2008

Investigation of the impact of various factors on the validity of customized norms

Xiaohui Zhao
University of Iowa

INVESTIGATION OF THE IMPACT OF VARIOUS FACTORS ON THE VALIDITY OF CUSTOMIZED NORMS

by

Xiaohui Zhao

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
(Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

August 2008

Thesis Supervisor: Associate Professor Timothy Ansley

ABSTRACT

Investigating the possibility of customizing off-the-shelf tests to provide various kinds of information is becoming increasingly interesting due to high information demands in the current testing environment. Comparing examinee achievement status on a national basis using such tests may provide a cost-effective solution for some practical problems. However, the normative estimates based on customized tests may be very different from those based on intact tests, and the validity of customized norms may be seriously compromised. The primary purpose of this study was to investigate the impact of various factors on the validity of customized norms. These factors included customizing strategy, estimating items, test length, correlations of latent abilities assessed by items from an intact test and new items, and test dimensional structures.

Monte Carlo simulation techniques were used to examine the accuracy of the customized norms. Both unidimensional and multidimensional data sets were generated and calibrated using unidimensional item response theory models. The five factors cited above were manipulated in a partially crossed design, with a total of 44 combinations of conditions. The outcomes of interest included estimated ability distributions and correlations, mean differences, mean absolute differences of ability and percentile estimates derived from intact tests and customized tests.

Based on the results of this study, it was concluded that: (1) customized instruments with all items from intact tests provided more accurate normative estimates than instruments having some items from intact tests removed; (2) using only items from intact tests to derive norms yielded more accurate estimates than using all items in customized tests; (3) lengthened customized tests yielded more accurate estimates than shortened tests; (4) the higher the correlation of latent abilities measured by items from intact tests and new items,

the more accurate the normative estimates; (5) the impacts of the various factors were small when the unidimensionality assumption was satisfied; the differences increased when data structures became more complicated.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

INVESTIGATION OF THE IMPACT OF VARIOUS FACTORS ON THE VALIDITY OF CUSTOMIZED NORMS

by

Xiaohui Zhao

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
(Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

August 2008

Thesis Supervisor: Associate Professor Timothy Ansley

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Xiaohui Zhao

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Psychological and Quantitative Foundations (Educational
Measurement and Statistics) at the August 2008 graduation.

Thesis Committee: _____
Timothy Ansley, Thesis Supervisor

_____
Stephen Dunbar

_____
Andrew Dean Ho

_____
Michael J. Kolen

_____
Mary Kathryn Cowles

ACKNOWLEDGMENTS

The past few years have been quite an adventure. There were many challenges, highs, and lows, but I am fortunate enough to have amazing support. All of you know who you are and I cannot possibly list all of your names here. Words alone alone cannot express my gratitude.

TABLE OF CONTENTS

LIST OF TABLES v

CHAPTER I. INTRODUCTION 1

Validity of Customized Norms 2
Factors Related to the Validity of Customized Norms 4
Customizing Strategies 4
Using Different Item Sets to Derive Customized Norms 6
Matching Tests on Dimensionality 7
Test Length 8
Correlation between Latent Abilities 9
Preview of the Study 9

CHAPTER II. REVIEW OF LITERATURE 12

Analyzing Multidimensional Data Using a Unidimensional Model 12
Data Simulation Using IRT Models 12
Previous Findings and Applications 16
Implications for the Present Study 19
Customized Tests and Customized Norms 19
Test Customizing Approaches 20
Matching Tests on Dimensionality 22
Test Length 24
Using Different Item Sets to Derive Customized Norms 26
Correlations of Latent Abilities 27
Implications for the Present Study 28

CHAPTER III. METHODS 31

Research Questions 31
Instruments 32
Intact Tests 32
Customized Instruments 34
Data Set Notations 36
Simulation Procedure 41
IRT Models for Simulation 41
Item Parameters 43
Simulees 45
Item Response Matrix 45
Research Design 46
Data Analysis Procedures 47
Data Analysis 47
Evaluation Criteria 49
Computer Programs 51

CHAPTER IV. RESULTS AND DISCUSSION 52

Description of Simulated Data and Item Parameter Estimates 52
Results for Unidimensional Data Sets 56

Using Different Estimation Items....57
Customizing Strategy....58
Test Length....58
Combined Effects of Customizing Strategies and Estimation Items....59
Results for Multidimensional Data Sets....60
Using Different Estimation Items....60
Customizing Strategy....61
Test Length....64
Combined Effects of Customizing Strategies and Estimation Items....64
Correlations of Latent Abilities....66
Comparing Unidimensional and Multidimensional Data Sets....70
Using Different Items for Norms Estimation....70
Customizing Strategies....71
Test Length....72
Summary....73

CHAPTER V. CONCLUSTIONS AND IMPLICATIONS....75

Findings....75
Test Dimensional Structures....75
Customizing Strategy....76
Items Used for Normative Estimation....77
Correlations of Latent Abilities....78
Test Length....78
Practical Implications....79
Group-Based Inferences....79
Individual Inferences....80
Strengths and Limitations....84
Strengths of the Study....84
Limitations....84
Recommendations for Future Studies....85
Conclusions....86

REFERENCES....88

APPENDIX. TABLES....92

LIST OF TABLES

Table 3-1. Descriptions of the Data Sets Representing Customized Tests.....33

Table 3-2. Notations of Averages and Individual Data Sets.....37

Table 3-3. Notation Related to Customizing Strategies under Different Dimensional Structures.....38

Table 3-4. Notation Related to Using Different Estimating Items under Different Dimensional Structures.....39

Table 3-5. Notation Related to Changing Test Length under Different Dimensional Structures.....40

Table 4-1. Descriptions and Notations for Unidimensional Data Sets Based on Manipulated Factors.....57

Table 4-2. Descriptions and Notations Related to Items for Normative Estimation.....61

Table 4-3. Descriptions and Notations Related to Customizing Strategies.....62

Table 4-4. Multidimensional Data Sets Related to the Combined Effects of Customizing Strategies and Estimation Items.....64

Table A-1. Item Parameters Used to Simulate Unidimensional and Multidimensional Data.....92

Table A-2. Summary Statistics of Unidimensional Item Parameter Estimates for Intact Tests and Customized Data Sets.....93

Table A-3. Unidimensional Item Parameter Estimates of Multidimensional Data Sets.....94

Table A-4. Estimated Ability Distributions of the Data Sets Based on the Unidimensional Intact Test.....95

Table A-5. Estimated Ability Distributions for the Two-dimensional Intact Test and Related Customized Tests.....96

Table A-6. Average Proportion Correct Scores for Data Sets Representing Intact and Customized Tests.....97

Table A-7. Average Reliability Estimates for Data Sets Representing Intact and Customized Tests.....98

Table A-8. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Sets Using Different Estimation Items under Unidimensionality.....99

Table A-9. Comparing Ability Estimates and Percentile Estimates Derived from Unidimensional Data Sets Using Different Items in Estimation.....99

Table A-10. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Sets Representing Different Customizing Strategies under Unidimensionality.....100

Table A-11. Comparing Ability Estimates and Percentile Estimates Derived from Data Sets Representing Different Customizing Strategies under Unidimensionality.....100

Table A-12. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Sets Representing Tests of Different Lengths under Unidimensionality.....101

Table A-13. Comparing Ability Estimates and Percentile Estimates Derived from Data Sets Representing Tests of Different Lengths under Unidimensionality.....101

Table A-14. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Sets Representing the Combined Effects of Customizing Strategies and Estimation Items under Unidimensionality.....102

Table A-15. Comparing Ability Estimates and Percentile Estimates Derived from Data Sets Representing the Combined Effects of Customizing Strategies and Estimation Items under Unidimensionality 102

Table A-16. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Multidimensional Data Sets Using Different Items to Estimate Norms.....103

Table A-17. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Using Different Items to Estimate Normative Information 104

Table A-18. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Multidimensional Data Sets Representing Different Customizing Strategies 105

Table A-19. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing Different Customizing Strategies.....106

Table A-20. Selected Percentiles of Estimated Ability Distributions Derived from Selected Multidimensional Data Sets Representing Tests of Different Length 107

Table A-21. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing Tests of Different Length.....107

Table A-22. Selected Percentiles of the Estimated Ability Distributions Derived from Multidimensional Data Sets Representing the Combined Effects of Customizing Strategies and Estimation Items 108

Table A-23. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing the Combined Effects of Customizing Strategies and Estimation Items 109

Table A-24. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Multidimensional Data Sets Representing Different Correlation Levels (Based on Unidimensional Intact Tests) 110

Table A-25. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing Different Correlation Levels (Based on Unidimensional Intact Tests).....111

Table A-26. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Multidimensional Data Sets Representing Different Correlation Levels (Based on Two-dimensional Intact Tests) 112

Table A-27. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing Different Correlation Levels (Based on Two-dimensional Intact Tests) 113

Table A-28. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Set Representing the Combined Effects of Correlations of Latent Abilities and Customizing Strategies 114

Table A-29. Comparing Ability Estimates and Percentile Estimates for the Combined Effects of Correlations of Latent Abilities and Customizing Strategies.....115

Table A-30. Selected Percentiles of Estimated Ability Distributions Derived from Data Sets Representing the Combined Effects of Correlations of Latent Abilities, Customizing Strategies and Estimation Items 116

Table A-31. Differences in Selected Percentiles Derived from Data Sets Representing the Combined Effects of Correlations of Latent Abilities, Customizing Strategies and Estimation Items 118

Table A-32. Comparing Ability and Percentile Estimates for the Combined Effects of Latent Ability Correlations, Customizing Strategies and Estimation Items....119

Table A-33. Selected Percentiles of Estimated Ability Distributions Derived from Data Sets Representing the Combined Effects of Correlations of Latent Abilities and Changing Test Length 120

Table A-34. Comparing Ability Estimates and Percentile Estimates for the Combined Effects of Correlations of Latent Abilities and Changing Test Length 121

Table A-35. Comparing the Effect of Using Different Items for Norms Estimation for Unidimensional and Multidimensional Data Sets.....122

Table A-36. Comparing Customizing Strategies for Unidimensional and Multidimensional Data Sets.....123

Table A-37. Comparing the Effect of Changing Test Length for Unidimensional and Multidimensional Data Sets 124

.....

CHAPTER I. INTRODUCTION

Assessing student achievement in K-12 settings has become increasingly complicated in recent years. Under the No Child Left Behind (NCLB) legislation, schools are required to provide various kinds of information from limited testing. A natural outgrowth of this environment has been a recent focus on the possibilities of customizing an off-the-shelf test. *Test customization* refers to the practice of customizing a commercially available off-the-shelf test to meet the increasing information demands of local educators. Depending on the nature of the test and the information in demand, an off-the-shelf test can be customized in various ways, including adding items measuring additional content domains to 1) a full-length intact test, 2) an intact test that is proportionally shortened in terms of content specifications, or 3) an intact test that is disproportionally shortened in terms of content specifications.

Compared with the possibility of administering several different tests, customizing an off-the-shelf test to satisfy multiple information needs has many practical advantages. Since the enactment of the NCLB legislation, the amount of testing at schools has increased dramatically, and the states alone are administering about 45 million tests each year (New York Times, March 18, 2006). Customizing an off-the-shelf test can reduce the costs of test development, satisfying the "desire on the part of consumers for more information from less testing time" (Ansley, Forsyth & Hoover, 1989, p.1). For local educators, test customization can save resources, instructional time, and administration efforts. This practice can also reduce the excessive test development pressure facing testing companies and help maintain quality control. When appropriately

developed and utilized, customized tests may be able to bring convenience to students, local educators, and testing companies.

Validity of Customized Norms

Among many issues related to test customization, the validity of customized norms has been widely discussed in the past twenty years. *Customized norms* refer to applying the normative information derived from an off-the-shelf test to a customized test through statistical calibrations (typically IRT calibrations). As a large-scale off-the-shelf test usually comes with national norms, valid customized norms may help local educators compare their students on a national basis. As the development of norms is expensive and time-consuming, deriving customized norms could be time efficient and cost effective if an intact test and a customized test will yield similar normative information.

Therefore, an essential question to be asked in this context is whether such normative scores are valid; that is, whether the normative information derived from an intact test and a customized test is essentially the same. According to the invariance properties of unidimensional item response theory (UIRT) calibrations, valid customized norms may be obtained when the UIRT assumptions are satisfied. Specifically, the invariance properties were described as person-free item calibration and item-free person measurement (Wright, 1968).

Person-free item calibration implies that item parameter estimates using different examinee samples differ only by sampling errors if the UIRT assumptions are satisfied. Thus, item parameter estimates obtained from a sample of examinees nationwide would be the same as those obtained from a sample of local examinees except for sampling errors.

Item free person measurement implies that the ability estimates based on different sets of items differ only by sampling errors if the IRT assumptions are satisfied. In customized testing situations, it may suggest that ability estimates obtained from an intact test and a customized test may be the same except for sampling errors. Therefore, it may not be necessary to differentiate ability estimates obtained from an intact test or a customized test.

Altogether, the invariance properties of UIRT calibrations imply that valid customized norms may be achieved if UIRT assumptions are met (Linn & Hambleton, 1991). However, these properties (person-free item calibration and item-free person measurement) are realized at the expense of strong statistical assumptions. In situations where the IRT assumptions are compromised, these properties may not hold and the validity of customized norms can be seriously threatened.

Many studies have investigated the validity of customized norms using simulated or real data sets. Although these studies differed from one another, they followed similar analysis procedures: 1) estimating examinee ability based on an intact test and related customized tests, 2) transforming the obtained ability estimates to the score scale of the intact test, 3) deriving “national norms” using the ability estimates obtained from the intact test, 4) deriving customized norms by comparing the ability estimates obtained from the customized tests to the “national norms”, and 5) investigating discrepancies between the obtained normative information.

The results of most studies showed systematic discrepancies between the customized norms and the normative information derived from the intact test (e.g., Allen, Ansley & Forsyth, 1987; Ansley, Forsyth & Hoover, 1989; Harris, 1987; Harris, 1988).

Such findings suggested that a variety of factors can contribute to discrepancies between the normative information derived from an intact test and a corresponding customized test. Therefore, investigating the impact of these factors is important to guard against distorted normative interpretations.

Factors Related to the Validity of Customized Norms

As shown in the previous studies, the validity of customized norms can be very complicated, and many factors could contribute to discrepancies between the normative information derived from an intact test and a corresponding customized test. To ensure valid normative interpretations, investigation of the impact of various factors on the validity of customized norms is important for the practitioners of customized testing.

Customizing Strategies

One important factor related to the validity of customized norms is the customizing strategy. When an off-the-shelf test is customized to meet local information needs, new items are usually added to assess content domains not covered by the original test. Usually, an off-the-shelf test is customized in one of three ways: 1) adding items to a full-length off-the-shelf test, 2) adding new items to an off-the-shelf test that is proportionally shortened in terms of content specifications, and 3) adding new items to an off-the-shelf test that is disproportionally shortened in terms of content specifications.

The first kind of customization may take place when clients have enough resources to administer a rather long test. In addition to a full-length off-the-shelf test, new items are included to measure content domains not covered by the original test. The latter two types of test customization may take place when the local testing resources are relatively limited. Although new items are added to measure content domains not covered by the original test, the resulting customized instrument is as long as the full-length off-the-shelf test. An intact test is, therefore, shortened according to the objectives of local

curricula. For the second kind of customization, an off-the-shelf test is proportionally shortened; that is, the proportional coverage of each content domain in the customized instrument remains the same as that of the off-the-shelf test. For the third kind of customization, an off-the-shelf test is disproportionally shortened; that is, the proportional coverage of each content domain of the customized instrument is adjusted based on objectives of local curricula. A choice between the latter two types of customizing strategies usually depends on the result of aligning the content specifications of an off-the-shelf test to the objectives of the local curricula.

The three types of customization strategies may represent a continuum in terms of the accuracy of the customized norms. For the first type of customization, a full-length off-the-shelf test is administered along with a set of new items measuring content domains not covered by the original test. The normative information derived from the resulting customized test may be very similar to that derived from the full-length off-the-shelf test. For the second type of customization, an off-the-shelf test is proportionally shortened. The standard error of measurement of the reduced set of intact items may become larger, which may contribute to larger discrepancies between the norms derived from the off-the-shelf test and the customized test. For the third type of customization, an off-the-shelf test is disproportionally shortened, which means that the reduced set of intact items may be different from the full-length off-the-shelf test in terms of dimensional structure. When customized norms are derived from the resulting instrument, even larger discrepancies between the norms may be observed.

Using Different Item Sets to Derive Customized Norms

As a typical customized test usually includes both intact items and items measuring content domains not covered by the original tests, complications can arise regarding deriving normative information using 1) only intact items or 2) all items in the customized instrument. Both procedures have been used in the literature.

For example, Harris (1990) constructed customized instruments by appending new items to the end of the full-length ACT test. The ability estimates used to derive normative information were obtained using 1) intact items in the customized instruments, or 2) all items in the customized instruments. As large discrepancies were observed for normative information derived from the customized and intact tests, the results of her study seemed to suggest that neither method would work reasonably well to provide valid customized norms.

Promising results, however, were observed in other studies on the same topic. For example, in Dungan's (1988) study, Grade 4 and 6 students took the complete Mathematics Tests (95 items) of the Metropolitan Achievement Test (MAT6) and a short 20-item locally constructed test. For each grade, there were five different locally constructed forms, each including 20 items that were administered to different groups of students together with the MAT6 test. The locally constructed items were calibrated to the MAT6 scale and substituted for the 20 easiest items in the MAT6 test. That is, the customized norm-referenced estimates were computed as if a student had taken 75 out of the 95 MAT6 items plus the 20 locally constructed items. The customized ability estimates were then compared to the scores obtained from the intact MAT6 tests. Very small differences between the estimates (relative to the standard error of measurement) were found in each case.

Based on the results of previous studies, it seems unclear which method would be more appropriate to derive customized norms. Note that when items measuring new content domains are added to the customized instrument, the dimensional structure of the resulting customized instrument may be different from that of the full-length off-the-shelf test. Therefore, the customized norms derived from the resulting instrument may be very different from that derived from the off-the-shelf test.

Matching Tests on Dimensionality

It is obvious that the above discussions were closely related to the concept of matching the intact and customized instruments in terms of dimensionality, which is an important condition related to the validity of the customized norms. This concept has been emphasized repeatedly by many authors (e.g., Linn & Hambleton, 1991; Hirsch & Keene, 1989).

