	170	-1.49	-0.61	3.65	3.53
70 < P ≤ 80	222	-2.46	-0.59	3.45	2.94
80 < P ≤ 90	179	-2.97	-0.26	3.61	2.86
90 < P < 100	344	-3.82	-0.15	3.96	2.16

Table 4.10

Differences Between Estimated and Actual MAT6 Raw Scores
By National Percentile Rank (P) of MAT6 Scores
With and Without Correction for Regression Effect,
Mathematics: Problem Solving, Grade 4

Percentile Group	N	Mean Difference		Mean Absolute Difference	
		$\hat{X} - X$ Uncorrected	Corrected	$ \hat{X} - X $ Uncorrected	Corrected
0 < P ≤ 10	112	2.34	0.90	3.18	2.77
10 < P ≤ 20	175	0.76	-0.25	3.03	2.96
20 < P ≤ 30	239	0.63	-0.09	2.97	2.96
30 < P ≤ 40	149	0.66	0.12	2.81	2.74
40 < P ≤ 50	313	0.59	0.23	2.79	2.78
50 < P ≤ 60	369	0.38	0.27	2.76	2.76
60 < P ≤ 70	221	0.10	0.17	2.74	2.75
70 < P ≤ 80	381	-0.04	0.22	2.60	2.64
80 < P ≤ 90	399	-0.09	0.41	2.32	2.40
90 < P < 100	513	-0.73	0.11	1.92	1.91

Table 4.11

Differences Between Estimated and Actual MAT6 Raw Scores
By National Percentile Rank (P) of MAT6 Scores
With and Without Correction for Regression Effect,
Mathematics: Problem Solving, Grade 6

Percentile Group	N	Mean Difference		Mean Absolute Difference	
		$\hat{X} - X$		$ \hat{X} - X $	
		Uncorrected	Corrected	Uncorrected	Corrected
0 < P ≤ 10	84	4.36	1.43	4.57	2.91
10 < P ≤ 20	104	2.89	0.89	3.28	2.12
20 < P ≤ 30	164	1.29	-0.03	2.29	2.09
30 < P ≤ 40	77	1.39	0.50	2.40	2.16
40 < P ≤ 50	255	0.96	0.41	2.25	2.12
50 < P ≤ 60	373	0.70	0.61	2.02	2.00
60 < P ≤ 70	223	0.58	0.82	1.65	1.75
70 < P ≤ 80	238	0.40	0.86	1.40	1.59
80 < P ≤ 90	248	-0.09	0.60	1.37	1.61
90 < P < 100	534	-0.34	0.75	0.98	1.22

Table 4.12

Differences Between Estimated and Actual MAT6 Raw Scores
By National Percentile Rank (P) of MAT6 Scores
With and Without Correction for Regression Effect,
Mathematics: Problem Solving, Grade 8

Percentile Group	N	Mean Difference		Mean Absolute Difference	
		$\hat{X} - X$		$ \hat{X} - X $	
		Uncorrected	Corrected	Uncorrected	Corrected
0 < P ≤ 10	55	4.16	0.95	4.38	2.66
10 < P ≤ 20	150	2.13	-0.04	2.85	2.15
20 < P ≤ 30	125	1.47	0.00	2.40	2.20
30 < P ≤ 40	155	0.89	-0.11	2.47	2.30
40 < P ≤ 50	219	0.46	-0.07	2.19	2.18
50 < P ≤ 60	265	0.32	0.27	2.21	2.21
60 < P ≤ 70	119	-0.66	-0.35	1.87	1.88
70 < P ≤ 80	131	-0.64	-0.09	1.57	1.56
80 < P ≤ 90	297	-0.96	-0.06	1.74	1.59
90 < P < 100	396	-1.16	0.33	1.53	1.40

CHAPTER 5

CONCLUSIONS

5.1 Discussion of the Results

All analyses in this investigation were done on six different datasets: Reading Comprehension at Grades 4, 6, and 8, and Mathematics: Problem Solving at Grades 4, 6, and 8. The results of the analyses are quite consistent across the datasets with two exceptions. The Grade 6 reading and Grade 6 mathematics datasets showed some unique patterns. In this section, the general results observed in each of the research areas are discussed first, and then these two exceptional datasets are discussed in more detail.