Hirsch and Keene (1989) investigated the impact of matching tests on dimensionality on equating results using both simulated and real data. For the simulated data sets, they constructed two tests that each had two underlying dimensions. When both tests had similar structure, the unidimensional IRT equating worked well. When the structure of the two simulated tests differed substantially, large errors were observed in estimated norm-referenced achievement levels derived from the customized instrument. Such findings were also observed in the real data results, where they found that the adequacy of the equating of the tests was closely related to the comparability of the dimensional structure of the tests to be equated.

In customized testing literature, the goal of matching tests in terms of dimensionality is usually achieved by matching tests by content specifications. For example, in Linn & Hambleton's (1991) article, they commented that "the content specifications of customized tests should be similar to those of the intact tests to ensure the two tests are matched in terms of dimensionality." (p. 204). The significance of matching tests in terms of dimensionality was also proposed by other researchers in various studies (e.g., Yen, et al., 1987; Hirsch & Keene, 1989).

The impact of matching tests by content specifications on the validity of customized norms was investigated by several studies (e.g., see, Way, et al., 1989; Allen, et al. 1987; Ansley, et al., 1989). In the Way et al. (1989) study, customized instruments were constructed by deleting items from the intact tests in two ways: 1) proportionally deleting items based on the table of specifications, or 2) disproportionally deleting items.

For either type of customization, systematic differences were observed in ability estimates and the corresponding normative information derived from intact tests and customized instruments. Similar results were also observed in other studies (Allen, et al. 1987; Ansley, et al., 1989).

Test Length

Another factor that seems to directly affect the validity of customized norms is test length. Although some researchers found that test length may be of little effect (e.g., Qualls-Payne et al, 1989), the findings of two of Harris' (1988, 1990) studies seemed to suggest test length could be directly related to the validity of customized norms.

Harris (1988, 1990) investigated the effect of shortening or lengthening an intact test while maintaining the proportional coverage of the content specifications. She found that changing test length would lead to different customized norms. To investigate the effect of shortening an intact test, she (1988) reduced the 40-item ACT Mathematics Test to 10, 20, and 30 items, maintaining the proportional coverage of the content categories of the intact test. To investigate the effect of lengthening an intact test, she (1990) appended items to the end of the full-length ACT English, Mathematics, Reading, and Science Tests. In both situations, sizable differences were observed between the ability estimates based on the intact tests and the customized instruments. The same was observed for the normative information obtained from the related instruments. She in turn concluded that, "test length, in and of itself, is a potent enough factor to make comparisons between total intact tests and shortened customized tests unwise" (Harris, 1988, p. 14).

Correlation between Latent Abilities

Besides the above mentioned factors, another issue related to the validity of customized norms may be the degree of correlation between latent abilities. The impact of latent ability correlation on various applications was investigated by many authors in

the literature (e.g., Ackerman, 1989; Ansley & Forsyth, 1985; Zhao et al., 2001; Tong & Kolen, 2006). For example, in Ansley and Forsyth's (1985) study, the impact of correlation between latent abilities on UIRT ability estimates was investigated using simulated data. They found that the unidimensional ability estimates were correlated with the average of the ability parameters for each dimension. The higher the correlation of the latent abilities, the closer the unidimensional ability estimate was to the MIRT ability average. Zhao et al. (2001) investigated this issue in computer adaptive testing. They found that a decrease in inter-dimensional ability correlation was associated with increasingly undesirable evaluative measures of simulation results.

Above all, the findings of previous studies suggested that different degrees of correlation between latent abilities may affect the results of a variety of measurement applications. Similar to ability estimation methods, although this issue has been investigated for many other applications, little effort has been directed towards the validity of customized norms. How this factor will influence the validity of customized norms needs to be explored.

Preview of the Study

When an off-the-shelf norm referenced test is customized to meet the information demands of local clients, deriving norms based on customized tests may be useful to local educators. However, the results of previous research have shown that the validity of customized norms can be very complicated. Rather than providing practical convenience, customized norms may lead to misleading normative interpretations, which can invalidate legitimate use of such information in spite of the good intentions. Carefully investigating the impact of a combination of factors on the validity of customized norms is, therefore, especially important to guard against distorted normative interpretations.

Although previous studies have provided guidelines regarding the validity of customized norms, the previous findings were limited in several ways. First, the literature

provided little information about topics such as the impact of correlations of latent abilities on customized norms. As discussed above, the results of previous studies suggested that different correlations might lead to different results in various applications. Its impact on the validity of customized norms, however, is inadequately investigated in the literature.

Second, for the factors that have been relatively widely investigated, the findings of previous studies seem to yield different conclusions, including the impact of test length and matching tests in terms of content specifications. Replications will provide further insights regarding the impact of these factors on the validity of customized norms.

Third, a comprehensive evaluation of the impact of various factors on the validity of customized norms is largely lacking. The literature has shown that the validity of customized norms can be very complicated, and many factors could simultaneously affect the validity of customized norms. To administer a customized test, important decisions need to be made related to a variety of factors, including using different item sets to derive norms, choosing a customizing strategy, and choosing an UIRT ability estimation method. The lack of a comprehensive study of the various factors on the validity of customized norms may make it difficult to make such decisions. Although the findings of previous studies could provide some guidelines, a comprehensive investigation of these factors in one study could provide additional insights on this issue.

Starting from the above observations, the present study investigated the impact of a combination of factors on the validity of customized norms. The factors investigated in this study were 1) customizing strategies, 2) using different items to derive customized normative information (using only items from an intact test or all items in a customized instrument), 3) test length, 4) test dimensional structures (whether items from an intact test and new items measure the same latent ability/abilities), 5) correlations of latent abilities. As it is difficult to manipulate the different factors simultaneously using real

data, data sets were simulated to represent typical customizing situations characterized by these factors.

CHAPTER II. REVIEW OF LITERATURE

This chapter is divided into two parts. The first part discusses the results of previous studies related to analyzing multidimensional data using a unidimensional model. As suggested previously, the validity of customized norms based on UIRT calibrations is especially a concern when the unidimensionality assumption is violated. Therefore, investigating the impact of analyzing multidimensional data using a unidimensional model in the context of customized testing is particularly relevant to the present study. The second part discusses the findings of previous studies specifically related to the validity of customized norms.

Analyzing Multidimensional Data Using a Unidimensional Model

In this section, discussions related to analyzing multidimensional data using a unidimensional model are divided into three parts. As this type of research has largely depended on simulated data, the first part discusses simulating data based on IRT models. Because the true item and ability parameters are known, simulated data sets allow the researcher to make comparisons between the true parameters and their estimates. However, the artificial nature of the simulated data requires that researchers take care to simulate realistic data. The second part describes various applications related to analyzing multidimensional data using a unidimensional model. The relationship of the present study to the findings of previous studies is discussed in the third part.

Data Simulation Using IRT Models

Rather than using item responses from a real data set, simulated data were used in the present study. This choice is due to the practical constraints of real data analysis in this context; that is, it is difficult or even impossible to manipulate various factors related to customized testing using real data. As simulation conditions are specified by the

researcher, it is possible to investigate the impact of various factors on the validity of customized testing. However, the artificial nature of simulated data makes it necessary to justify the simulation procedure and the parameter values chosen for data simulation.

When unidimensional data sets are generated using a UIRT model, there has been little disagreement regarding the choice of a UIRT model; that is, the three-parameter logistic model (3-PL) is typically used in the literature. However, there has been less agreement when multidimensional data sets are generated. This section discusses simulating data using multidimensional item response theory (MIRT) models.

To generate multidimensional data sets, the first step is to choose an MIRT model. In the past twenty years, several MIRT models have been proposed. They can be classified into two categories: compensatory and noncompensatory models. A compensatory MIRT model assumes that high ability on one dimension can compensate for low ability on another dimension in terms of the estimated probability of a correct response. In comparison, a noncompensatory model limits the extent to which high ability on one dimension can compensate for the lack of ability on another dimension.

One frequently cited noncompensatory model was proposed by Sympson (1978). This model was an extension of the unidimensional three-parameter logistic model, and was specified as follows:

$$P_i(\Theta_j) = c_i + \frac{(1-c_i)}{\prod_{k=1}^{m}\{1+\exp[-1.7a_{ik}(\theta_{jk}-b_{ik})]\}}, \quad (1)$$

where

$P_i(\Theta_j)$ is the probability of a correct response to item *i* by examinee *j* whose location in the latent space is described by $\Theta_j = (\theta_{j1}, \theta_{j2}, ..., \theta_{jm})$,

θ_{jk} is the ability parameter for person *j* on dimension *k*,

a_{ik} is the discrimination parameter for item *i* on dimension *k*,

b_{ik} is the difficulty parameter for item *i* on dimension *k*, and

c_i is the guessing parameter for item *i*.

Several compensatory models were also proposed in the literature. A compensatory multidimensional extension of the 3-PL model was described by several researchers (Hattie, 1981; McKinley & Reckase, 1983; Doody-Bogan & Yen, 1983). The model proposed by Doody-Bogan and Yen (1983) was as follows:

$$P_i(\Theta_j) = c_i + \frac{1 - c_i}{1 + \exp[-1.7(\sum_{k=1}^{m} a_{ik}(\theta_{jk} - b_{ik}))]}, \quad (2)$$

where

all parameters are defined previously.

Reckase modified this model by including a scalar difficulty term rather than a specific difficulty parameter on each dimension. This model was defined as

$$P_i(\Theta_j) = c_i + \frac{1 - c_i}{1 + \exp(-\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i)}, \quad (3)$$

where

the constant 1.7 is omitted,

d_i is a scalar difficulty parameter ($d_i = \sum_{k=1}^{m} a_{ik} b_{ik}$), and

all other parameters are defined previously.

The reader is referred to Reckase and McKinley (1985) for a comprehensive review of the models proposed in the literature.

The difference between compensatory and noncompensatory models can be found in their denominators. In noncompensatory models, the denominator is the product of the denominators for each dimension, while in compensatory models only the exponential terms in the denominators are multiplied. In noncompensatory models, the least probability of a correct response represents an upper bound on the probability of a correct response pooled across all dimensions. In compensatory models, the effects of the different dimensions are additive.

Whether a compensatory or noncompensatory model better represents the data in a typical testing situation is unclear. For example, Ansley and Forsyth (1985) argued that noncompensatory models had greater intuitive appeal when more than one latent ability was required to answer an item correctly. Wang (1985), however, considered Sympson's model to be excessively restrictive as the probability of a correct response to an item was bounded from above by the smallest marginal probability.

It is possible that noncompensatory and compensatory models may represent data differently depending on the nature of the test. For example, for math problem solving items, both reading and math abilities are required to answer an item correctly. To solve such a problem, it is reasonable to assume that high ability in reading may not compensate for low ability in math. This assumption, however, may not apply for reading items that are heavily loaded on content knowledge. For example, for a reading item related to a biology article, the knowledge of the biological topic may increase an examinee's chance of a correct response; that is, the examinee may be able to answer the item correctly even though his / her reading ability is relatively low.

Due to the availability of estimation software, however, MIRT compensatory models have been used more often in the literature. Based on this observation and the

above discussion of compensatory and noncompensatory MIRT models, a compensatory MIRT model was chosen to simulate MIRT data in the present study.

To make the generated data sets realistic, the results of item parameter estimation based on real data are often used as item parameters for simulation. Although different procedures have been used in the literature, a common procedure is using a combination of the descriptive statistics of discrimination and difficulty parameter estimates and the exact values of the guessing parameter estimates. For example, Tong & Kolen (2006) estimated UIRT item parameters using the 3-PL model based on real data. The item parameters used to generate multidimensional data included the mean and standard deviation of the item discrimination and difficulty estimates, and the exact values of the guessing parameter estimates. The generated discrimination and difficulty parameters on different dimensions were assumed to be independent. A similar simulation procedure was used in this study to generate multidimensional data sets.

Previous Findings and Applications

When tests are calibrated and scored using a UIRT model, it is assumed that only one latent ability is required to answer items correctly. This assumption, however, may be violated. This section describes various applications related to analyzing multidimensional data using a unidimensional model.

Impact on Parameter Estimation

Several studies examined the consequences of analyzing multidimensional data using a unidimensional model using simulated data sets (e.g., Ansley & Forsyth, 1985; Yen, 1985; Doody, 1985; Way, Ansley, & Forsyth, 1988; Ackerman, 1989). For example, Ansley and Forsyth (1985) investigated the nature of unidimensional parameter estimates using two-dimensional data sets generated using Sympson's noncompensatory model. Their results demonstrated that violations of the assumption of unidimensionality did have an effect on parameter estimates. They found that a) the $\hat{a}$ values estimated

from a unidimensional IRT model were best considered as the average of the true a_1 and a_2 values in a two-dimensional latent space; b) the $\hat{b}$ values seemed best thought of as overestimates of the true b_1 values; and c) $\hat{\theta}$ seemed to be highly correlated with the average of the true θ values.

An approximate relationship between unidimensional logistic parameter estimates and true multidimensional parameters was derived by Yen (1985) using a least square approach. The derived relationship showed that $\hat{\theta}$ estimated from a unidimensional model could be considered as a weighted combination of the underlying traits. The weights were proportional to the trait's importance as determined by the item discrimination on a particular dimension.

The analytic developments by Wang (1985) concurred with Yen's finding. She concluded that,

> "the unidimensional estimate is in fact a linear combination of the underlying latent variables. However, depending on the influence of each trait on the items (as implied in the item discrimination parameters), it may appear to estimate the dominant dimension, the average of the dimensions, or any weighted average of the dimensions. The particular composite obtained for a given data set is determined by the structure of the matrix of "true" item parameters (p. 31).

Generally speaking, the findings of these studies suggested that multidimensionality did affect UIRT parameter estimates. More discussions on this topic regarding the impact of θ-correlation on UIRT ability estimates are given later.

Applications Related to Analyzing Multidimensional Data using a Unidimensional Model

Previous studies have investigated several applications from the perspective of analyzing multidimensional data using a unidimensional model, including equating, scaling and computer adaptive testing. Although the purposes of those studies were different from that of the present study, they provide useful information in terms of 1) how the concept of analyzing multidimensional data using unidimensional models was

approached in various measurement applications, and 2) how the present study was related to findings of the previous studies.

Hirsch and Keene (1989) investigated the impact of analyzing multidimensional data using a unidimensional model in an equating setting. They constructed two tests (referred to as an intact test and a customized test) that each had two underlying dimensions. Based on the simulated data, they found that unidimensional IRT equating worked well when both tests had similar dimensional structure. However, large errors in estimated norm-referenced achievement levels derived from the customized instruments were identified when the dimensional structures of the intact and customized instruments differed substantially. Their findings suggested that the adequacy of the equating was closely related to the comparability of the dimensional structure of the tests to be equated.

Several studies investigated the consequence of analyzing multidimensional data using a unidimensional model in the context of vertical scaling (e.g., Yen, 1986; Camilli, 1998; Tong and Kolen, 2006). For example, Yen (1986) noted that when the 3-PL model was used for tests varying widely in difficulty across grades, the variance of the ability scale usually decreased over grades as the tests became more difficult. She suggested that scale shrinkage could be a result of the dimensionality change across grades. Similar ideas were proposed by Camilli (1998). He suggested that the between-grade shrinkage could be a result of analyzing multidimensional data using a unidimensional model.

The impact of analyzing multidimensional data using a unidimensional model was also investigated in the context of computer adaptive testing (De Ayala, 1992; Folk & Green, 1989; Weiss & Suhadolnik, 1982; Zhao, et. al., 2001). For example, Zhao et. al. (2001) investigated the impact of the correlation between latent abilities and the correlation between MIRT difficulty parameters on parameter estimates. They found that the correlation between MIRT difficulties seemed to have little effect on ability estimates. The decrease in the inter-dimensional ability correlation was associated with increasingly undesirable evaluative measures of the simulation results. Consistent with

previous literature, their findings suggested that caution should be exercised in unidimensional CAT applications when the unidimensionality assumption was likely to be violated.

Implications for the Present Study

In summary, the findings of previous studies regarding analyzing multidimensional data using a unidimensional model provided two useful pieces of information. First, the findings of these studies suggested that unidimensional estimates were related to two or more latent dimensions when a data set was multidimensional. This may be of particular concern in investigating the validity of customized norms.

Second, unidimensional IRT calibrations may be useful even though the unidimensionality assumption is not satisfied. This observation is identified for various applications such as equating and scaling. As deriving customized norms is the application of interest for the present study, whether UIRT calibrations will be useful for this application needs to be further explored.

Customized Tests and Customized Norms

In this section, the discussions of previous studies related to the validity of customized norms are divided into seven parts. The first six parts discuss relevant factors as identified in the literature, including: 1) customizing strategies, 2) matching tests on dimensionality, 3) test length, 4) using different item sets to derive customized norms, 5) UIRT ability estimation methods, and 6) correlation between latent abilities. The relationship of the present study to the findings of previous studies is discussed in the last part.

Test Customizing Approaches

In the current testing environment, four commonly used customizing strategies are 1) norm-referenced test (NRT) only (NRT-only), 2) NRT-based, 3) criterion-

referenced test (CRT) based (CRT-based), and 4) CRT-only (Linn & Hambleton, 1991; Wilson & Hiscos, 1984).

If the customization takes place in the NRT-only approach, a full-length off-the-shelf test is administered. The customization only takes place in reporting the scores related to local objectives. That is, the score reports are constructed based on the clusters of test items that correspond to local objectives. The advantage of the NRT-only approach is that the intact test is not changed. The disadvantage of this approach is that it is impossible to assess the standards in the local curriculum not covered by the intact test. Even for the objectives covered in the intact test, the precision of the information totally depends on the degree to which these objectives are emphasized in the intact tests, which may or may not match the demands as specified by the local curriculum.

For the second type of customization, the NRT-based approach, a full-length off-the-shelf test is administered. In addition, new items are added to cover topics sparsely covered or not covered by the intact tests. Usually the added items are not used to derive the norm-referenced scores. The customization takes place in test construction in that new items are added to assess additional content domains that are not covered or not adequately covered by the off-the-shelf test. Compared with the NRT-only approach, the NRT-based approach provides a means to respond to the issue of missing topics or mismatch of emphases on the intact test. The customized instrument, however, is longer than the off-the-shelf test. As all items in the original test and new items are included in a single administration, the customized instrument might be too long in some circumstances.

The third type of customizing strategy, the CRT-based approach, is similar to the NRT-based approach in that both intact items and items measuring content domains not covered by the original test are included in a customized test. Instead of administering all of the intact items, only a reduced set of intact items are included in this type of customized instrument. The normative information can be obtained from the reduced set

of intact items or from a combination of intact items and new items in the customized instrument. Compared with the first two approaches (NRT-only and NRT-based), the CRT-based customization emphasizes local objectives rather than normative comparisons. The customization takes place in test construction where items are selected based on the objectives of local curricula. The resulting customized instrument is usually more able to evaluate the local curricula than the first two approaches. However, the precision of the normative information derived from this type of customization may be threatened as only a reduced set of intact items are used to derive normative information.

For the fourth type of customization, the CRT-only model, a customized test includes only items assessing local objectives. The norm-referenced scores are obtained without actually administering any intact items, and equating or linking is used to derive norms for customized test. A variety of equating designs can be used to obtain the normative information. Two commonly used designs are: 1) common examinee design in which both tests to be equated are administered to the same group of students, and 2) random group design in which two randomly equivalent groups may be formed and one of the two tests is administered to each group. Both classical equating and IRT procedures can be used to provide the basis for generating NRT scale score estimates that can be converted to various types of norm-referenced scores.

The four customizing models form a continuum, which represents different compromises between the competing requirements for norm-referenced and curriculum-specific information. At the NRT end of the continuum, the test is built for norm-referenced interpretations from which some curriculum-specific information can be reported. The normative information derived from this type of customization is complete, but the information about specific curriculum objectives might be incomplete and less than ideal. At the CRT end of the continuum, the test is built to provide curriculum-based information from which norm-referenced scores may be inferred. The objective

information might be valid and comprehensive, but the normative information is apt to be less accurate.

Due to the above considerations, this study investigates the impacts of the two less extreme conditions in the four models: the NRT-based and CRT-based approaches. The two models represent feasible customizing strategies to meet the dual information needs of assessing local curricula and providing normative information. The resulting customized instruments include both intact items and new items assessing content domains not covered by the intact test.

Depending on the adequacy of local resources, clients can administer a) a full-length intact test or b) a shortened intact test plus items measuring new content domains. When the intact test is shortened, it can be shortened proportionally or disproportionally in terms of content specifications. Therefore, three customizing strategies are investigated in this study: 1) adding items to a full-length intact test, 2) adding items to an intact test that is shortened proportionally in terms of content specifications, and 3) adding items to an intact test that is shortened disproportionally in terms of content specifications. The first customizing strategy relates to the NRT-based approach, and the last two strategies relate to the CRT-based approach as discussed in this section.

Matching Tests on Dimensionality

Neither typical NRTs nor typical CRTs tend to be unidimensional (see, e.g., Hambleton, et al., 1989; Linn, 1990; Linn & Hambleton, 1991; Yen, Green, & Burket, 1987). However, this does not mean that unidimensional IRT calibrations are useless for psychometric applications such as scaling, equating and linking. In the context of customized testing, matching tests in terms of dimensionality is one essential condition related to valid customized norms (see, e.g., Linn, 1990; Linn & Hambleton, 1991; Yen, Green, & Burket, 1987). As Yen, Green, and Burket (1987) stated, “Multidimensionality

does not preclude the use of a unidimensional procedure ... However, it is essential that the tests be matched for multidimensionality" (p. 11).

Hirsch and Keene (1989) used an equating method to investigate the effects of matching tests in terms in dimensionality. They simulated two data sets to represent an intact test and a customized instrument that each had two underlying dimensions. When both tests had similar dimensional structures, the unidimensional IRT equating worked well. When the structures of the two simulated tests differed substantially, large errors were observed in estimated norm-referenced achievement levels derived from the customized instrument. Similar findings were also observed based on the real data sets. That is, the adequacy of the equating of the tests was closely related to the comparability of the dimensional structures of the tests to be equated.

The notion of matching tests in terms of dimensionality is closely related to the suggestions of other researchers (e.g., Linn & Hambleton, 1991; Holmes, 1986; Lenke, 1989). As Linn & Hambleton (1991) commented, "(the) content coverage of a customized test needs to be carefully matched to the content of the NRT to which it is being equated. Customized norms are apt to be distorted when a content category is disproportionally represented on the customized test" (p. 196).

In the context of customized testing, matching tests in terms of dimensionality is usually achieved in terms of content specifications. It is assumed that the intact and customized tests will have similar dimensional structure if they are matched in terms of content specifications. When the validity of customized norms is of interest, "it is important to assure that the customized test and the NRT have proportional coverage of the content categories" (Linn & Hambleton, 1991, p. 196).

Several studies have addressed this issue by comparing the normative information derived from a full-length intact test and that derived from a subset of intact items. Harris (1987), for example, constructed three customized subtests by selecting items from either three or four of the six content categories of the ACT Mathematics Test. In general, there

was relatively poor agreement in estimated scores obtained from the customized tests and the full-length intact test.

Way, Forsyth & Ansley (1989) investigated the effect of content match by selecting items from four different subtests of the Iowa Tests of Basic Skills (ITBS) (Hieronymous, Hoover & Lindquist, 1985). Two types of customized instruments were constructed by deleting items from the intact tests. To construct a content customized test, items that did not match the local objectives were deleted. Therefore, the content coverage of the customized instrument was different from that of the intact test. To construct a representative customized test, items were deleted proportionally so that the remaining items formed a representative sample of the content domains of the intact tests. For each examinee, ability estimates were obtained from the intact test, the content customized test, and the representative customized test. Customized norms were obtained by comparing the ability estimates obtained from the customized instruments to the percentile rank table constructed from the intact tests. For the four ITBS subtests, the normative information derived from the representative customized tests showed systematically higher degrees of agreement with the actual normative information than the norms derived from the content customized tests. Their findings also indicated that the customized test and intact test should have proportional coverage of the content categories.

Test Length

Of the many factors identified in the literature, test length seems to directly affect the validity of customized norms. Several studies have investigated the impact of changing test length on the normative information derived from an intact test or corresponding customized instruments.

Harris (1988) investigated the impact of shortening a test on the validity of the customized norms. The instrument used in her study was the 40-item ACT Mathematics

Test. The customized instruments were constructed by shortening the tests to 10, 20, and 30 items, maintaining the proportional coverage of the content categories of the intact test to the extent possible. Both real and contrived data sets (where examinees were administered the intact test, and the item responses of selected items were deleted to represent the customization), as well as simulated unidimensional and multidimensional data sets were used in her study. As sizable differences were observed between the ability estimates based on the intact tests and the reduced-length instruments, she concluded that, "test length, in and of itself, is a potent enough factor to make comparisons between total intact tests and shortened customized tests unwise" (p. 14).

Harris (1990) also investigated the impact of lengthening the test on the validity of the customized norms. The instrument used in that study included the ACT English, Mathematics, Reading, and Science Tests. The customized instruments were constructed by appending items to the end of the intact item sections to minimize the context effects. Sizable differences were again identified between the ability estimates based on the intact and the lengthened instruments. She, therefore, concluded that it was difficult to achieve valid customized norms based on "a shortened, lengthened, or otherwise modified version of the intact test" (p. 15).

In spite of the negative results as shown in Harris's studies, positive findings were observed in other studies regarding the effect of test length. For example, Qualls-Payne et al (1989) used short versions of one form of a norm-referenced test (Form B) to estimate the proportion correct scores for an alternate form of the test (Form A). Three shortened forms containing 10, 20, and 30 items from Form B were constructed, providing proportional content coverage and average item difficulties that were approximately equal to those of Form A. The national proportion correct scores were estimated from scaling Form B items together with Form A items. The scores obtained from the shortened forms were compared to the proportion correct scores obtained from Form A. The findings suggested that very good estimates of the proportion correct scores

could be obtained using the alternate form of the test (for even the shortest instruments) using IRT scaling methods.

Using Different Item Sets to Derive Customized Norms

When a customized instrument includes both intact items and items measuring content domains not covered by the original test, examinee ability could be estimated using: 1) intact items, or 2) a combination of intact items and new items. In previous literature both procedures have been used. There seems to be inadequate information to evaluate which procedure will provide more valid normative information in this context.

Deriving normative information using entirely intact items has been frequently used in research related to customizing tests by the NRT-based approach; that is, new items are added to a full-length intact test. For this type of customization, new items are usually contained in a separate booklet to assess content domains not covered by the original test. Although the added items could be used for other purposes such as score reporting, they are usually not used to derive norm-referenced scores (Linn & Hambleton, 1991).

Deriving normative information using a combination of intact items and new items has also been used in the literature (e.g., Harris, 1989; Harris, 1990; Green, 1987; Dungan, 1988). For example, Dungan (1988) used a combination of intact items and new items to derive customized norms. In that study, Grade 4 and 6 students took the complete Mathematics Tests (95 items) of the Metropolitan Achievement Test (6th Edition) (MAT6) and a short 20-item CRT test. For each grade, there were five different CRT forms, each including 20 items. The CRT forms were administered to different groups of students together with the MAT6 test. The CRT items were calibrated to the MAT6 scale and substituted for the 20 easiest items in the MAT6 test. That is, the customized norm-referenced estimates were computed as if a student had taken 75 out of the 95 MAT6 items plus the 20 CRT items. The customized estimates were then

compared to the scores obtained from the MAT6 tests. The results showed very small differences between the estimates in each case in comparison with the standard error of measurement for the scale score.

Correlations of Latent Abilities

Another factor that may be related to the validity of customized norms is the correlation between latent abilities (θ-correlation). Several studies have investigated the impact of the correlation between latent abilities on various applications, including UIRT ability estimation, vertical scaling, and computer adaptive testing.

Ansley and Forsyth (1985) investigated the impact of correlation between latent abilities on UIRT ability estimates. They generated data using the two-dimensional version of Sympson's (1978) noncompensatory model. They found that $\hat{\theta}$ seemed to be highly correlated with the average of the true θ values.

> "It is almost always true that as the vectors of θs became more highly correlated, the values of statistics derived from the two-dimensional data sets approached the values of statistics derived from the unidimensional data sets" (p. 47).

The unidimensional parameter estimates varied as a result of the correlation between θs. Clear disparities were found even if the abilities were highly correlated. Their findings were confirmed by Way, Ansley, and Forsyth (1988) who generated data using both Sympson's (1978) and Doody-Bogan and Yen's (1983) models. Similar findings were also reported by Ackerman (1989) who generated data using McKinley and Reckase's (1982) model.

Tong & Kolen (2006) investigated the impact of θ-correlation on vertical scales using simulated data sets. They generated data sets using a compensatory model with differing degrees of correlations: .1, .3, .5, .7, .9 and 1. They found that different degrees of θ-correlation did have an impact on the resulting vertical scales. Given the way the data sets were generated, the multidimensional data results with the θ-correlation of 1 were used as the criteria. For mean estimates across grades, they found that the smaller

the θ-correlation, the more deviate the mean estimates were from the criterion. The standard deviation estimates were relatively similar to one another across various conditions of θ-correlation. For the effect size estimate, they found that the smaller the θ-correlation, the more deviate the estimates were from the criterion.

Several studies investigated the impact of θ-correlation in computer adaptive settings. For example, Lau (1996) investigated the effect θ-correlation in a computer adaptive licensure test setting. He found that θ-correlation had little effect on average error rate (Type I and Type II error), but was negatively related to the number of items used to make a mastery decision. The larger the correlation, the fewer items would be used to make a decision. Zhao et al. (2001) also investigated the impact of θ-correlation on computer adaptive testing. They found that a decrease in inter-dimensional ability correlation was associated with increasingly undesirable evaluative measures of simulation results.

Above all, the findings of previous studies seemed to suggest that different degrees of θ-correlation could affect the results of various applications. Similar to the research about UIRT ability estimation methods, although this issue has been investigated for different applications, whether such findings can be generalized to the validity of customized norms is not clear. Investigations specifically addressing this topic will provide more insight about this issue.

Implications for the Present Study

As suggested by findings of previous studies, the validity of customized norms could be very complicated. Although valid customized norms may be achieved through appropriate calibrations, this expectation can not always be realistic given the strong UIRT statistical assumptions. The findings of previous studies have provided important guidelines regarding the validity of customized norms. However, they were limited in the following aspects.

First, previous studies provided little information about the impact of several factors, including: 1) correlations between latent abilities, and 2) customizing strategies. As for the first factor, although previous studies focusing on other applications (e.g., equating and scaling) suggested that they can lead to different results, their impact on the validity of customized norms was not adequately investigated. For the second factor, the effect of different customizing strategies, the procedure commonly used in previous studies was modifying an intact test based on one or at most two strategies. For example, in the studies by Allen, et al. (1987), Ansley, et al. (1989), and Way, et al. (1989), the customized instruments were constructed by deleting items from an intact test. In the Forsyth et al. (1992) study, the customized instruments were constructed by a) deleting intact items and b) adding new items to the intact test. A systematic investigation of the three customizing strategies as mentioned previously has not been conducted in the past.

Second, for the factors that have been relatively widely investigated, the findings of previous studies yielded different conclusions regarding their impact on customized norms. This is observed for the factors of test length and matching tests in terms of content specifications. For example, as for matching tests, both positive and negative findings were observed in the literature. Hirsch & Keene (1989) and Qualls-Payne et al. (1989) identified small differences (relative to the standard error of measurement) between the normative information derived from intact tests and that from customized tests. That seems to suggest that valid customized norms could be obtained if the intact and customized instruments have similar structures. On the other hand, Allen, et al. (1987), Ansley, et al. (1989), and Way, et al. (1989) observed sizable discrepancies in the normative information derived from an intact test and that derived from a customized instrument proportionally representing the intact test. Similar observations were also identified for the effect of test length on customized norms. Due to the uncertainty of the previous findings, replications will provide further insights on these issues.

Third, although previous studies have investigated one or two factors related to customized tests, investigating the impact of various factors simultaneously in one single study has not been done. As previously mentioned, customizing an off-the-shelf test can be very complicated. The customization process often requires clients to make decisions about a variety of factors. Investigating the impact of various factors simultaneously could provide further insights for clients in those circumstances.

The present study investigates the impact of a combination of factors on the validity of customized norms. Specifically, the factors investigated in this study include 1) customizing strategies, 2) length of a customized test, 3) items used for ability estimation (intact items or all items in a customized instrument), 4) correlation between latent abilities, and 5) dimensionality of intact items and added items. Although many studies have investigated the effects of one or two of the factors listed above, none investigated them simultaneously in one single study. As customizing an off-the-shelf test is complicated and many factors need to be considered at the same time, the findings of the present study may provide practitioners additional information to assist in evaluating customized norms. The details of the methods used to carry out the investigation are described in the next chapter.

CHAPTER III. METHODS

This chapter is divided into five parts. The first part defines the research questions addressed in this study. The second part describes the data sets generated to represent intact tests and corresponding customized instruments. The third part describes in detail the data generation procedures used in this study. A summary of the research design is given in the fourth part. The last part describes the data analysis procedures, including a brief description of the computer programs used in this study.

Research Questions

This study investigated the impact of a combination of factors on the validity of customized norms. As unidimensional IRT calibrations are typically used in the current testing environment, in this study the normative information is based on unidimensional proficiency estimates obtained under a variety of conditions.

The factors investigated in the present study included: 1) customizing strategies, 2) items used for ability estimation (intact items or all items in the customized instruments), 3) test length, 4) correlation between latent abilities, and 5) dimensionality of intact tests and added items. Specifically, the study addresses the following research questions:

1. Will customizing strategies (adding items to full-length intact tests, proportionally shortened tests and disproportionally shortened tests) affect the validity of customized norms?
2. Will the choice of different item sets used for ability estimation (items from an intact test or all items in a customized instrument) affect the validity of customized norms?
3. Will test length affect the validity of customized norms (based on 20, 40, or 60 items)?

4. If items from intact tests and added items measure different latent abilities, will different degrees of correlation between latent abilities (.00, .30, .60 and .90) affect the validity of customized norms?
5. If the items from intact tests and items added in customization measure different latent abilities, will the dimensionality of the intact test (unidimensional or two-dimensional) affect the validity of customized norms?
6. Will various combinations of the five factors affect the validity of customized norms?

Instruments

In this study data sets were simulated to represent both intact tests and customized tests. This section describes the generation of "intact test" data sets and how they were modified to represent typical customized testing situations. Descriptions of the "intact" data sets, "customized" data set IDs, customizing strategies, and dimensional structures of "customized" data sets are listed in Table 3-1.

Intact Tests

One data set representing a unidimensional intact test and one data set representing a two-dimensional intact test, each consisting of 40 items, were generated in this study. The unidimensional "intact test" data set included 40 items measuring the same dimension. The two-dimensional "intact test" data set included 20 items measuring Dimension I and 20 items measuring Dimension II.

Table 3-1. Descriptions of the Data Sets Representing Customized Tests

Intact Tests	"Customized" Data Set ID	Customizing Strategy	Number of Items in each Category				
Uni-dimensional intact test	U60U	Customizing Strategy I (control case): Adding 20 items to the full-length intact test (both intact items and added items measuring the same latent ability)	Total: 60	Intact:	40	(DI)	60
				New:	20	(DII)	0
	U60M	Customizing Strategy I: Adding 20 items to the full-length intact test (intact items and added items measuring different latent abilities)	Total: 60	Intact:	40	(DI)	40
				New:	20	(DII)	20
	U40U	Customizing Strategy II&III (control case): Adding 20 items to a shortened intact test (both intact items and added items measuring the same latent ability)	Total: 40	Intact:	20	(DI)	40
				New:	20	(DII)	0
	U40M	Customizing Strategy II: Adding 20 items to a shortened intact test (intact items and added items measuring different latent abilities)	Total: 40	Intact:	20	(DI)	20
				New:	20	(DII)	20
Two-dimensional intact test	M60M	Customizing Strategy I: Adding 20 items to the full-length intact test (20 intact items measuring Dimension I, 20 intact items measuring Dimension II, and added 20 items measuring Dimension III)	Total: 60	Intact:	40	(DI)	20
						(DII)	20
				New:	20	(DIII)	20
	M40M_PROP	Customizing Strategy II: Adding 20 items to the intact test that was proportionally shortened in terms of dimensionality (Deleting 10 intact items measuring Dimension I, deleting 10 intact items measuring Dimension II, and adding 20 items measuring Dimension III)	Total: 40	Intact:	20	(DI)	10
						(DII)	10
				New:	20	(DIII)	20
	M40M_DISP	Customizing Strategy III: Adding 20 items to the intact test that was disproportionally shortened in terms of dimensionality (Deleting 5 intact items measuring Dimension I, deleting 15 intact items measuring Dimension II, and adding 20 items measuring Dimension III)	Total: 40	Intact:	20	(DI)	15
						(DII)	5
				New:	20	(DIII)	20

Customized Instruments

When tests are customized to meet the information needs of local clients, items are usually added to measure content domains not covered by the original test. As noted earlier, common customizing strategies include: (a) adding items to a full-length intact test, (b) adding items to an intact test that is shortened proportionally in terms of content specifications, and (c) adding items to an intact test that is shortened disproportionally in terms of content specifications.