5.1.1 Research Area #1: Extent of Error/Bias for Groups and Individuals

- a. What is the extent of the error present in individual Metropolitan Achievement Test, Sixth Edition (MAT6) scores as estimated from Connecticut Mastery Test (CMT) performance?
- b. What is the extent and direction of the error present in group MAT6 scores as estimated from CMT performance?

As is evident in the mean absolute differences reported in Table 4.1, individual estimates of MAT6 scores deviate from actual MAT6 scores an average of more than five score points on the 60-item Reading Comprehension Subtest and an average of more than two score points on the 30-item Mathematics: Problem Solving Subtest. The standard error of substitution observed when using CMT estimates was compared to the expected standard error of substitution for parallel

forms in Tables 4.2 and 4.3. This comparison showed that individual differences were surprisingly similar whether a parallel form or a CMT estimate were substituted for the MAT6 subtest. In fact, in three of the six datasets, the standard error was even greater in the case of parallel tests than in the case of the CMT estimate.

As is shown in Table 5.1, the standard error of substitution, not surprisingly, is closely related to the correlations between the two tests. In all cases except one, where the correlation coefficient was higher for the parallel forms, the standard error of substitution was lower for the parallel forms. Conversely, when the correlation coefficient was higher between the CMT estimate and the actual MAT6 score, that standard error of substitution was lower.

Table 5.1
Relationship Between Correlation Coefficients
and Standard Error of Substitution

Subtest	MAT6 and CMT Estimate Correlation Coefficient	Standard Error of Substitution	MAT6 (L) and MAT6 (M) Correlation Coefficient	Standard Error of Substitution
Reading Comprehension				
Grade 4	.79	7.92	.87*	6.21**
Grade 6	.77	7.50	.86*	6.04**
Grade 8	.82	6.19**	.85*	6.42
Mathematics: Problem Solving				
Grade 4	.81	3.31	.82*	3.22**
Grade 6	.86*	2.37**	.84	3.20
Grade 8	.85*	2.74**	.83	3.42

* Higher correlation coefficient

** Lower standard error of substitution

The bias for groups of examinees which is present in the estimation of MAT6 norms from CMT performance is shown by the mean differences in Table 4.1. For all subtests, the differences were less than one score point, and other than the exceptional cases, the differences are remarkably close to zero. This indicates that group data, such as that used to evaluate Chapter 1 programs, was accurately obtained through the use of this model.

In Connecticut's application of the CRT-only model, group means estimated from the CRT are very close to means observed in an actual NRT administration. On an individual level, Connecticut's estimation of NRT scores from the CRT seems to approximate the substitution of a parallel form of the NRT. How can the apparent success of the CRT-only model in Connecticut be reconciled with the many concerns and cautions associated with the model in current literature?

Keene and Holmes (1987) warn that "any norm-referenced scores computed with the CRT-only model must be used with extreme caution" (p. 22). Their major concerns are that the content on the two tests is likely to differ and that the criterion-referenced test is likely to be substantially easier than the norm-referenced test. Yen, Green, and Burket (1987) state regarding this model that "the local IRT calibration produces results that are NRT-equivalent for that sample of examinees at that time" (p. 10), but warn that this equivalence may not hold up for another group of examinees or even for the same examinees at another point in time. Perhaps, the manner in which these concerns have been addressed in Connecticut has contributed to their successful application of the CRT-only model.

The issue of content match is related to the item response theory (IRT) assumption of unidimensionality. Although it is not reasonable to think of a standardized achievement test as unidimensional, the theoretical success of IRT equating requires that the two tests be matched in their "multidimensionality." That is, they must measure similar content. In Connecticut, analyses were done to ensure adequate content similarity. For the problem solving subtest, CMT mathematics items were specifically selected for inclusion in the equating which were the best predictors of the MAT6 score.

The issue of differential difficulty is a very reasonable concern since CRTs do tend to be much easier than NRTs, and such differences in the score distribution will affect the equatability of the tests. In Connecticut's case, however, the Connecticut Mastery Test is an unusually challenging criterion-referenced test. As is reported in Table 3.10, the mean p-values for the CMT and the MAT6 subtests are quite similar, and, in fact, for five of the six datasets, the CMT subtest was more difficult than the MAT6 subtest.

The issue of population dependence questions whether the relationship between two tests established for one group of examinees is generalizable to other groups of examinees. The sample used in this study was representative of the students in Connecticut and independent of the sample used to establish the equating. The results of this study indicate the generalizability of these equating results to Connecticut students in general. There is no indication, however, that the CMT would be a good estimator of MAT6 for populations outside of Connecticut.

The issue of time dependence raises concern that local instruction is likely to be more aligned with the local CRT than with the more general NRT. In that case, performance on the CRT may be more sensitive to local instruction than the NRT. Consequently, instruction which takes place after the equating has been done may result in greater gains on the CRT than on the NRT, causing the equating to no longer be valid. This issue is not addressed in this study. However, until more is known about the effect of the local curricular emphases on the link between the CMT and MAT6 over time, this link will continue to be established annually. That is, the equating data is collected at the same time that the tests are taken to which the equating results will be applied.

5.1.2 Research Area #2: Variation in Error Across the Score Distribution

- a. Are there differences in the degree of error present in individual MAT6 norms as estimated from CMT performance among three ability groups: low, middle, high?
- b. Are there differences in the degree or direction of error present in group MAT6 scores as estimated from CMT performance among three ability groups: low, middle, high?
- c. What error patterns can be observed across the score distribution?

After correcting for a regression effect which was inherent in the research design, mean group differences between actual MAT6 scores and MAT6 scores which were estimated from the CMT were fairly consistent for groups across the score distribution (See Tables 4.5 through 4.12). Except in the cases of Grade 6 reading and Grade 6 mathematics, which will be addressed later, the group differences are

of reasonable magnitude across the score distribution with no particular relationship apparent between degree or direction of error and position on the score distribution. (Note: There are apparently some invalid scores at the lowest percentiles (0-10); they should be ignored except as they may have influenced other data.)

For individuals, a very dramatic pattern is apparent. For all six datasets, the mean absolute differences between estimated and actual MAT6 scores is greatest for the lowest score levels and steadily decreases as the score level increases. This model is estimating more accurately for better performing students in Connecticut.

This finding is contrary to work done by Schattgen and Osterlind (1989). In their study, Grade 3 reading tests, an NRT and a CRT, were administered to both an equating sample and a cross-validation sample. Equipercetile methodology was used to equate the two tests, making both actual and estimated NRT scores available for subjects in the cross-validation sample. These estimated and actual NRT scores were then used to select students for placement into Chapter 1 programs (at or below the 45th percentile) and for placement into gifted programs (at or above the 90th percentile). They found a much greater degree of agreement between placement decisions at the 45th percentile than they did at the 90th percentile.

The reason that Schattgen and Osterlind's data shows better estimation at lower percentile ranges and Connecticut's data shows better estimation at the higher percentile ranges appears to be related to the score distributions of the equated tests. In the Schattgen and Osterlind study, there was a significant ceiling effect in the CRT distribution. The CRT was much easier than the NRT in

their sample and the CRT was more negatively skewed. These distribution indicators are contrasted with those of the subtests which were equated in Connecticut in Table 5.2. In Connecticut, in all cases except Grade 4 mathematics, the NRT was the easier test with a higher mean p-value and a more negatively skewed distribution. In all Connecticut datasets, there is a steady increase in estimation accuracy as the percentile rank is increased; however, this trend is less dramatic in the fourth grade mathematics dataset.