For the first customizing situation, the content coverage of the intact test is relevant and appropriate. With adequate resources to administer a rather long instrument, it might be reasonable to administer the full-length intact test plus new items measuring a content area not covered by the intact test. The resulting customized instrument includes more items than the original test.

Three data sets, **U60U, U60M**, and **M60M**represent the first situation where 20 items measuring a new dimension were added to the full-length intact test. The **U60U** data set, which included all items (intact items and added items) measuring the same dimension, was created as a frame of reference for the first type of customization. For the **U60M** data set, 20 items measuring a new dimension not covered by the original test were added to the unidimensional intact test. For the **M60M** data set, 20 items were added to the two-dimensional intact test to measure a new dimension not covered by the original test. The resulting data sets had 20 more items than the original “intact test” data set.

For the second type of customization, an intact test was shortened proportionally based on content specifications. New items were added to measure a new content domain not covered by the original test. The final customized instrument was the same length as the intact test.

The data sets, **U40M** and **M40M_PROP,** represent this type of customization. The data set **M40M_PROP** was constructed based on the two-dimensional intact test.

The intact items were removed in such a manner that the dimensionality structure of the shortened test was proportionally the same as that of the intact test. Specifically, two sets of 10 items (10 item measuring Dimension I, 10 items measuring Dimension II) were removed from the intact test, and 20 new items were added to measure a dimension not covered by the intact test. The final data set included 40 items, which was the same length as the intact test. The data set **U40M,** which was constructed based on the unidimensional intact test, is a simplified version of **M40M_PROP**. As intact items all measured the same dimension, the dimensional structure of the shortened intact test was the same as that of the original test. Specifically, 20 items were removed from the intact test and 20 items were added to measure a new dimension not covered by the original test. The resulting data set included 40 items, which is the same length as the "intact test" data set.

For the third type of customization, the intact tests were shortened disproportionally with respect to content domains. This represents situations in which clients evaluate the content coverage of the intact test to be relevant, but they want the proportional coverage of each content domain to be adjusted according to local curricula. Additionally, they are interested in assessing a new content area not covered by the original tests.

The data set **M40M_DISP** represents this situation. The proportional content coverage of the intact test was adjusted and items were added to measure a new dimension. Specifically, 5 items measuring Dimension I and 15 items measuring Dimension II were removed and 20 items measuring a new dimension were added to the instrument. The resulting data set included 40 items, which is the same length as the intact tests.

The latter two types of customization can be described briefly as adding items to a shortened intact test. They are similar to the first type of customization in that new items are added to measure a content domain not covered by the original test. Different from

the first strategy, however, the resulting instruments are the same length as the intact test. The data set **U40U,** which included items (intact items and new items) measuring the same latent ability, was constructed as a frame of reference for this situation.

Data Set Notations

The previous section described how data sets were constructed to represent intact tests and customized tests using different customizing strategies. This section describes the notation used to distinguish the various versions of these data sets. To answer the research questions, a variety of versions of these data sets were required. Generally speaking, the intact tests were represented by U_Intact or M_Intact depending on the dimensionality of the intact tests. The customized data sets were represented by notations related to the investigated factors, including customizing strategies, estimating items and correlations of latent abilities.

To represent individual data sets, a 'three-part' notation scheme was used that included customizing strategy, estimation items and correlations of latent abilities. (1) The first part of the notation represented the customizing strategy, which was made up of: a) the dimensionality of the intact test based on which the customized data set was constructed, b) the number of items in the customized test, and c) the dimensionality of customized test. Take U60U for example. The letter 'U' in the beginning suggested it came from a unidimensional intact test; the number '60' suggested 60 items were included in the customized test; and the last letter 'U' suggested that the customized data set was also unidimensional. (2) The second part of the individual data set notation scheme was related to the item sets used to estimate normative information: only items from intact tests or all items in customized tests (including both items from intact tests and new items). If only items from intact tests were used in estimation, the suffix _Part was used; if all items were used, the suffix _All was used. (3) The third part of the individual data set notation scheme was related to the correlations of latent abilities

measured by the items from intact tests and new items. The suffixes _.0, _.3, _.6, _.9 were used to represent the different correlations. Note that the correlations of latent abilities were irrelevant to unidimensional customized data sets, and the _.0, _.3, _.6, _.9 suffixes were only used for multidimensional customized tests (i.e., new items measuring a different latent ability). Take U60M_All_.0 for an example, which represented the customized data set constructed by adding 20 new items measuring a different latent ability to a unidimensional intact test when the correlation of the latent abilities was .0.

Table 3-2. Notations of Averages and Individual Data Sets

Test Dimensionality	Level 1 (avg. over correlations and estimating items) Customizing Strategy	Level 2 (avg. over correlations) Customizing Strategy X Estimating Items	Level 3 (individual customized data sets) Customizing Strategy X Estimating Items X Correlations
U_Intact →1-D cust. test (Unidimensionality)	U60U	U60U_Part U60U_All	
	U40U	U40U_Part U40U_All	
U_Intact →2-D cust. test	U60M	U60M_Part U60M_All	U60M_Part_.0; _.3;_.6; _.9 U60M_All_.0; _.3; _.6; _.9
	U40M	U40M_Part U40M_All	U40M_Part_.0; _.3; _.6; _.9 U40M_All_.0; _.3; _.6; _.9
M_Intact →3-D cust. test	M60M	M60M_Part M60M_All	M60M_Part _.0; _.3; _.6; _.9 M60M_All_.0; _.3; _.6; _.9
	M40M_Prop	M40M_Prop_Part M40M_Prop_All	M40M_Prop_Part_.0; _.3; _.6; _.9 M40M_Prop_All_.0; _.3; _.6; _.9
	M40M_Disp	M40M_Disp_Part M40M_Disp_All	M40M_Disp_Part_.0; _.3; _.6; _.9 M40M_Disp_All_.0; _.3; _.6; _.9

To represent the average of a set of individual data sets, which was usually related to a main effect, the notation included fewer parts than that of an individual data set. Take U60M_All for an example, which represented an average over different correlations.

Take U60M for another example, which represented an average over both correlations and estimating items. Table 3-2 lists the notations for averages and individual data sets representing intact and customized tests. The Level 3 notations represented individual data sets. Level 1and Level 2 notations represented averages of data sets.

The above discussions described how to represent individual data sets or averages over a set of data sets. To answer the research questions of this study, a variety of versions of data sets were required. Table 3-3 to Table 3-6 list the notations that were used to describe effects of customizing strategy, estimating items, test length and correlations of latent abilities. Table 3-3 lists the notation related to customizing strategies. The second column lists the average effect representing various customizing strategies under different test dimensional structures, which were averaged over the individual data sets in the third column.

Table 3-3. Notation Related to Customizing Strategies under Different Dimensional Structures

Test Dimensionality	Customizing Strategy	Related Individual Data Sets
U_Intact →1-D cust. test	U60U	U60U_Part; U60U_All
	U40U	U40U_Part; U40U_All
U_Intact →2-D cust. test	U60M	U60M_Part_.0; _.3; _.6; _.9; U60M_All_.0; _.3; _.6; _.9;
	U40M	U40M_Part_.0; _.3; _.6; _.9; U40M_All_.0; _.3; _.6; _.9
M_Intact →3-D cust. test	M60M	M60M_Part_.0; _.3; _.6; _.9; M60M_All_.0; _.3; _.6; _.9
	M40M_Prop	M40M_Prop_Part_.0; _.3; _.6; _.9; M40M_Prop_All_.0; _.3; _.6; _.9
	M40M_Disp	M40M_Disp_Part_.0; _.3; _.6; _.9; M40M_Disp_All_.0; _.3; _.6; _.9

Note. The suffixes _.0; _.3; _.6; _.9 were not relevant to U60U_Part, U60U_All, U40U_All, or U40U_Part as they were not subject to changes in correlations between latent abilities based on the items from intact tests and the new items.

Table 3-4 lists the notation related to estimating items: using all items or only items from intact tests. The second column lists the average effects of using different

estimating items under different test dimensional structures, which were averaged over the individual data sets in the third column.

Table 3-4. Notation Related to Using Different Estimating Items under Different Dimensional Structures

Test Dimensionality	Items for Norm Estimation	Related Individual Data Sets
U_Intact →1-D cust. test	Unidimensional_All	U60U_All, U40U_All
	Unidimensional_Part	U60U_Part, U40U_Part
U_Intact →2-D cust. test	U_All	U60M_All_.0; _.3; _.6; _.9 U40M_All_.0; _.3; _.6; _.9
	U_Part	U60M_Part_.0; _.3; _.6; _.9 U40M_Part_.0; _.3; _.6; _.9
M_Intact →3-D cust. test	M_All	M60M_All_.0; _.3; _.6; _.9 M40M_Prop_All_.0; _.3; _.6; _.9 M40M_Disp_All_.0; _.3; _.6; _.9
	M_Part	M60M_Part_.0; _.3;_.6;_.9 M40M_Prop_Part_.0; _.3; _.6;_.9 M40M_Disp_Part_.0; _.3; _.6;_.9

Note. The suffixes _.0; _.3; _.6; _.9 were not relevant to U60U_Part, U60U_All, U40U_All, or U40U_Part as they were not subject to changes in correlations between latent abilities based on the items from intact tests and the new items.

Table 3-5 lists the notation related to the effect of changing test length based on 40-item intact tests. As unidimensional and two-dimensional intact tests were investigated, the dimensional structures related to this effect reflected this consideration. Although there were other customized data sets with 20 or 60 items, the selected data sets were the most similar to intact test in terms of dimensionality.

Table 3-5. Notation Related to Changing Test Length under Different Dimensional Structures

Test Dimensionality	Customizing Strategy	Test Length (Item No.)	Related Individual Data Sets
1-D Intact test	U40U_Part	20	
	U60U_All	60	
2-D Intact test	M40M_Prop_Part	20	M40M_Prop_Part_.0; _.3; _.6; _.9
	M60M_All	60	M60M_.0; _.3; _.6; _.9

Note. The suffixes _.0; _.3; _.6; _.9 were not relevant to U40U_Part or U60U_All as they were not subject to changes in correlations of latent abilities based on items from intact tests and new items.

Table 3-6 lists the notation related to the correlations of latent abilities based on unidimensional and two-dimensional intact tests. This comparison was only valid for the customized data sets whose new items measured a different latent ability than the items from intact tests. The customized data sets under unidimensionality were not related to this discussion. Similar to previous tables, the second column list the average effects representing different correlations of latent abilities based on unidimensional and two dimensional intact tests, which were averaged over the individual data sets in the third column.

Table 3-6 Notation Representing Different Correlations of Latent Abilities under Different Dimensional Structures

Test Dimensionality	Correlation Of Latent Abilities	Individual Data Sets Representing Other Manipulated Factors
U_Intact →2-D cust. test	U_.0	U60M_All_.0, U40M_All_.0, U60M_Part_.0, U40M_Part_.0
	U_.3	U60M_All_.3, U40M_All_.3, U60M_Part_.3, U40M_Part_.3
	U_.6	U60M_All_.6, U40M_All_.6, U60M_Part_.6, U40M_Part_.6
	U_.9	U60M_All_.9, U40M_All_.9, U60M_Part_.9, U40M_Part_.9
M_Intact →3-D cust. test	M_.0	M60M_All_.0, M40M_Prop_All_.0, M40M_Disp_All_.0, M60M_Part_.0, M40M_Prop_Part_.0, M40M_Prop_Part_.0
	M_.3	M60M_All_.3, M40M_Prop_All_.3, M40M_Disp_All_.3, M60M_Part_.3, M40M_Prop_Part_.3, M40M_Prop_Part_.3
	M_.6	M60M_All_.6, M40M_Prop_All_.6, M40M_Disp_All_.6, M60M_Part_.6, M40M_Prop_Part_.6, M40M_Prop_Part_.6,
	M_.9	M60M_All_.9, M40M_Prop_All_.9, M40M_Disp_All_.9, M60M_Part_.9, M40M_Prop_Part_.9, M40M_Prop_Part_.9,

Simulation Procedure

This study was based on simulated data sets rather than selected responses from real testing data for two reasons: (1) it is very difficult or even impossible to manipulate the factors of interest based on real data sets; and (2) large data sets can be generated. This section describes procedures related to data simulation, including the selection of IRT models, item parameters, and the simulee ability distribution.

IRT Models for Simulation

To generate data sets, the first step was to choose an IRT model. To generate multidimensional data sets, Doody-Bogan and Yen's (1983) compensatory model (shown below) was used in this study.

$$P_i(\Theta_j) = c_i + \frac{1 - c_i}{1 + \exp[-1.7(\sum_{k=1}^{m} a_{ik}(\theta_{jk} - b_{ik}))]}, \quad (4)$$

where

$P_i(\Theta_j)$ is the probability of a correct response to item *i* by examinee *j* whose location in an *m*-dimensional latent space is described by the proficiency vector $\Theta_j = (\theta_{j1}, \theta_{j2}, ..., \theta_{jm})$,

θ_{jk} is the ability parameter for person *j* on dimension *k*,

a_{ik} is the discrimination parameter of item *i* on dimension *k*,

b_{ik} is the difficulty parameter of item *i* on dimension *k*, and

c_i is the pseudo-chance-level parameter of item *i*.

As the three-parameter logistic model (3-PL) (shown below) is often used in unidimensional IRT calibrations, it was used to simulate unidimensional data sets.

$$P_i(\theta_j) = c_i + \frac{1 + c_i}{1 + \exp[-1.7a_i(\theta_j - b_i)]}, \quad (5)$$

where

$P_i(\theta_j)$ is the probability of a correct response to item *i* by examinee *j* whose location in the unidimensional latent space is described by θ_j,

a_i is the discrimination parameter for item *i*,

b_i is the difficulty parameter for item *i*, and

c_i is the pseudo-chance-level parameter for item *i*.

The 3-PL model was also used to estimate item characteristics and examinee proficiencies for the unidimensional IRT calibrations in this study.

Item Parameters

To make simulated data realistic, item parameter estimates obtained from real data analyses were used to generate item parameters. The results of item analyses from a

recent national administration of the ACT Math Test were used as the source of item parameters. Specifically, the simulation parameter values included the means and standard deviations of the estimated item discrimination and difficulty parameters, and the estimated guessing parameters obtained from the real data analysis using the 3-PL model.

Generating Multidimensional Data

Doody-Bogan and Yen's (1983) compensatory model was used to generate multidimensional data sets. To simulate items that measured a certain dimension, the discrimination parameters were generated as

$$a_{ik} \sim N(\mu_a, \sigma_a), \tag{6}$$

where

μ_a and σ_a are the mean and standard deviation of the *a* estimates from the ACT math test, and

a_{ik} is defined previously.

The discrimination parameters on different dimensions were generated independently of each other.

Likewise, the difficulty parameters on different dimensions were also independently generated as

$$b_{ik} \sim N(\mu_b, \sigma_b), \tag{7}$$

where

μ_b and σ_b are the mean and standard deviation of the *b* parameter estimates from the ACT math test, and

b_{ik} is defined previously.

The guessing parameters were not generated; that is, the exact values of the estimated guessing parameters based on the 3-PL model using the ACT data were used as the guessing parameters.

Generating Unidimensional Data

The unidimensional data sets were generated based on the 3-PL model. The discrimination parameters were generated as

$$a_i \sim N(\mu_a, \sigma_a), \tag{8}$$

where

μ_a and σ_a are the mean and standard deviation of the *a* parameter estimates from the ACT math test using the 3-PL model, and

a_i is the discrimination parameter for item *i*.

The difficulty parameters were generated as

$$b_i \sim N(\mu_b, \sigma_b), \tag{9}$$

where

μ_b and σ_b are the mean and standard deviation of the *b* parameter estimates from the ACT math test using the unidimensional 3-PL model, and

b_i is the difficulty parameter for item *i*.

The guessing parameters were not generated; that is, the exact values of the estimated guessing parameters based on the 3-PL model using the ACT data were used as the guessing parameters.

Simulees

To generate a multidimensional data set, the person ability vectors were generated from a standard multivariate normal distribution with $\rho(\theta_k, \theta_l)$ = .00, .30, .60, and .90. The different $\rho(\theta_k, \theta_l)$ were chosen to represent different correlations of latent abilities. Note that in most testing situations, correlations among ability parameters are likely to be at least .6. Correlations smaller than .6 may be unrealistic and were included for comparison purposes.

To generate a unidimensional data set, the person ability parameters were generated from a standard normal distribution.

To achieve relatively stable estimates, a sample size of 1,000 was simulated for each unidimensional and two-dimensional intact test. For each repetition, a different set of person ability vectors was simulated.

Item Response Matrix

For each combination of ability and item parameters, the probability of correctly answering item *i* was computed using the selected IRT models. To generate a multidimensional data set, the probability was calculated using Doody-Bogan and Yen's (1983) compensatory model. To generate a unidimensional data set, the probability was calculated using the 3-PL model.

The calculated probability was then compared to a random number that was generated from a uniform distribution on the range (0, 1). If the probability was less than the random number, a response of 0 to item *i* was generated, otherwise, a response of 1 was generated. The resulting dichotomous item response data sets were used for item and ability calibrations.

Research Design

To account for sampling error, data generation and analysis were replicated 50 times. In order to answer the research questions, a partially crossed factorial design was used in this study..

The manipulated conditions were the following:

1. dimensionality of the intact test:
    1) unidimensional intact test;
    2) two-dimensional intact test;
2. customizing strategies:
    1) adding items to a full-length intact test (applicable to both unidimensional and two-dimensional intact tests);
    2) adding items to an intact test that was shortened proportionally in terms of dimensionality (only applicable to a two-dimensional intact test);
    3) adding items to an intact test that was shortened disproportionally in terms of dimensionality (only applicable to a two-dimensional intact test);
3. items used for ability estimation:
    1) only items from intact tests;
    2) all items in customized tests;
4. correlation between latent abilities:
    1) .00;
    2) .30;
    3) .60;
    4) .90;
5. test length of the customized instruments based on 40-item intact tests (which refers to the number of items used to estimate norms)

1) 20 items;
2) 40 items:
3) 60 items.