Table 5.2

Mean p-Values and Skewness Indicators
for Tests Used in Connecticut and Tests
Used by Schattgen and Osterlind (1989)

Subtest	Criterion-Referenced Test		Norm-Referenced Test	
	Mean p-Value	Skewness	Mean p-Value	Skewness
Connecticut Tests				
Reading Comprehension				
Grade 4	.70	-0.62	.73*	-0.76**
Grade 6	.68	-0.57	.70*	-0.75**
Grade 8	.74	-0.96	.79*	-1.07**
Mathematics: Problem Solving				
Grade 4	.76*	-0.86**	.68	-0.39
Grade 6	.74	-0.65	.80*	-1.10**
Grade 8	.71	-0.42	.76*	-0.67**

Schattgen and Osterlind	.79*	-1.12**	.64	-0.37

* Higher mean p-value

** More negatively skewed

5.1.3 Exceptional Dataset: Grade 6 Reading Comprehension

The first data set which exhibited patterns that varied from what was generally observed was Grade 6 Reading Comprehension. The mean difference between estimated and actual MAT6 raw scores ($\hat{X} - X$) is reported in Table 4.1 as -0.67. That is, on average, MAT6 Reading Comprehension raw scores were underestimated by .67 points. Although this may not seem to be a dramatic degree of error, it is noticeably more extreme than the Grade 4 and Grade 8 datasets. The mean absolute difference, an indicator of estimation accuracy for individuals, is not different for the Grade 6 dataset. Table 4.8 shows another unique phenomenon; the underestimation is especially extreme at the lower percentile groups. This type of interaction between estimation error and position on the score distribution was not apparent in any of the other five datasets. Another curious bit of information is presented in Table 5.3; the same bias was not evident when the same analysis was done using the NCE scale instead of the raw score scale.

Table 5.3

Mean Difference Between Estimated MAT6 Scores
and Actual MAT6 Scores ($\hat{X} - X$)
Raw Scores and Normal Curve Equivalents (NCE)

Reading Comprehension

	Raw Score Difference	NCE Difference
Grade 4	-0.11	-1.94
Grade 6	-0.67	-1.71
Grade 8	-0.26	-3.13

In an attempt to find clues as to the cause of these variant patterns, all conversion tables were rechecked and test characteristics were reviewed. No errors were found in the conversion tables. As indicated in Table 3.10, the distributions of the two tests, CMT and MAT6, for Grade 6 Reading Comprehension are very similar both in terms of mean p-value and skewness indicators. The only indicator which is somewhat weaker for this dataset is the correlation between the two tests. As is reported in Table 3.11, the Grade 6 Reading Comprehension Subtests had the lowest correlation of the six datasets, .76 for raw scores on the two tests and .77 for estimated and actual MAT6 raw scores.

5.1.4 Exceptional Dataset: Grade 6 Mathematics Problem Solving

The second dataset which exhibited patterns that varied from what was generally observed was Grade 6 Mathematics: Problem Solving. As reported in Table 4.1, the mean absolute difference is very small for Grade 6 Mathematics, but the mean difference, group bias, is much more extreme at Grade 6 than at Grades 4 and 8. The mean difference of 0.66 indicates that on average group scores are overestimated by about .66 points. Table 4.11 shows that this overestimation is fairly consistent across the score distribution.

In reviewing the characteristics of the two problem solving tests, it was found that this dataset had the most highly correlated tests. As Table 3.11 shows, the correlation between the Grade 6 Mathematics: Problem Solving CMT and MAT6 raw scores is .85, and the correlation between the estimated and actual MAT6 raw scores is .86. These

correlations are even higher than the correlation between parallel MAT6 subtests, Forms L and M, which is .84 (See Table 3.3).