To facilitate comparison of the factors of interest, two unidimensional data sets, **U60U** and **U40U**, that included intact items and new items both measuring the same latent ability were constructed. The two data sets were constructed as a frame of reference to investigate the impact of customizing tests by (1) adding items to a full-length intact test, or (2) adding items to a shortened intact test, when the unidimensionality assumption was met.

Data Analysis Procedures

Descriptions of the data analysis procedures are divided into three parts. The first part describes how ability estimates and normative information were obtained based on data sets representing intact tests and the corresponding customized tests. The second part describes how the obtained ability estimates and corresponding normative information were compared. A brief description of the computer programs used is given in the last part.

Data Analysis

To investigate the validity of customized norms, the data analyses were conducted as follows: 1) estimate examinee ability based on an intact test and the corresponding customized tests; 2) create tables of percentile ranks (PR) based on the ability estimates obtained from the intact test; 3) derive PR estimates for the corresponding customized tests by comparing the corresponding ability estimates to the PR table constructed in the previous step.

Item and Ability Calibrations

Item parameter and ability estimates were obtained in two steps in this study. First, item parameters were estimated using the data sets representing the intact test, which were considered nationally calibrated in this study. Second, examinee proficiency was estimated based on the "customized" data sets by fixing the parameters of the intact items (only the parameters of new items were estimated in this step). By doing so, the ability estimates obtained from the "customized" data sets were put on the score scale of the intact test.

The ML estimator was used to estimate simulee proficiency. The ML proficiency estimator of examinee *j* is the value of θ that maximizes the likelihood function (shown below)

$$\log_e L(\theta_j) = \sum_{i=1}^{n} \{x_{ij} \log_e P_i(\theta_j) + (1 - x_{ij}) \log_e [1 - P_i(\theta_j)]\}, \quad (10)$$

where

$P_i(\theta_j)$ is the probability of a correct response to item *i* by examinee *j*,

x_{ij} is the score of examinee *j* on item *i* (*i* = 1, 2,…, *n*), and

$L(\theta_j)$ is the likelihood function for examinee *j*.

The implicit likelihood equation to be solved is

$$\frac{\partial \log L(\theta_j)}{\partial \theta_j} = \sum_{i=1}^{n} \frac{x_{ij} - P_i(\theta_j)}{P_i(\theta_j)[1 - P_i(\theta_j)]} \cdot \frac{\partial P_i(\theta_j)}{\partial \theta_j} = 0, \quad (11)$$

where

all variables were defined previously.

The ML estimator was obtained by setting the first derivative of the log likelihood equal to zero. The computer program BILOG-MG (Zimoski, Mislevey, and Bock, 2002 was used for estimation.

Creating National Norms Based on the Intact Tests

PR tables were constructed using the proficiency estimates obtained from the intact tests. The resulting PR tables were referred to as the "national norms" in the study.

Deriving Customized Norms

To derive normative information based on customized instruments, the ability estimates obtained from the "customized" data sets were compared to the PR table constructed in the previous step. The resulting PR estimates are referred to as the "customized norms" in the present study.

Evaluation Criteria

The primary focus of this study was to investigate the degree of similarity between the normative information derived from an intact test and that derived from customized instruments. Based on the data set representing the unidimensional intact test, four data sets (**U60U, U60M, U40U,** and **U40M**) were constructed to represent four customized testing situations. Based on the data set representing the two-dimensional intact test, three data sets (**M60M, M40M_PROP,** and **M40M_DISP**) were constructed to represent three customized testing situations. The customized norms were derived by comparing the corresponding ability estimates to the PR table constructed from "intact test" data sets.

To evaluate the validity of customized norms, both ability estimates and PR estimates derived from the intact test and customized instruments were compared in terms of: correlation, mean difference (bias), and mean absolute difference. The comparison process is described next.

Comparing Ability Distributions

To compare the estimated ability distributions, selected percentiles ($P_{10}, P_{20}, P_{30}, P_{40}, P_{50}, P_{60}, P_{70}, P_{80}, P_{90}$) of the ability distribution obtained from the

customized data sets were compared to those obtained from the intact data sets. The percentile difference, which is defined as the difference between the same percentile of the two distributions on the theta score scale, was used to evaluate differences between distributions.

Comparing Ability Estimates

For each examinee, several ability estimates were obtained from the data sets representing the intact and corresponding customized instruments. The ability estimates obtained from the "customized" data sets were compared to those obtained from the "intact test" data set based on three criteria:

1) Pearson correlations between ability estimates;

2) mean differences;

3) mean absolute differences.

Comparing Percentile Rank Estimates

For all examinees, the PRs were estimated by comparing the ability estimates obtained from the "customized" data sets to the PR tables constructed from the "intact test" data set. The PR estimates obtained from the "customized" data sets were compared to those obtained from the "intact test" data set based on three criteria:

1) Pearson correlations between PR estimates;

2) mean differences;

3) mean absolute differences.

Computer Programs

Several computer programs were used in this study. They included (a) a program developed by the author to generate data sets representing intact tests and customized tests using SAS Interactive Matrix Language (SAS/IML) (SAS institute, 1990) and the SAS MACRO language; (b) BILOG-MG (Zimoski, Mislevey, and Bock, 2002) for

unidimensional IRT item and ability estimates; (c) a program developed by the author to compare ability and PR estimates using the SAS MACRO language.

CHAPTER IV. RESULTS AND DISCUSSION

The purpose of this study was to investigate the impact of various factors on the validity of customized norms. The study was conducted using the methods described in Chapter III. The factors investigated in this study were: 1) customizing strategies (adding items to a full-length intact test, a proportionally shortened intact test or a disproportionally shortened intact test), 2) using different items to estimate normative information (only the items from an intact test or all items from a customized instrument), 3) test length (number of items used to estimate norms, 20, 40, 60 items), 4) dimensional structure of customized tests (adding items measuring the same latent ability to a unidimensional intact test, adding items measuring a different latent ability to a unidimensional intact test or to a two-dimensional intact test), 5) correlations of latent abilities (r = .00, .30, .60, .90) for multidimensional tests.

This chapter is made up of four sections. The first section describes the characteristics of the simulated data sets representing intact and customized tests. Section 2 describes the results when the unidimensionality assumption was satisfied (assuming new items and intact items measured the same latent ability), which includes the main effects (using different item sets to estimate normative information, changing customizing strategy and varying test length) and their interaction effects. The results for multidimensional data sets are described in Section 3, including the main and interaction effects studied in Section 2 and the effect of changing the correlations of the latent abilities (assuming new items and intact items measured different latent abilities). A summary of the results is given in the Section 4.

Description of Simulated Data and Item Parameter Estimates

Table A-1 gives a summary of the item parameters (discrimination and difficulty) used to simulate the data sets. A simple structure design was used; that is,

each item measured one latent ability and its discrimination for other dimensions was zero. The item discrimination parameters were simulated based on a lognormal distribution with mean of 1.04 and standard deviation of .40. The item difficulty parameters were simulated using a normal distribution with mean of -.25 and standard deviation of 1.03.

To demonstrate the effect of applying the 3-PL model to data sets of different structures, Table A-2 lists the summary statistics of the unidimensional item parameter estimates for the unidimensional data sets and those of the multidimensional data sets averaged across correlation levels. Several issues are worth mentioning. First, the item parameter estimates of the three unidimensional data sets (U_Intact, U60U, and U40U) were similar to the simulation parameters. For item discrimination estimates, the mean of the unidimensional tests was about 1.14 (the parameter was 1.04) and the standard deviation was about .35 (the parameter was .40). For item difficulty estimates, the mean was around -.25 (the parameter was -.25) and the standard deviation was .96 (the parameter was 1.03).

Second, the mean item discrimination estimates were different for the unidimensional and multidimensional data sets. The three unidimensional tests (U_Intact, U60U, and U40U) were very similar, and the mean discrimination estimates ranged from 1.12 to 1.15. The six multidimensional data sets (U60M, U40M, M_Intact, M60M, M40M_Prop, and M40M_Disp) were somewhat different depending on how they were constructed. It seemed that the more complicated the dimensional structure, the lower the average unidimensional discrimination estimates were. Generally speaking, the multidimensional data sets were less discriminating than the unidimensional data sets. The mean discrimination estimates of the multidimensional data sets ranged from .84 to .95. However, the standard deviations of the discrimination estimates were very similar for the unidimensional and multidimensional data sets (ranging from .30 to .39).

Third, all data sets (intact and customized tests) were similar in difficulty. The magnitude of the mean difficulty estimates was similar for all data sets, ranging from -.23 to -.27. Note, however, these data sets differed in the variability of the unidimensional difficulty estimates. Generally speaking, the unidimensional tests (U_Intact, U60U, and U40U) were less variable (the mean standard deviation ranged from .94 to .98) than the multidimensional tests (U60M, U40M, M_Intact, M60M, M40M_Prop, M40M_Disp) (the mean standard deviation ranged from 1.25 to 1.40). Last, the guessing parameter estimates of all data sets were similar; the mean guessing parameter estimates ranged from .18 to .22.

Table A-3 lists the summary statistics of the unidimensional item parameter estimates at each correlation level (.00, .30, .60, .90) for the six multidimensional data sets, U40M, U60M, M_Intact, M60M, M40M_Prop, M40M_Disp (the intact items and new items measured different latent abilities). These data sets were similar in difficulty (the mean difficulty estimates ranging from -.30 to -.24). The standard deviations of the difficulty estimates decreased as the correlations increased from .00 to .90 (around 2 when the correlations were .00, and around 1 when the correlations were .90). The mean item discrimination and guessing parameter estimates differed across correlation levels. For item discrimination estimates, the magnitude of the means generally increased as the correlations increased. For example, the mean discrimination estimate of U40M increased from .86 to 1.06 when the correlation increased from .00 to .90. Similar patterns were also observed for other data sets except M_Intact, which exhibited a minor deviation from the pattern. The standard deviations of the discrimination estimates generally decreased when the correlations increased. When the correlations were .00, the average standard deviations ranged from .35 to .55. When the correlations increased to .90, the average standard deviations ranged from .30 to .33. For guessing parameter estimates, the magnitudes of the mean and standard deviation tended to decrease as the correlations increased. For example, for the data set U40M, the mean guessing parameters decreased

from .24 to .19 as the correlations increased from .0 to .9. Similar patterns were also observed for other data sets. In general, whether the unidimensionality assumption was satisfied seemed to have an impact on the item parameter estimates. As the correlations of latent abilities increased (data sets became more unidimensional), the unidimensional item parameter estimates became more similar to the simulated item parameter values, and the standard errors of item parameters became smaller.

To provide a general picture of the ability estimates, Table A-4 describes the estimated ability distributions based on the unidimensional intact test (U_Intact) and related data sets. Table A-5 describes the estimated ability distributions based on the two-dimensional intact test (M_Intact) and related data sets. In both tables, the suffix _All was used when all items in the customized data sets were used, and the suffix _Part was used when only items from intact tests were used for norms estimation. In both Tables A-4 and A-5, selected percentiles (10, 20, 30, 40, 50, 60, 70, 80, and 90) are averaged across data sets. Generally speaking, these estimated ability distributions were very similar. There seemed to be no systematic differences at different correlation levels. More detailed information about the estimated ability distributions is presented later in this chapter.

Table A-6 lists the mean proportion correct scores for the data sets representing intact and customized tests across the 50 replications. Consistent with the results for the ability estimates shown in Tables A-4 and A-5, the average proportion correct scores for all tests were very similar, ranging from .60 to .64. No systematic differences were observed at different correlation levels.

Table A-7 lists the average reliability estimates (Cronbach's Alpha) of the data sets representing intact and customized tests. Generally speaking, the reliability estimates reasonably reflected the impact of changing test length and dimensional structure. If data sets had the same dimensional structure, the reliability estimates increased as the test length increased. For example, the data sets U60U_All, U_Intact, U40U_All, U40U_Part had 60, 40, 40, 20 items respectively, and their reliabilities were .93, .90, .90, and .81,

respectively. If data sets had the same number of items, the reliability estimates of unidimensional tests were higher than those of the multidimensional tests. Consider the unidimensional data set U60U_All and the multidimensional data set U60M_All for example: U60U_All had a higher reliability estimate than U60M_All. The same pattern was also observed for the other data sets (e.g., U40U_All, U40M_All). Additionally, for multidimensional data sets (U60M, U40M, M_Intact, M60M, M40M_Prop, M40M_Disp), the reliability estimates increased as the magnitude of correlations of the latent abilities increased. For example, the reliability estimates of U60M_All increased from .86 to .92 as the correlations increased from .00 to .90. The same pattern was observed for other data sets.

In short, the simulated data sets appeared to be reasonable given the simulation conditions. Results based on these data sets are described in following sections.

Results for Unidimensional Data Sets

In this section, the validity of customized norms was investigated when the unidimensionality assumption was satisfied (i.e., new items measured the same latent ability as items from an intact test). Table 4-1 lists the manipulated factors, brief comments and abbreviations of the related data sets. Three main effects (items for norms estimation, customizing strategy, test length) and the interaction effect of changing items for norms estimation and customizing strategy were investigated. Note that the interaction effect of test length and other factors were not included due to the design of this study. Based on the design, tests of different length were represented by particular combinations of estimation items and customizing strategies. Specifically, the data set U60U_All represented the 60-item test length condition and the data set U40U_Part represented the 20-item condition.

Using Different Estimation Items

When the unidimensionality assumption was satisfied, estimating normative information using only items from intact tests (Unidimensional_Part) or all items from customized tests (Unidimensional_All) each yielded accurate normative estimates (Table

Table 4-1. Descriptions and Notations for Unidimensional Data Sets Based on Manipulated Factors

Effect	Abbreviations of Data Sets	Description
Estimating items	Unidimensional_All Unidimensional_Part	Estimating normative information (a) using all items in customized instruments (Unidimensional_All, which is the average of U60U_All, U40U_All), or (b) only items from the intact test (Unidimensional_Part, which is the average of U60U_Part, U40U_Part).
Customizing strategy	U60U U40U	Adding 20 items measuring the same dimension as the intact items to (a) full-length intact tests, or (b) shortened intact tests
Test length	U40U_Part (20-item) U_Intact (40-item) U60U_All (60-item)	Number of items used to estimate norms (20, 40, and 60 items)
Interaction of estimation items and customizing strategies	U40U_All U40U_Part U60U_All U60U_Part	The combined effects of using different estimation items and customizing strategies

Note: The suffix _All/_Part is used to indicate the norms estimation procedure: using all items or only items from intact tests. If the suffix is not present in the name of a data set, the results were averaged across the two estimation procedures.

A-8, Table A-9). Nearly all criterion measures yielded very similar results except for a minor deviance in the absolute differences of percentile estimates (5.00 for Unidimensional_All and 3.69 for Unidimensional_Part). This suggested the average discrepancies of the percentile estimates in either direction (underestimation or overestimation) based on customized tests and intact tests were about 5 if all items were used to estimate norms, while the average discrepancies were about 4 if only items from intact tests were used. The differences were very small based on the magnitudes of

standard errors (0.36 for Unidimensional_All and 0.41 for Unidimensional_Part). Using intact items and using all items of the customized instruments seemed to yield comparable results when the unidimensionality assumption was satisfied.

Customizing Strategy

Data set U60U represented adding items to full-length intact tests, and data set U40U represented adding items to shortened intact tests. Both conditions yielded good ability and percentile estimates, but the data set U60U yielded slightly more accurate normative estimates than U40U (Tables A-10, A-11). The differences were most obvious in the absolute differences of ability estimates (.08 versus .25) and absolute differences of percentile estimates (1.92 versus 6.23) for U60U and U40U.

Note that the customized instrument U60U yielded virtually the same results as the intact tests. The discrepancies of the estimated ability distributions were less than .01 for selected percentiles. The correlation of ability estimates was close to 1 and the average absolute difference was .08. At the same time, the correlation of percentile estimates was close to 1 and the average absolute difference was 1.92.

Data set U40U also yielded good normative estimates, but the results were slightly less accurate than those of U60U. The average absolute difference of ability estimates was .25 (compared to .08 for U60U), and the average absolute difference of percentile estimates was 6.23 (compared to 1.92 for U60U).

Test Length

When customization was conducted based on 40-item intact tests under unidimensionality, the data sets U40U_Part and U60U_All represented the two conditions of changing test length: changing test length from 20 to 60 items. The results suggested that changing test length had an impact on the accuracy of normative estimates even if the unidimensionality assumption was satisfied. Generally speaking, lengthened tests (U60U_All) yielded more accurate normative estimates than shortened tests

(U40U_Part) (Table A-12, Table A-13). For the 20-item U40U_Part and 60-item U60U_All, the differences were observed in the absolute differences of ability estimates (compared to the 40-item intact test; .30 versus .15) and the absolute differences of percentile estimates (7.38 versus 3.84).

Combined Effects of Customizing Strategies and Estimation Items

Four data sets were related to the combined effect of customizing strategies and norms estimation. These were ranked as U60U_Part >U60U_All > U40U_All > U40U_Part by the decreasing accuracy of the norms estimates. The differences were observed in the absolute differences of ability estimates and percentile estimates, but not in other measures (Table A-14, Table A-15).

Recall that the results from the main effects indicated that (a) customizing full-length intact tests yielded better estimates than customizing reduced-length intact tests, and (b) using only items from intact tests for norms estimation yielded more accurate estimates than using all items. The combined effects were similar to the main effects. By design, the results of U60U_Part were the same as the intact tests. The results of U60U_All were only slightly worse than U60U_Part, which were the best among the rest. The average absolute difference of ability estimates for U60U_All was .15, and the average absolute difference of percentile estimates was 3.84. In comparison, the results of U40U_All and U40U_Part were noticeably worse. The average absolute differences of ability estimates for U40U_Part and U40U_All were .27 and .30, and the average absolute differences of percentile estimates were 6.66 and 7.38.

The above ranking suggested an interaction effect of customizing strategy and estimation items. Maintaining all items from intact tests was very important in customizing. The results of customization based on full-length intact tests (U60U_Part, U60U_All) were more accurate than those based on reduced-length intact tests

(U40U_Part, U40U_All). The normative estimates based on reduced-length customized data sets were not as accurate as those based on full-length customized data sets, which is true even if the new items were used in estimation. The effect of using different estimation items was more pronounced for the 60-item data sets.

Results for Multidimensional Data Sets

The previous section describes the results under unidimensionality. In this section, the results are described when the unidimensionality assumption was violated. The multidimensional data sets included customized data sets based on unidimensional and two-dimensional intact tests. Different from the previous section, the new items and items from intact tests measured different latent abilities. Together with the factors investigated in the previous section (customizing strategy, items used for normative estimation, test length), the impact of the correlations of latent abilities are also investigated.

Using Different Estimation Items

When new items measure a different latent ability, they can either be included or not included to estimate normative information. Similar to the previous section, as both unidimensional intact tests and two-dimensional intact tests were customized, the results are described accordingly. Table 4-2 lists the abbreviations of the conditions related to the effects of estimation items. The prefixes U_ and M_ represent the dimensionality of intact tests: unidimensional intact tests or two-dimensional intact tests. Specifically, the notation U_All represents the average of U60M_All and U40M_All; U_Part represents the average of U60M_ Part and U40M_ Part; M_All represents the average of M60M_ All, M40M_Prop_All and M40M_Disp_All; and M_Part represents the average of M60M_ Part, M40M_Prop_ Part and 40M_Disp_ Part. The suffixes _All and _Part represent the norms estimation procedure: using all items in customized tests or only items from intact tests.

Table 4-2. Descriptions and Notations Related to Items for Normative Estimation

Customized Tests	Intact Tests	Using All items For Norms Estimation	Using Items from Intact Tests for Norms Estimation
Two-dimensional	Unidimensional	U_All	U_Part
Three-dimensional	Two-dimensional	M_All	M_Part

When unidimensional intact tests were customized by adding items measuring a different latent ability, using only items from intact tests led to more accurate results than using all items. This was observed in absolute differences of ability estimates and percentile estimates. The average absolute differences of ability estimates were .15 versus .26 for U_Part and U_All, and the average absolute differences of percentile estimates were 3.69 versus 6.43 (Table A-17). The differences were not observed in estimated ability distributions, correlations and mean differences of ability estimates and percentile estimates (close to zero) (Tables A-16, A-17).

When two-dimensional intact tests were customized, using only items from intact tests in estimation (M_Part) yielded more accurate results than using all items (M_All). This was similar to the previous condition, but the differences between the two estimation procedures were much larger. The differences were observed in the correlations and absolute differences of ability estimates and percentile estimates. For M_Part and M_All, the correlations of ability estimates were .98 and .82, and the average absolute differences were .22 and .49. Their correlations of percentile estimates were .82 and .98, and the absolute differences were 5.55 and 13.33 (also Tables A-16, A-17).

Customizing Strategy

When intact tests were unidimensional, two customizing strategies were investigated: adding items to full-length intact tests (U60M) and adding items to shortened intact tests (U40M). When intact tests were two-dimensional, three customizing strategies were investigated: adding new items to full-length intact tests

(M60M), adding items to proportionally shortened tests (M40M_Prop), and adding items to disproportionally shortened tests (M40M_Disp). Table 4-3 lists the customizing strategies based on the unidimensional and two-dimensional intact tests.

Table 4-3. Descriptions and Notations Related to Customizing Strategies

Customized Test Dimensionality	Intact Test Dimensionality	Customizing Strategy	
		Customizing Full-length Intact Tests	Customizing Reduced-length Intact Tests
Two-dimensional	Unidimensional	U60M	U40M
Three-dimensional	Two-dimensional	M60M	M40M_Prop M40M_Disp

Note:

U60M stands for customizing based on full-length unidimensional intact tests, and U40M for customizing based on reduced-length intact tests.

M60M stands for customizing based on full-length two-dimensional intact tests, M40M_Prop for customizing based on proportionally reduced-length intact tests, and M40M_Disp for customizing based on disproportionally reduced-length intact tests.

When unidimensional intact tests were customized, data sets based on full-length intact tests (U60M) yielded more accurate results than those based on shortened intact tests (U40M) (Tables A-18, A-19). The differences were observed in the absolute differences of ability estimates and percentile estimates. Compared to U40M, the data set U60M yielded lower absolute differences of ability estimates (.08 compared to .31) and lower absolute differences of percentile estimates (1.80 compared with 7.77). There were essentially no differences in the estimated ability distributions, mean differences (close to zero) and correlation measures of ability estimates and percentile estimates (both very high).

It is interesting that customized data sets based on full-length intact tests, U60M, actually led to very good normative estimates even though the unidimensionality assumption was violated. The average absolute difference of percentile estimates was

1.80, which means that on average the percentile estimates based on intact tests and customized tests differed by approximately 1.80 in either direction.

When intact tests were two-dimensional, customized data sets based on full-length intact tests (M60M) also yielded more accurate results than those based on shortened intact tests (M40M_Prop, M40M_Disp), but the differences were much larger in this condition (Tables A-18, A-19). This was observed in the correlations and average absolute differences of ability estimates and percentile estimates. For M60M, M40M_Prop and M40M_Disp, the correlations of ability estimates were .98, .89, .88, respectively and the average absolute differences were 0.14, .41, and .42. The corresponding correlations of percentile estimates were .99, .90, .87, respectively and the average absolute differences were 3.61, 11.20 and 11.28. Note that the discrepancies arising from customization based on full-length or reduced-length intact tests were very large in this situation. The customized data set based on a full-length intact test, M60M, yielded fairly good results; its average absolute difference of percentile estimates was 3.61. The customized data sets based on reduced-length intact tests, M40M_Prop and M40M_Disp, yielded poor results; their average absolute differences of percentile estimates were about 11.

An interesting observation is that proportionally or disproportionally shortening two-dimensional intact tests seemed to have a very small impact on the normative estimates. The results of M40M_Prop and M40M_Disp were very similar and no systematic differences were identified by any criterion measure.

Test Length

As the intact tests included 40 items, 20-item and 60-item data sets, M40M_Prop_Part and M60M_All, were investigated. These data sets represented tests of changed length: As shown in Tables A-20 and A-21, the normative estimates based on the 20-item and 60-item data sets were both different from those based on intact tests. It

seemed to suggest that changing test length would affect the accuracy of the norm estimates.

The 20-item data set M40M_Prop_Part yielded slightly less accurate normative estimates than the 60-item data set M60M_All. The differences were shown by the average absolute differences of percentile and ability estimates. For the 20-item data set M40M_Prop_Part and 60-item data set M60M_All, the absolute differences of ability estimates were .36 versus .28 and the absolute differences of percentile estimates were 9.16 versus 7.22.

Combined Effects of Customizing Strategies and Estimation Items

The combined effects of customizing strategies and estimation items based on multidimensional data sets are described in this section. Table 4-4 lists the data sets related to the combined effects.

Table 4-4. Multidimensional Data Sets Related to the Combined Effects of Customizing Strategies and Estimation Items

		Customizing Strategy		
		Customizing Full-length Tests	Customizing Shortened Tests	
Items for Estimation	All Items	U60M_All	U40M_All	
	Items from Intact Test	U60M_Part	U40M_Part	
	All Items	M60M_All	M40M_Prop_All	M40M_Disp_All
	Items from Intact Test	M60M_Part	M40M_Prop_Part	M40M_Disp_Part

The results for customized unidimensional intact tests are listed in Tables A-22, A-23. Recall that the main effects implied that a) full-length customized tests yielded more accurate results than reduced-length tests, and b) using only items from intact tests yielded more accurate results than using all items. The simple effects were very similar to the main effects. For both data sets constructed using different customizing strategies

(U60M and U40M), using only items from intact tests yielded more accurate results than using all items in customized tests. That is, U60M_Part yielded more accurate results than U60M_All, U40M_Part more accurate than U40M_All. Likewise, in both estimation situations (using all items or only items from intact tests to estimate norms), the customized data sets based on full-length intact tests produced more accurate results than those based on reduced-length intact tests. That is, U60M_All yielded more accurate results than U40M_All and U60M_Part more accurate than U40M_Part. The effects were observed in the correlations and absolute differences of ability estimates and percentile estimates.

The data sets were ordered according to the decreasing accuracy: U60M_Part, U60M_All, U40M_Part, U40M_All.There was some indication of an interaction between customization strategy and items used to estimate norms in that the differences between estimation items conditions were larger for full-length customized tests.

The results for the customized two-dimensional intact tests are listed in Tables A-22, A-23. Again the simple effects were very similar to the main effects. For the three data sets constructed based on different customizing strategies (M60M, M40M_Prop, M40M_Disp), using only items from intact tests yielded more accurate results than using all items in customized tests. That is, M60M_Part yielded more accurate results than M60M_All, M40M_Prop_Part more accurate than M40M_Prop_All, and M40M_Disp_Part more accurate than M40M_Disp_All. Likewise, in both estimation situations (using all items or only items from intact tests), customizations based on full-length intact tests (M60M) produced more accurate results than those based on reduced-length intact tests (M40M_Prop, M40M_Disp). That is, M60M_Part yielded more accurate results than M40M_Prop_Part, M40M_Disp_Part; and M60M_All, more accurate than M40M_Prop_All, M40M_Disp_All. The effects were observed in the correlations and absolute differences of ability estimates and percentile estimates.

The data sets formed four groups according to the decreasing accuracy of the results: (1) M60M_Part, (2) M60M_All, (3) M40M_Prop_Part, M40M_Disp_Part, (4) M40M_Prop_All, M40M_Disp_All. By design, the data set M60M_Part yielded the same results as the intact tests. The second group M60M_All yielded the most accurate results among the rest. The third group, M40M_Prop_Part and M40M_Disp_Part, yielded slightly less accurate results than the second group, but the differences were not large. The data sets in the last groups yielded the worst estimates.

There was very little evidence of an interaction between customizing strategies and estimation items. The differences between using all items or only items from the intact tests were fairly consistent for both customization strategies,

Correlations of Latent Abilities

Recall that new items measure a different latent ability than items from intact tests for the multidimensional data sets. In this section, the results of changing the correlations of latent abilities measured by new items and items from intact tests are discussed. The main effects of correlations of latent abilities are described first, then the combined effects of correlation, customizing strategy, estimation items, and finally the combined effects of correlations and test length. As both unidimensional and two-dimensional intact tests were customized, the results are described accordingly.

Main Effects of Correlations of Latent Abilities

When intact tests were unidimensional, the correlations of latent abilities measured by the items from intact tests and items added in customization seemed to have a small impact. The magnitudes of all criterion measures were very similar at different correlation levels. No systematic trends were observed when the correlation increased from .00 to .90 (Tables A-24, A-25).

When intact tests were two-dimensional, the correlations of latent abilities had a relatively large impact on the estimates. The higher the correlations of latent abilities, the

more accurate were the normative estimates. This was observed in the correlations and absolute differences of ability estimates and percentile estimates. As the correlation of latent abilities increased from .00 to .90, the correlations of ability estimates and percentile estimates increased and the absolute differences decreased. Specifically, when the correlations of latent abilities increased (.00, .30, 60, .90), the correlation of the ability estimates based on customized tests and intact tests increased (.90, .92, .96, .97), the absolute differences decreased (.34, .31, .23, .17). At the same time the correlations of percentile estimates increased (.91, .92, .96, .98), and the absolute differences decreased (9.39, 8.39, 6.22, 4.42) (Tables A-26, A-27). As mentioned previously, correlations among ability parameters are likely to be .6 or higher in most testing situations. When the correlation was .60 or higher, the ability estimates and percentile estimates were rather accurate although data sets with higher correlations (.90) had better accuracy in normative estimates than those with lower correlation (.60).

Combined Effects of Correlations of Latent Abilities and Customizing Strategies

When unidimensional intact tests were customized, U60M_All and U40M_All represented the combined effects of customizing strategies and estimation procedures. Note that the data sets U60M_Part and U40M_Part are not relevant to this situation because in those data sets the new items measuring a different ability were not used in estimation. Their results are, therefore, not subject to the changes in correlations of latent abilities measured by items from intact tests and new items so only the combined effects of correlations and customization strategies are relevant for these customized unidimensional data sets. For the data sets U60M_All and U40M_All, changing the correlations of latent abilities seemed to have a very small impact on the results (Tables A-28, A-29). The results of U60M_All were similar at different correlations levels and no

systematic differences were found. The same was true for the results for U40M_All at different correlation levels.

When two-dimensional intact tests were customized by adding items measuring a different latent ability, higher correlations of latent abilities led to more accurate estimates (Tables A-30, A-31, and A-32). This was true for M40M_Prop_Part, M40M_Prop_All, M40M_Disp_Part and M40M_Disp_All. The trend was observed in the correlations and average absolute differences for ability estimates and percentile estimates. Take M40M_Prop_All as an example. When the correlation increased from .00 to .90, the correlation of ability estimates increased from .49 to .92 and mean absolute difference decreased from .80 to .31. The correlations of percentile estimate changed from .49 to .93 and mean absolute differences decreased from 22.46 to 7.96. Note, however, the results of M60M_All were different from the others, and the increase of correlation seemed to have little impact on the results. The average absolute differences of ability estimates and percentile estimates were similar at different correlation levels.

There were also combined effects of correlations of latent abilities, customizing strategies and estimation items. The impact of correlations of latent abilities was very large if all items in a reduced-length customized test were used to estimate norms, as seen in to the data sets M40M_ Prop_All and M40M_ Disp_All. The impact of correlations was moderate if only items from intact tests in a reduced-length customized test were used to estimate norms, as seen in the data sets M40M_ Prop_Part and M40M_ Disp_Part. The impact was very small if customization was based on full-length intact tests, as observed in the data set M60M_All. When the correlations of latent abilities increased from .00 to .90, the average absolute differences of percentiles changed from about 22 to 8 for M40M_Prop_All and M40M_Disp_All. In comparison, the changes were from about 10 to 7 for M40M_Prop_Part and M40M_Disp_Part.

For customized data sets based on reduced-length intact tests, there were combined effects of correlations of latent abilities and estimation items (M40M_Prop_All

versus _Part, M40M_Disp_All versus _Part). When the correlations of latent abilities were low (.00, .30), the results based on all items and only items from intact tests were very different. When the correlations of latent abilities were high (.90), the differences were much smaller. Take M40M_Prop_All and M40M_Prop_Part for example (Table A-32). When the correlations of latent abilities were .00, the results were very different. Their correlations of percentile estimates were .49 versus .88 and the average absolute differences were 22.46 versus 10.29. When the correlations of latent abilities increased to .90, the differences virtually disappeared. The correlations of percentiles were .93 versus .94, and the absolute differences were 7.96 versus 7.38 (Table A-32).

An interesting observation is that the correlations of latent abilities also affected the magnitudes of standard errors of criterion measures with higher correlations leading to smaller standard errors. This was observed for the correlations, differences and average absolute differences of ability estimates and percentile estimates, and was especially dramatic for average absolute differences of percentile estimates. Consider M60M_All for example. The standard errors of absolute differences of percentile estimates decreased from 2.73 to .43 when the correlation of latent abilities increased from .00 to .90.

Combined Effects of Correlations of Latent Abilities and Test Length

There were combined effects of correlations of latent abilities and changing test length. For the shorter tests, there was a clear pattern of increased accuracy as the correlations increased. The pattern was less obvious for the longer tests. This was observed in the correlations and average absolute differences of ability estimates and percentile estimates (Tables A-33, A-34).

For the 20-item data set M40M_Prop_Part, when the correlations of latent abilities increased from .00 to .90, the correlations of ability estimates increased from .86 to .92 and the average absolute differences decreased from .40 to .30. The correlations of

percentile estimates increased from .88 to .94 and the average absolute differences changed from 10.29 to 7.38. The trends for the corresponding 60-item data sets were not nearly as uniform.

Comparing Unidimensional and Multidimensional Data Sets

In the previous two sections, results based on unidimensional data sets and multidimensional data sets were described. In this section the results for the unidimensional and multidimensional data sets are compared. Three effects are described: items used for norms estimation, customizing strategies and changing test length. For multidimensional data sets, the results were averaged across different correlations of latent abilities.

Using Different Items for Norms Estimation

Based on either unidimensional or multidimensional data sets, using only items from intact tests yielded more accurate norms estimates than using all items in customized data sets (Table A-35). When unidimensionality was satisfied, Unidimensional_Part yielded more accurate results than Unidimensional_All. For the multidimensional data sets based on unidimensional intact tests, U_Part yielded more accurate results than U_All. For the data sets based on two-dimensional intact tests, M_Part yielded more accurate results than M_All. The effects were observed in the correlations and absolute differences for both ability and percentile estimates.

Note that the results of unidimensional data sets were more accurate than their multidimensional counterparts. Unidimensional_All yielded slightly better results than U_All (its multidimensional counterpart based on unidimensional intact tests) and much better results than M_All (its counterpart based on two-dimensional intact tests). The average absolute differences were 5.00, 6.43, and 13.33 for Unidimensional_All, U_All, and M_All, respectively. The same was observed for Unidimensional_Part, U_Part, and

M_Part. Their average absolute differences of percentile estimates were 3.69, 3.69 and 5.55.

There were combined effects of estimation items and data set dimensional structures. When unidimensionality was satisfied, whether new items were used to estimate norms made little difference in the results. When multidimensional tests were customized based on unidimensional intact tests, the differences increased. When customization was made based on two-dimensional intact tests, the differences increased further. Consider average absolute differences of percentile estimates for example. They were 5.00 versus 3.69 for unidimensional data sets (Unidimensional_All and for Unidimensional_Part), 6.43 versus 3.69 averaged across the multidimensional data sets based on unidimensional intact tests (U_All, U_Part), and 13.33 versus 5.55 for the multidimensional data sets based on two-dimensional intact tests (M_All vs. M_Part).

Customizing Strategies

Based on either unidimensional or multidimensional data sets, customized data sets based on full-length intact tests yielded more accurate normative estimates than those based on reduced-length intact tests (Table A-36). When unidimensionality was satisfied, U60U yielded more accurate results than U40U. When multidimensional customization was made based on unidimensional intact tests, U60M yielded much more accurate results than U40M. When customization was made based on two-dimensional intact tests, M60M yielded more accurate estimates than M40M_All or M40M_Disp. The effects were observed in the correlations and average absolute differences of ability estimates and percentile estimates.

Note that the results based on unidimensional intact tests were more accurate than those based on two-dimensional data sets. For customized data sets based on full-length intact tests, the average absolute differences of percentiles were 1.92, 1.80, and 3.61 for U60U, U60M, and M60M, respectively. For customized data sets based on

reduced-length intact tests, the average absolute differences of percentiles were 6.23, 7.77, 11.20, and 11.28 for U40U, U40M, M40M_Prop, and M40M_Disp, respectively.