The distributions of the two subtests are somewhat different, however. As Table 3.10 indicates, the Grade 6 Mathematics: Problem Solving Subtest has a skewness indicator of -1.10. This is substantially more extreme than the CMT skewness indicator of -0.65.

5.2 Implications of the Study

For Chapter 1 students in Connecticut, the results of this investigation offer hope for a less intrusive and more cohesive program evaluation design. All fourth, sixth, and eighth grade students in Connecticut, including Chapter 1 students, are required by state legislation to take the CMT, and these results are the primary indicator of educational success in Connecticut. In addition, according to federal evaluation guidelines, Chapter 1 students must take a norm-referenced test each year. This creates a situation where this group of students, who are most likely to be traumatized by testing and who can least afford to give up instructional time, are subjected to twice the testing of the general population. Furthermore, it creates a situation where success is not clearly defined; is the real criteria for success the CMT or the norm-referenced test? With the model examined in this study, CMT performance alone can yield criterion-referenced information as well as the norm-referenced information which is needed for the federal evaluation.

For the more general body of students, teachers, and educational administrators, this work in Connecticut offers a model of one

methodology for obtaining both criterion-referenced and norm-referenced information from one test administration. This study shows that under certain circumstances this model can be used successfully. This study also shows that factors such as the correlations between the equated tests and the similarity of their score distributions can affect the accuracy with which norm-referenced scores are estimated. The sharing of this work offers other educators a basis on which to design a model to meet their needs.

For the field of psychometrics, this study provides strong evidence that under certain conditions a local criterion-referenced test can be used to provide national norms with reasonable accuracy. Hopefully, this work will stimulate the psychometric community to take a closer look at this and other models of customized testing, to actually examine the success of the models in practice rather than prematurely dismiss them on theoretical grounds.

Another important contribution of this study is the introduction of a new variation of the CRT-only model. Keene and Holmes (1987) describe the CRT-only model as a CRT being equated to an NRT so that norms can be estimated from CMT performance. They also describe a CRT-based model in which selected NRT items are embedded in a CRT to provide estimated norms. The approach used in Connecticut in mathematics is a new variation which worked well.

The CMT has a very large number of mathematics items which are all used for criterion-referenced score reporting. However, a subset of those mathematics items was selected to be equated to the MAT6 Mathematics: Problem Solving Subtest; these items were selected on the basis of a content review and a stepwise regression analysis to

determine which set of items was the best predictor of the MAT6 score. This procedure was very successful, yielding high correlations between the MAT6 subtest and the selected set of CMT mathematics items; at two of the three grade levels, the correlations were higher than the correlation between parallel forms of MAT6. This variation on the CRT-only model yielded very accurate estimation and is a promising area for future work.

5.3 Limitations of the Study

The generalizability of this study could be limited to the methodology that was employed in the study. Of the many methods of test customization which were discussed in the literature review (Chapter 2), this study examines only the CRT-only model. Furthermore, this study examined an application of one parameter item response theory (IRT) equating; it may not necessarily generalize to equipercentile equating or even to other IRT models of equating.

The generalizability of this study could also be limited by the characteristics of the instruments which were used, the CMT and the MAT6. These two tests have certain psychometric properties, and the relationship between the two tests (e.g., correlation) has certain characteristics. If the methodology in this study was applied to another set of tests with different psychometric properties, the results could be different.

This study is also limited by all of the uncontrollable factors which may be present in real data. Some data at the very lowest percentile ranks of the MAT6 distribution were rather bizarre (e.g., a score of 1 NCE on the MAT6 and 84 NCEs on the CMT). This real person

scored very well on one test and very poorly on the other. Rather than equating error, there is clearly a human factor which is impossible to identify and, therefore, impossible to control. It could be that an examinee had a serious change of mood, or that on one test the examinee lost his/her place on the answer sheet, or that he/she knew that the MAT6 didn't really "count". In working with real human beings, there are more factors at play than are defined by the research design.