There were combined effects of customizing strategies and data set dimensional structures. When unidimensionality was satisfied, relatively small differences were observed between U60U and U40U. A very similar pattern was observed between the multidimensional data sets based on unidimensional intact tests (U60M, U40M). For customized data sets based on two-dimensional intact tests, the differences were much larger. Take absolute differences of percentile estimates for example. They were 1.92 versus 6.23 for the unidimensional data sets (U60U, U40U), 1.80 versus 7.77 for the multidimensional data sets based on unidimensional intact tests (U60M, U40M), and 3.61 versus 11.20, and 11.28 for the multidimensional data sets based on two-dimensional intact tests (M60M, M40M_Prop, M40M_Disp).

Test Length

Based on either unidimensional or multidimensional data sets, the lengthened 60-item customized data sets yielded more accurate results than the shortened 20-item data sets. When unidimensionality was satisfied, the results of the 60-item data set U60U_All were more accurate than those for the 20-item data set U40U_Part with average absolute differences of percentile estimates, 3.84 versus 7.38. When customizing was conducted based on two-dimensional intact tests, the results of the 60-item data set M60M_All were more accurate than those for the 20-item data set M40M_Prop_Part with the absolute differences of percentile estimates, 7.12 versus 9.16 (Table A-37).

The results of unidimensional data sets were more accurate than their multi-dimensional counterparts. The unidimensional 60-item data set U60U_All yielded more accurate results than the multidimensional data set M60M_All with the average absolute differences of percentile estimates being 3.84 versus 7.22. Similarly, the unidimensional 20-item data set U40U_Part yielded more accurate results than the multidimensional data

set M40M_Prop_Part with the average absolute differences of percentile estimates, 7.38 versus 9.16.

Summary

In this chapter, the properties of simulated data sets were described. The results of examinee proficiency estimates and percentile ranks based on customized data sets and intact tests were described. Four factors were manipulated to investigate the accuracy of customized norms, including customizing strategies, items used for normative estimation, correlations of latent abilities, and test dimensional structures.

There is a mixed picture about the accuracy of customized norms. On one hand, the manipulated conditions had little impact on some measures, including estimated ability distributions, mean differences of ability estimates and mean differences of percentile estimates. The normative estimates based on customized tests and intact tests were very similar based on the three measures. On the other hand, the manipulated conditions affected other measures, including correlations and average absolute differences of ability estimates and percentile estimates. As the magnitudes of the criterion measures were sometimes very large, the validity of the results should be very carefully considered. A more detailed summary of the results and their practical implications are presented in the next chapter.

CHAPTER V. CONCLUSTIONS AND IMPLICATIONS

Customizing standardized achievement tests has long been of interest because any single test is unlikely to measure all of the objectives specified by schools or states. It would appear to be to a user's advantage to customize those tests to better match local objectives. This raises the important question of whether the norms based on the standardized test are still relevant with the customized instrument. This study investigated the impact of several factors on the validity of customized norms. The manipulated factors included: 1) items used for normative estimation, 2) customizing strategy, 3) test length (number of items used to estimate the norms), 4) correlation of latent abilities, as well as 5) test dimensional structures. As it would have been very difficult to manipulate these factors using real data sets, Monte Carlo techniques were used in this study.

The procedures used to customize intact tests and derive normative estimates were outlined in Chapter III, and the results were presented in Chapter IV. In this chapter, the conclusions and the practical implications are discussed. The strengths and limitations of the present study are discussed, and recommendations for further research are also listed. Five sections are included in this chapter: (1) Findings, (2) Practical Implications, (3) Strengths and Limitations, (4) Recommendations for Future Studies, and (5) Conclusions.

Findings

The major conclusions related to the manipulated factors are described in this section.

Test Dimensional Structures

This study investigated the effects of various factors under different dimensional conditions depending on whether the unidimensionality assumption was satisfied. The unidimensional data sets were constructed by adding items measuring the same latent

ability. The multidimensional data sets were constructed by adding items measuring a different latent ability. Another level of dimensionality was created by adding a new dimension to two-dimensional intact tests.

Results across conditions conformed to patterns according to the complexity of the data structure. Results were most accurate for unidimensional data sets and least accurate for two-dimensional tests with a third dimension added. Specific examples of these findings are presented in the next few sections in conjunction with other manipulated factors.

Customizing Strategy

Generally speaking, for the unidimensional data sets, customizing based on all items from the intact test yielded more accurate normative estimates than customizing based on only some items from the intact test. The differences were observed in the correlations and average absolute differences of ability estimates, correlations and average absolute differences of percentile rank estimates. However, the estimated ability distributions and the mean differences of ability estimates and percentile estimates were very similar.

When the unidimensionality assumption was satisfied, the differences between data sets based on different customizing strategies (i.e., adding items to full-length or reduced-length intact tests) were trivial (U60U, U40U). When new items measuring a different latent ability were added to unidimensional intact tests, the differences increased (U60M, U40M). The differences increased further when two-dimensional intact tests were customized (M60M, M40M_Prop, M40M_Disp).

The above observations suggested a combined effect of customizing strategies and test dimensional structures. When unidimensionality was satisfied, the differences between customization strategies were very small. The differences increased when unidimensionality was violated and test dimensional structures became more

complicated. The differences were the largest when new items and items from intact tests measured three different latent abilities.

An interesting outcome was observed when two-dimensional intact tests were shortened differently in terms of the test dimensional structure. It seemed that whether intact tests were shortened proportionally or disproportionally (M40M_Prop, M40M_Disp) had very little impact, based on all outcome measures.

Items Used for Normative Estimation

Generally speaking, the normative estimates based on only the items from intact tests were more accurate than those based on all items in customized tests. Again, the differences were observed in the correlations and average absolute differences of ability estimates, correlations and average absolute differences of percentile estimates. The estimated ability distributions and the mean differences of ability estimates and percentile estimates were similar.

Similar to customizing strategies, the effects of using different estimating items were investigated under three dimensional structures: all items measuring the same latent ability (unidimensionality satisfied), adding items measuring a different latent ability to unidimensional intact tests, adding items measuring a different dimension to two-dimensional intact tests. When the unidimensionality assumption was satisfied, using all items in customized instruments or only the items from intact tests yielded similar results. When unidimensionality was violated by adding items measuring a different latent ability to unidimensional intact tests, the differences increased. The differences increased further when two-dimensional intact tests were customized.

The above observations suggested a combined effect of estimating items and test dimensional structures. When unidimensionality was satisfied, the differences in normative results based on items from intact tests or all items were small. The differences increased when unidimensionality was violated and test dimensional structures became

more complicated. The differences were the largest when two-dimensional intact tests were customized by adding items measuring a different latent ability.

Correlations of Latent Abilities

When items from intact tests and new items measured different latent abilities, the effects of the correlations of latent abilities depended on the dimensional structure of the intact tests. When intact tests were unidimensional, the correlations of latent abilities seemed to have a very small impact. The results were similar at different correlation levels (.0, .3, .6, .9). When intact tests were two-dimensional, the correlations of latent abilities had a larger impact. The higher the correlations of latent abilities, the more accurate were the normative estimates. Data sets with correlations of .60 or higher yielded rather accurate ability and normative estimates, but increased correlation still led to more accurate results (the results of .90 were more accurate than those of .60). Again the differences were observed in the correlations and average absolute differences of ability estimates and correlations and average absolute differences of percentile estimates. The estimated ability distributions and the mean differences of ability estimates and percentile estimates were again similar.

Test Length

Changing test length affected the accuracy of the normative estimates. The estimated norms based on fewer items were less accurate than those based on more items. Similar to the previous condition, the differences were observed in the correlations and average absolute differences of ability estimates and correlations and average absolute differences of percentile estimates. The estimated ability distributions and the mean differences of ability estimates and percentile estimates were similar.

Practical Implications

When normative information is derived from locally developed customized tests, the normative estimates may be used for group-based or individual inferences. Although the two kinds of inferences may be used independently depending on the information demands of local educators, they are both typically of interest. Based on the results of this study, there seems to be a mixed picture regarding the validity of these two inferences. Particular caution should be exercised if individual inferences are the major focus of test users.

Group-Based Inferences

When the focus is investigating the overall achievement status of a large group of examinees, three outcome measures reported here might be appropriate: the estimated ability distribution, the mean ability estimates and the mean percentile estimates. There seems to be largely random impact on group-level inferences for the investigated factors, at least based on the results of the three measures used here. The average ability and normative estimates based on customized tests and intact tests were very similar, which was consistent in all conditions in this study. The effect of overestimation and underestimation for ability estimates and percentile estimates appeared to cancel out each other, which yielded mean differences very close to zero. The estimated ability distributions based on intact tests and customized tests were very similar, and the overall distributions overlapped with each other at all selected percentiles ($P_{10,}P_{20,}P_{30,}P_{40,}P_{50,}P_{60,}P_{70},P_{80,}P_{90}$). Although extreme percentiles (P_1, P_5, P_{95}, P_{99}) were not addressed, examination of the graphs associated with these distributions showed that there were no obvious inconsistencies at those values. The results seemed to suggest that group-based normative inferences may be appropriate in the circumstances investigated in this study.

Individual Inferences

When the focus is comparing achievement status of individual examinees, four outcome measures were used to evaluate the accuracy of the individual normative estimates: the mean absolute differences of ability estimates, the mean absolute differences of percentile estimates, the correlations of ability estimates, and the correlations of percentile estimates. The magnitudes of these outcome measures changed significantly in various conditions, and the mean absolute differences of percentile rank estimates based on intact tests and customized tests were as high as 16. Although there are no universally accepted criteria for evaluating the measures used in this study, the results suggested that individual normative inferences should be used very cautiously in real customized testing situations.

Using Different Items to Estimate Normative Information

Customized tests typically involve replacing items from an intact test with new items measuring additional content domains not measured by the original test. As the customized instrument includes both new items and items from the original test, using only the items from the intact test for normative estimation appears to be important. This is especially true if 1) the unidimensionality assumption is violated, or 2) some items from the intact test are removed from customized tests (i.e., customizing situations similar to U40M, M40M_Prop and M40M_Disp). The normative estimates based on items from intact tests were more accurate than those based on all items.

This finding was consistent with the suggestion of Linn and Hambleton (1991): “additional content areas or extra coverage of content that is sparsely covered by the NRT (norm referenced test) may be added…, but they should not be part of the calculation of norm-referenced scores” (p. 204). Although other previous findings suggested accurate normative estimates may be obtained using both items from off-the-shelf tests and new

items (Dungan, 1988; Green, 1987), practitioners may want to exercise more caution in choosing an estimating procedure.

Customizing Strategy

When new items measuring additional content domains are added to customize off-the-shelf tests, different customizing strategies may be used regarding the combination of items from off-the-shelf tests and new items measuring additional content domains. If testing resources permit, it might be desirable to keep all items from off-the-shelf tests in customization. The full-length customized data sets would probably yield more accurate normative estimates than the reduced-length tests (i.e., removing some items from off-the-shelf tests). This was particularly important if the unidimensionality assumption was violated. This suggestion is consistent with the classification of customizing strategies by Linn and Hambleton (1991).

Although it may be ideal to administer all items from off-the-shelf tests, practical constraints often make it impossible. How to shorten an off-the-shelf test when the unidimensionality assumption is violated, therefore, is important. In this study, two-dimensional intact tests were shortened proportionally or disproportionally in terms of the test dimensional structure. Compared with the data sets including all items from off-the-shelf tests, the normative estimates derived from tests shortened either proportionally or disproportionally were less accurate. An interesting observation is that the estimates derived from the tests shortened either proportionally or disproportionally were very similar to each other. Findings in the literature suggest that customized tests should resemble off-the-shelf tests in test dimensional structure as much as possible. If the dimensional structures of the two tests are identical, it may be appropriate to make normative inferences based on the customized tests (e.g., Linn and Hambleton, 1991; Yen et al., 1987; Holmes, 1986, Lenke, 1989). The results of this study seem to suggest that reduced-length customized tests would yield less accurate normative estimates than full-

length customized tests even if the customized tests and off-the-shelf tests have the same dimensional structure. Similar findings were also reported by Way et al. (1989). In their study, they found that the normative estimates based on content representativecustomized tests were systematically different from those based on the off-the-shelf tests. It appears that extreme caution should be used even if the tests are shortened strictly based on the same table of specifications. This is especially relevant if the latent abilities are not highly correlated (assuming the items from off-the-shelf tests and new items measure different latent abilities). Note, however, the simulation design of this study may not completely reflect situations involving shortening an intact test based on content specifications. A study with real data would be required to insure realistic and generalizable results. Whether the conclusions of this study may be generalized needs to be further investigated.

Test Length

When off-the-shelf tests are customized, the customized tests may have different lengths than the original tests. The results of this study showed that changing test length had an impact on the validity of customized norms, which was true even if the unidimensionality assumption was satisfied. This suggested that the normative estimates derived from customized tests of different length should be used with caution. When the customized tests are much shorter than off-the-shelf tests, normative information for individuals should be used with extreme caution. This finding was consistent with Harris's results (1988, 1990), where sizable differences were observed for normative estimates based on off-the-shelf tests and customized tests of different lengths. She concluded that "test length, in and of itself, is a potent enough factor to make comparisons between total intact tests and shortened customized tests unwise" (1988, p.14).

Although maintaining the original length of the off -the-shelf test is ideal, the customized instrument will likely include a different number of items than the off-the-shelf tests. Generally speaking, increasing test length had less impact on the accuracy of the normative estimates than shortening tests (at least for the lengths examined in this study). Lengthened customized instruments provided more accurate normative estimates than shortened instruments, which was true whether or not the unidimensionality assumption was satisfied.

Correlations of Latent Abilities

When tests are calibrated based on Item Response Theory, unidimensional IRT models are often used. Although the unidimensionality assumption may be violated using real testing data, IRT calibrations based on unidimensional models may be robust for purposes like scaling and equating (e.g., Hirsh and Keene, 1989; Yen et al., 1987). This, however, may not be the case when drawing individual normative inferences is the focus.

When an off-the-shelf test is tailored towards local curriculum, new items are usually added to cover additional content domains not measured by the original test. When new items and the items from the off-the-shelf test measure different latent abilities, it is desirable that the latent abilities are highly correlated. The results of the study showed that higher correlations were associated with more accurate normative estimates based on the customized tests. If the correlation of the latent abilities was very low, individual normative estimates based on customized instruments were very different from the estimates based on off-the-shelf test. It may not, therefore, be appropriate to draw normative inferences in this situation. Note, however, the correlations among ability parameters are likely to be .6 or higher in most testing situations, and normative estimates were reasonably accurate in these situations.

Strengths and Limitations

Strengths of the Study

This study is part of an ongoing effort to investigate the validity of customized norms. Although a large number of studies have been conducted, this study makes additional contributions that might help practitioners administering customized tests. As only a small number of factors were investigated in previous studies, this study attempted to investigate the impact of various factors related to customized tests. Specifically, the following factors were investigated: (1) customizing strategy, (2) items used for normative estimation, (3) test length, (4) correlations of latent abilities (assuming new items and items from an off-the-shelf test measured different latent abilities), and (5) test dimensionality considerations. They represent potentially important factors when off-the-shelf tests are customized to assess local curricula.

Limitations

As is typically the case, this study does not explain all possible factors, and several limitations were unavoidable.

Data Sets

In this study, data sets were simulated using the descriptive statistics (means and standard deviations) of item parameter estimates (3-PL model) derived from a real data set. As is typically true of simulation studies, the representativeness and generalizability of the simulated item parameters is unknown. Additionally, the multidimensional data sets were simulated using a simple structure model, which might not be generalizable to more complex multidimensional situations.

Test Conditions and Generalization of Results

Implementing customized tests and estimating normative information are very complicated in a real testing environment. In addition to factors investigated in this study,

other factors that may affect the validity of customized norms may include: estimation methods (Bayesian versus MLE), context effects, administration modes, item positions, and the effect of speededness. Although the simulation design of this study aimed to mimic real testing situations, it merely represented some complications that may arise in reality. The conclusions of this study are restricted to the limited conditions. Whether the results might be generalized to more complicated testing situations needs to be investigated.

Recommendations for Future Studies

Investigating the validity of customized norms in real testing environments is complicated as many factors may affect the accuracy of the normative inferences. This study represents a small part of the continuing research in this area. As shown in the previous section, there were some limitations of the present study in terms of its simulation procedures and research design. Further studies could be conducted by taking these limitations into consideration.

Data sets

There are several ways to make the simulated data sets more representative and comprehensive. For example, real item parameter estimates from a variety of different types of tests (e.g. Reading, Math, and Science) might be used to simulate item parameters. In addition to means and standard deviations of item parameter estimates, other factors might be incorporated, including different numbers of items, different ratios of new items and items from off-the-shelf tests, and changing item positions. To create multidimensional data sets, more complicated multidimensional models might be used. It would be interesting to investigate whether the results of this study could be replicated in different settings.

Test conditions

More variations of customizing strategies and test length constraints should be investigated in future studies. In this study, the ratios of the items from an off-the-shelf test and new items were 2:1 (40 old items and 20 new items) and 1:1 (20 old items and 20 new items) based on a 40-item off-the-shelf test. In future studies, different ratios should be investigated. Other factors that might be of interest are ability estimation procedures (MLE, Bayes), unidimensional IRT models (1-PL, 3-PL), and context effects.

Real data analysis

As is common to all simulation studies, investigating whether the results may be replicated using real data sets is always of interest for future studies. It would be useful to see whether the conclusions of the present study can be generalized to real data situations.

Conclusions

The practice of test customization will likely continue to increase in educational settings. With growing accountability pressures, schools will be looking for ways to limit testing time and expense while still assessing their educational standards and objectives. Customizing a nationally standardized test offers the possibility of obtaining national normative information while measuring only the objectives relevant to a particular school or state. This study has offered some evidence that the test customization process might not always yield accurate estimates of examinees' abilities. If the users of customized tests are primarily interested in aggregated score reporting, the normative estimates based on customized tests and intact tests tend to be consistent based on the results of this study. If practitioners are primarily interested in reporting normative estimates for individuals, the estimates based on customized tests could be very different from those based on intact tests depending on how an intact test is customized. Practitioners should be fully aware of the risks of using customized norms; at the very least, they should use

caution when making customization decisions. More research is needed to address the areas of concern noted in this study.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113-127.

Allen, N. A., Ansley, T. N., & Forsyth, R. A. (1987). The effects of deleting content-related items on IRT ability parameters. *Educational and Psychological Measurement, 47,* 1141-1152.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two dimensional data. *Applied Psychological Measurement, 9,* 39-48.

Ansley, T. N., Forsyth, R. A., & Hoover, H. D. (1989, March). *Test customization: Can we have our cake and eat it too?* Paper presented at the annual meeting of the National council on Measurement in Education, San Francisco.

Arenson, K. A. (2006). Testing errors prompt calls for oversight. *New York Times,* March 18, 2006.