The results of this investigation show strong support for the use of the CRT-only model under conditions similar to those in Connecticut. Variations in those conditions (e.g., different tests, different populations, different methodologies) could affect the success of this model. Which factors are critical and the degree to which variation could affect the estimation accuracy is not well understood at this time. This understanding is emerging from a collection of studies similar to this one done under different circumstances (Dungan, 1988; Green, 1987; Harris, 1987; Qualls-Payne, Raju, & Groth, 1989; Schattgen & Osterlind, 1989).

5.4 Recommendations for Further Research

The results of this study provide strong evidence that a criterion-referenced test can be used to estimate national norms under certain conditions; however, several questions remain. Four areas are identified below which would be meaningful research to follow this study.

1. Stability over time. This study applied equating results to data which was collected at the same time as the equating study data. It does not look at the effect of time and instruction on the relationship of the two tests.
2. Effect of group size. This study looked at the accuracy of estimated NRT scores for individuals and for very large groups. It would be interesting to examine the stability of estimation for groups of various sizes.
3. Effect of equating methodology. This study used one equating methodology, the one parameter (Rasch) IRT model. A comparison of different equating methods, equipercentile and the various IRT methods, would be valuable.
4. Effect of test characteristics. One pair of tests was used in this study with a given set of psychometric characteristics. Simulation research where tests were created according to a specified set of characteristics (dimensionality, skewness, mean p-value, correlation) and the equating methodology was applied to these carefully designed test score distributions would be very revealing.

BIBLIOGRAPHY

- Allen, N. L., Ansley, T. N., & Forsyth, R. A. (1987). The effect of deleting content-related items on IRT ability estimates. Educational and Psychological Measurement, 47, 1141-1153.
- American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author.
- Connecticut State Department of Education. (1987). Connecticut mastery test technical manual. Hartford, CT: Author.
- Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In Hambleton, R. K. (Ed.), Applications of item response theory (pp. 175-195). Vancouver, BC: Educational Institute of British Columbia.
- Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. International Journal of Education Research, 13, 161-173.
- Crocker, L., & Algina, J. (1986). Introduction to Classical & Modern Test Theory. New York, NY: Holt, Rinehart, & Winston.
- Diamond, E. E. (1984, April). Content considerations vs. growth in achievement testing: Hobson's choice? Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.
- Dungan, L. A. (1988, April). Norm-referenced test customization: Validation of individual score interpretation. Paper presented at the annual meeting of the National Council on Measurement in Education in New Orleans, LA.
- Goldsby, C. (1988, April). Norm-referenced test customization: Curricular considerations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Good, R. H., & Salvia, J. (1988). Curriculum bias in published, norm-referenced reading tests: Demonstrable effects. School Psychology Review, 17, 51-60.
- Green, D. R. (1987). Local versus national calibrations. Paper presented at the annual meeting of the American Educational Research Association in Washington, DC.
- Green, D. R., & Yen, W. M. (1984, April). Content and construct validity of norm-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

- Gulliksen, Harold (1950). Theory of mental tests. New York, NY: John Wiley & Sons, Inc.
- Hambleton, R. K. (1985). Criterion-referenced assessment of individual differences. In C. E. Reynolds, & V. L. Wilson (Eds.), Methodological and statistical advances in the study of individual differences. New York, NY; Plenum Press.
- Hambleton, R. K., & Martois, J. S. (1983). Evaluation of a test score prediction model based upon item response model principles and procedures. In Hambleton, R. K. (Ed.), Applications of item response theory (pp. 196-211). Vancouver, BC: Educational Institute of British Columbia.
- Hambleton, R. K., & Rogers, H. J. (1989). Solving Criterion-referenced measurement problems with item response models. International Journal of Educational Research, 13, 145-160.
- Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.
- Harris, D. J. (1987, April). Estimating examinee achievement using a customized test. Paper presented at the annual meeting of the American Educational Research Association in Washington, DC.
- Hieronymous, A. N., & Hoover, H. D. (1986). Iowa tests of basic skills. Chicago, IL: Riverside Publishing Co.
- Hirsch, T. M., & Keene, J. M. (1989, March). Derived norms for customized tests: An examination of content dimensionality. Paper presented at the annual meeting of the National Council on Measurement in Education in San Francisco, CA.
- Jaeger, R., & Tittle, C. (Eds.) (1980). Minimum competency achievement testing: Motives, models, measures, and consequences. Berkeley, CA: McCutchan.
- Jolly, S. J., & Gramenz, G. W. (1984). Customizing a norm-referenced achievement test to achieve curricular validity: A case study. Educational Measurement: Issues and Practices, 3, 16-18.
- Kean, M. H. (1986, April). Testing and the curriculum: A publisher's perspective. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.
- Keene, J. M., & Holmes, S. E. (1987, April). Obtaining norm-referenced test information for local objective-referenced tests: Issues and challenges. Paper presented at the meeting of the National Council on Measurement in Education, Washington, DC.

- Linn, R. L., & Hambleton, R. K. (1990). Customized tests and customized norms. (CRESST Technical Report). Los Angeles, CA: UCLA School of Education.
- Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Menlo Park, CA: Addison-Wesley Publishing Co.
- Mehrens, W. A. (1984). National tests and local curriculum: Match or mismatch? Educational Measurement: Issues and Practices, 3, 9-15.
- Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences on achievement test data. Journal of Educational Measurement, 23, 185-196.
- Osterlind, S. J. (1987). Missouri mastery and achievement tests. Jefferson City, MO: Missouri Department of Elementary and Secondary Education.
- Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.
- Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. (1986). Metropolitan achievement tests national norms booklet. San Antonio, TX: The Psychological Corporation.
- Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. (1988). Metropolitan achievement tests technical manual. San Antonio, TX: The Psychological Corporation.
- Qualls-Payne, A. L., Raju, N.S., & Groth, M.A. (1989, March). Accuracy of the estimation of national item p-values of a customized test as a function of core test length, sample size and IRT model. Paper presented at the annual meeting of the American Educational Research Association in San Francisco, CA.
- Roudabush, G. E. (1975, April). Estimating normative scores from a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association in Washington, DC.
- Scannell, D. P. (1986). Tests of achievement and proficiency. Chicago, IL: Riverside Publishing Co.
- Schattgen, S. F., & Osterlind, S. J. (1988, April). Estimating norm-referenced information from a criterion-referenced test: an application of the ORT only model. Paper presented at the annual meeting of the National Council on Measurement in Education in New Orleans, LA.

- Schattgen, S. F., & Osterlind, S. J. (1989, March). The validity of norm-referenced information obtained from an objective-referenced test using the ORT only model. Paper presented at the annual meeting of the National Council on Measurement in Education in San Francisco, CA.
- Schmidt, W. H. (1983). Content biases in achievement tests. Journal of Educational Measurement, 20, 165-178.
- Taleporos, B., Canner, J., Strum, I., & Faulkner, D. (1988, April). The process of customization of the Metropolitan Achievement Test (MAT6) in mathematics for New York City public school students. Paper presented at the annual meeting of the American Educational Research Association in New Orleans, LA.
- Texas Education Agency. (1986). Report on providing national comparative data on the TEAMS test. Austin, TX: Author.
- Way, W. D., Forsyth, R. A., & Ansley, T. N. (1989). IRT ability estimates from customized achievement tests without representative content sampling. Applied Measurement in Education, 2, 15-35.
- Wilson, S. M., & Hiscox, M. D. (1984). Using standardized tests for assessing local learning objectives. Educational Measurement: Issues and Practices, 3, 19-22.
- Yen, W. M., Green, D. R., & Burket, G. R. (1987). Valid normative information from customized achievement tests. Educational Measurement: Issues and Practices, 6, 7-13.