Bock, R. D., & Mislevy, R. J. (1981). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological measurement, 6,*431-444.

Camilli, G. (1988). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement, 36(1),* 73-78.

DeAyala, R. J. (1992). The influence of dimensionality on CAT ability estimation. *Educational and Psychological Measurement, 52,* 513-528.

Doody-Bogan, E. N. & Yen, W. M. (1983). *Detecting multidimensionality and examining its effects on vertical equating with the three parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Doody, E. N. (1985). *Examining the effects of multidimensional data on ability and item parameter estimation using the three parameter logistic model*. Paper presented at the annual meeting of the America Educational Research Association, Chicago.

Dungan, L. A. (1988, April). *Norm-referenced test customization: Validation of individual score interpretations*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13,* 373-389.

Forsyth, R. A., Twing. J. S., & Ansley, T. N. (1992). Three applications of customized testing in local school districts. *Applied Measurement in Education. 5(2),* 111-22.

Green, D. R. (1987, April). *Local versus national calibrations*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Hambleton, R. K. (Eds), *Handbook of modern item response theory*. New York: Springer.

Hambleton, R. K., Gower, C., & Rogers, H. J. (1989, March). *Customized norm-referenced testing: A review of issues and methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Harris, D. J. (1987, April). *Estimating examinee achievement using a customized test*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Harris, D. J. (1988, April). *An examination of the effect of test length on customized testing using item response theory*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Harris, D. J. (1990, April). *Applying national norms to a lengthened customized test*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Hattie, J. (1981). *Decision criteria for determining unidimensionality*. Unpublished doctoral dissertation. University of Toronto.

Hieronymous, A. N., Hoover, H. D., & Lindquist, E. F. (1985). *Iowa tests of basic skills*, Forms G and H. Chicago: Riverside Publishing.

Hirsch, T. M., & Keene, J. M. (1989, March). *An examination of the effects different dimensional test structures have on test equating*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Holmes, S. E. (1986, April). *Multi-purpose tests: A solution to test proliferation*. Paper presented at the annual meting of the National Council on Measurement in Education, San Francisco.

Kolen, M.J. & Brennan, R. L. (2004). *Testing equating, scaling and linking*. NY: Springer.

Lau, A. C. (1996). *Robustness of a unidimensional computerized testing mastery procedure with multidimensional testing data*. Unpublished doctoral dissertation. University of Iowa.

Lenke, J. M. (1989, March). *Norm-referenced scores for customized tests: Issues and applications*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Linn, R. L. & Hambleton. (1991). Customized test and customized norms. *Applied Measurement in Education, 4(3),* 185-207.

McKinley, R. L.,& Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space*. American College Testing Programs: Research Report ONR83-2.

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 models.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference*. Minneapolis: University of Minnesota Department of Psychology, Computerized Adaptive Testing Laboratory.

SAS/IML (version 6) [Computer programming language] (1990). Cary, NC: SAS Institute.

Toit, M. D. (Eds). (2002). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Chicago: Scientific Software International, Inc.

Tong, Y. & Kolen, M. J. (2006). *Vertical scaling and scaling shrinkage*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Tong, Y. & Kolen, M. J. (in press). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*.

Wang, M. M. (1985). *Fitting a unidimensional model to multidimensional item response theory data: The effects of latent space misspecification on the application of IRT*. Unpublished manuscript.

Way, W.D., Forsyth, R. A., & Ansley T. N. (1989). IRT ability estimates from customized achievement tests without content sampling. *Applied Measurement in Education, 2*. 15-35.

Weiss D. J., & Suhadolnik, D. (1982). Robustness of adaptive testing to multidimensionality. In D. J. Weiss (Ed.), *Item Response Theory and Computerized Adaptive Testing Conference proceedings* (pp. 248-280). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Wilson, S. M., & Hiscos, M. D. (1984). Using standardized tests for assessing local learning objectives. *Educational Measurement: Issues and Practice*, *3*, 19-22.

Wright, B. D. (1968), Sample free test calibration and person measurement. *Proceedings of the 1967 ETS invitational conference on testing problems*. (pp. 85-101). Princeton, NJ: Educational Testing Service.

Yen, W. M. (1985). Increasing item complexity: a possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*, 399-410.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23(4)*, 299-325.

Yen, W. M., Green, E. R., & Burket, G. R. (1987). Valid normative information from customized achievement tests. *Educational Measurement: Issues and practice. 6*, 7-13.

Zhao, J. C., McMorris, R. F., & Chen, R. (2002, April). *The robustness of the unidimensional 3PL IRT model when applied to two-dimensional data in computerized adaptive testing*. Paper presented at the 2002 Annual Meeting of the American Educational Research Association, New Orleans, LA.

Zimowski, M. F., Muraki, E., Mislevey, R. J., & Bock. R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer program]. Chicago: Scientific Software International, Inc.

APPENDIX. TABLES

Table A-1. Item Parameters Used to Simulate Unidimensional and Multidimensional Data

	Data Set Abbreviation	Dimensionality (Item No.)		Dimension I	Dimension II	Dimension III
Items from Intact Tests	U_Intact	D1 (40)	Discrimination	Lognormal E(x)=1.04 Std(x)=.40		
			Difficulty	Normal E(x) = -.25 Std(x) =1.03		
New Items	U60U U40U	D1 (20)	Discrimination	Lognormal E(x)=1.04 Std(x)=.40		
			Difficulty	Normal E(x)= -.25 Std(x)=1.03		
	U60M U40M	D2 (20)	Discrimination		Lognormal E(x)=1.04 Std(x)=.40	
			Difficulty		Normal E(x)= -.25 Std(x)=1.03	
Items from Intact Tests	M_Intact	D1(20)	Discrimination	Lognormal E(x)=1.04 Std(x)=.40		
			Difficulty	Normal E(x)= -.25 Std(x)=1.03		
		D2(20)	Discrimination		Lognormal E(x)=1.04 Std(x)=.40	
			Difficulty		Normal E(x)= -.25 Std(x)=1.03	
New Items	M60M M40M_Prop M40M_Disp	D3(20)	Discrimination			Lognormal E(x)=1.04 Std(x)=.40
			Difficulty			Normal E(x)= -.25 Std(x)=1.03

Note: E(x) refers to the expected value of a distribution and Std (x) refers to the standard deviation of a distribution.

Table A-2. Summary Statistics of Unidimensional Item Parameter Estimates for Intact Tests and Customized Data Sets

			U_ Intact	U60U	U40U	U60M	U40M	M_ Intact	M60M	M40M _Prop	M40M _Disp
a	Mean	Mean	1.14	1.12	1.15	0.95	0.94	0.95	0.84	0.90	0.88
		Std. Err.	0.06	0.05	0.06	0.09	0.13	0.12	0.14	0.15	0.15
	Std	Mean	0.35	0.35	0.36	0.39	0.37	0.35	0.33	0.32	0.30
		Std. Err.	0.11	0.09	0.11	0.10	0.12	0.16	0.14	0.14	0.10
b	Mean	Mean	-0.26	-0.25	-0.23	-0.27	-0.26	-0.26	-0.26	-0.25	-0.25
		Std. Err.	0.16	0.12	0.16	0.17	0.20	0.20	0.17	0.20	0.21
	Std	Mean	0.94	0.98	0.94	1.36	1.28	1.25	1.40	1.29	1.35
		Std. Err.	0.13	0.12	0.13	0.43	0.39	0.39	0.48	0.48	0.50
c	Mean	Mean	0.19	0.19	0.18	0.21	0.21	0.22	0.22	0.21	0.20
		Std. Err.	0.01	0.01	0.01	0.02	0.03	0.03	0.03	0.03	0.02
	Std	Mean	0.07	0.06	0.06	0.07	0.07	0.08	0.07	0.07	0.07
		Std. Err.	0.01	0.01	0.01	0.01	0.02	0.03	0.02	0.02	0.02

Note:

The summary statistics were averaged across different correlations of latent abilities for each multidimensional data set. Note also that all items were used to compute the summary statistics.

Table A-3. Unidimensional Item Parameter Estimates of Multidimensional Data Sets

Theta Corr.	Para.			U40M	U60M	M_Intact	M60M	M40M_ Prop	M40M_ Disp
.00	a	Mean	Mean	0.86	0.83	0.92	0.73	0.80	0.73
			Std. Err.	0.05	0.12	0.14	0.08	0.15	0.10
		Std	Mean	0.48	0.49	0.55	0.51	0.44	0.35
			Std. Err.	0.08	0.10	0.15	0.12	0.17	0.12
	b	Mean	Mean	-0.27	-0.30	-0.28	-0.26	-0.26	-0.29
			Std. Err.	0.24	0.28	0.29	0.24	0.31	0.32
		Std	Mean	1.99	1.84	1.80	2.15	1.96	2.08
			Std. Err.	0.30	0.29	0.31	0.25	0.46	0.44
	c	Mean	Mean	0.24	0.25	0.25	0.26	0.24	0.22
			Std. Err.	0.01	0.03	0.01	0.01	0.03	0.02
		Std	Mean	0.09	0.10	0.11	0.10	0.10	0.10
			Std. Err.	0.01	0.02	0.03	0.02	0.03	0.02
.30	a	Mean	Mean	0.89	0.86	0.86	0.74	0.81	0.79
			Std. Err.	0.05	0.08	0.09	0.06	0.08	0.06
		Std	Mean	0.42	0.34	0.30	0.25	0.27	0.26
			Std. Err.	0.07	0.11	0.11	0.09	0.10	0.09
	b	Mean	Mean	-0.28	-0.26	-0.25	-0.24	-0.24	-0.24
			Std. Err.	0.16	0.18	0.18	0.15	0.17	0.17
		Std	Mean	1.37	1.25	1.18	1.33	1.23	1.29
			Std. Err.	0.19	0.16	0.18	0.14	0.18	0.15
	c	Mean	Mean	0.21	0.21	0.22	0.22	0.21	0.20
			Std. Err.	0.01	0.02	0.02	0.01	0.02	0.02
		Std	Mean	0.07	0.07	0.08	0.07	0.07	0.07
			Std. Err.	0.01	0.02	0.02	0.02	0.02	0.02
.60	a	Mean	Mean	0.96	0.96	0.95	0.86	0.92	0.91
			Std. Err.	0.05	0.06	0.05	0.04	0.05	0.05
		Std	Mean	0.35	0.31	0.26	0.23	0.25	0.25
			Std. Err.	0.07	0.09	0.06	0.05	0.08	0.08
	b	Mean	Mean	-0.26	-0.26	-0.26	-0.26	-0.25	-0.25
			Std. Err.	0.13	0.17	0.16	0.13	0.15	0.16
		Std	Mean	1.10	1.07	1.05	1.13	1.06	1.08
			Std. Err.	0.11	0.13	0.18	0.14	0.14	0.13
	c	Mean	Mean	0.20	0.20	0.20	0.20	0.19	0.19
			Std. Err.	0.01	0.01	0.01	0.01	0.01	0.01
		Std	Mean	0.06	0.07	0.07	0.06	0.06	0.06
			Std. Err.	0.01	0.01	0.01	0.01	0.01	0.01
.90	a	Mean	Mean	1.06	1.09	1.08	1.04	1.08	1.09
			Std. Err.	0.06	0.06	0.06	0.05	0.06	0.06
		Std	Mean	0.33	0.33	0.31	0.30	0.31	0.32
			Std. Err.	0.08	0.10	0.07	0.06	0.09	0.09
	b	Mean	Mean	-0.26	-0.25	-0.26	-0.26	-0.25	-0.24
			Std. Err.	0.13	0.16	0.15	0.12	0.14	0.14
		Std	Mean	0.99	0.96	0.96	1.00	0.93	0.95
			Std. Err.	0.10	0.12	0.15	0.12	0.12	0.10
	c	Mean	Mean	0.19	0.19	0.19	0.19	0.19	0.19
			Std. Err.	0.01	0.01	0.01	0.01	0.01	0.01
		Std	Mean	0.07	0.07	0.06	0.06	0.06	0.06
			Std. Err.	0.01	0.01	0.01	0.01	0.01	0.01

Table A-4. Estimated Ability Distributions of the Data Sets Based on the Unidimensional Intact Test

Dimensional Structures	Correlations of Latent Abilities	Data Sets	Percentiles								
			10	20	30	40	50	60	70	80	90
Intact test		U_Intact	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.51	0.82	1.26
1-D Data Sets		U60U_All	-1.25	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.82	1.26
		U40U_All	-1.23	-0.78	-0.47	-0.22	0.02	0.26	0.51	0.81	1.26
		U40U_Part	-1.23	-0.77	-0.47	-0.21	0.03	0.27	0.52	0.83	1.30
2-D Data Sets	**r=.00**	U60M_All	-1.25	-0.80	-0.48	-0.22	0.02	0.26	0.52	0.83	1.27
		U40M_All	-1.27	-0.80	-0.48	-0.21	0.03	0.28	0.53	0.83	1.26
	r=.30	U60M_All	-1.26	-0.81	-0.48	-0.22	0.03	0.27	0.53	0.83	1.28
		U40M_All	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.81	1.26
	r=.60	U60M_All	-1.26	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.83	1.27
		U40M_All	-1.23	-0.79	-0.48	-0.22	0.01	0.26	0.51	0.81	1.27
	r=.90	U60M_All	-1.25	-0.81	-0.49	-0.23	0.02	0.26	0.52	0.83	1.26
		U40M_All	-1.23	-0.78	-0.48	-0.22	0.02	0.26	0.51	0.82	1.26
	Across Correlations	U60M_All	-1.25	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.83	1.27
		U40M_All	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26

Note:

1. This table lists customized data sets based on the unidimensional intact test: U60U, U40U, U60M, and U40M.

2. The suffix _All is used if all items were used to estimate normative information, and __Part is used if only items from the intact test were used.

3. The results of U60U_Part, U60M_Part, and U40M_Part are not listed in the table as they included the same items as intact tests or existing entities in the table, which, therefore, yielded the same results by design. Specifically U60U_Part and U60M_Part are the same as U_Intact and U40M_Part is the same as U40U_Part.

Table A-5. Estimated Ability Distributions for the Two-dimensional Intact Test and Related Customized Tests

Correlations of Latent Abilities	Data Sets	Percentiles								
		10	20	30	40	50	60	70	80	90
r=.00	M_Intact	-1.26	-0.81	-0.49	-0.22	0.02	0.27	0.54	0.84	1.28
	M60M_All	-1.25	-0.81	-0.49	-0.22	0.02	0.27	0.54	0.85	1.28
	M40M_Prop_All	-1.22	-0.80	-0.49	-0.22	0.02	0.25	0.50	0.81	1.27
	M40M_Prop_Part	-1.26	-0.78	-0.46	-0.21	0.02	0.26	0.52	0.82	1.25
	M40M_Disp_All	-1.24	-0.80	-0.48	-0.22	0.03	0.27	0.53	0.83	1.26
	M40M_Disp_Part	-1.28	-0.76	-0.45	-0.19	0.03	0.26	0.53	0.85	1.31
r=30	M_Intact	-1.27	-0.80	-0.48	-0.21	0.03	0.26	0.51	0.82	1.26
	M60M_All	-1.25	-0.80	-0.47	-0.21	0.03	0.27	0.52	0.82	1.26
	M40M_Prop_All	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.51	0.82	1.26
	M40M_Prop_Part	-1.24	-0.79	-0.47	-0.21	0.02	0.25	0.49	0.79	1.23
	M40M_Disp_All	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.51	0.81	1.27
	M40M_Disp_Part	-1.27	-0.79	-0.46	-0.20	0.03	0.26	0.51	0.80	1.25
r=.60	M_Intact	-1.25	-0.80	-0.49	-0.23	0.01	0.25	0.51	0.81	1.26
	M60M_All	-1.25	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.82	1.28
	M40M_Prop_All	-1.23	-0.78	-0.48	-0.23	0.01	0.25	0.50	0.81	1.27
	M40M_Prop_Part	-1.23	-0.77	-0.46	-0.21	0.02	0.24	0.49	0.79	1.27
	M40M_Disp_All	-1.24	-0.78	-0.49	-0.22	0.01	0.25	0.51	0.81	1.27
	M40M_Disp_Part	-1.24	-0.78	-0.47	-0.21	0.03	0.25	0.51	0.81	1.28
r=.90	M_Intact	-1.24	-0.79	-0.48	-0.22	0.01	0.25	0.51	0.82	1.26
	M60M_All	-1.25	-0.80	-0.49	-0.22	0.01	0.26	0.52	0.82	1.27
	M40M_Prop_All	-1.23	-0.79	-0.48	-0.22	0.02	0.26	0.51	0.80	1.26
	M40M_Prop_Part	-1.22	-0.78	-0.47	-0.21	0.02	0.25	0.50	0.80	1.27
	M40M_Disp_All	-1.23	-0.79	-0.48	-0.22	0.01	0.25	0.51	0.81	1.26
	M40M_Disp_Part	-1.23	-0.78	-0.47	-0.22	0.02	0.25	0.51	0.81	1.30
Across Correlations	M_Intact	-1.25	-0.80	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	M60M_All	-1.25	-0.80	-0.49	-0.22	0.02	0.27	0.53	0.83	1.27
	M40M_Prop_All	-1.23	-0.79	-0.48	-0.22	0.02	0.25	0.51	0.81	1.27
	M40M_Prop_Part	-1.24	-0.78	-0.47	-0.21	0.02	0.25	0.50	0.80	1.26
	M40M_Disp_All	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.51	0.81	1.26
	M40M_Disp_Part	-1.26	-0.78	-0.46	-0.21	0.03	0.26	0.51	0.82	1.28

Note: The results of M60M_Part are not listed as it included the same items as the intact test M_Intact, and therefore yielded the same results as M_Intact by design.

Table A-6. Average Proportion Correct Scores for Data Sets Representing Intact and Customized Tests

Data Set	Across Correlations	r = .00	r = .30	r = .60	r = .90
U_Intact	0.64				
U60U_All	0.63				
U40U_All	0.61				
U40U_Part	0.64				
U60M_All	0.62	0.62	0.62	0.62	0.62
U40M_All	0.61	0.61	0.61	0.61	0.61
M_Intact	0.64	0.64	0.64	0.64	0.64
M60M_All	0.64	0.64	0.64	0.64	0.64
M40M_Prop_All	0.60	0.60	0.60	0.60	0.60
M40M_Prop_Part	0.63	0.63	0.63	0.63	0.63
M40M_Disp_All	0.61	0.61	0.61	0.61	0.61
M40M_Disp_Part	0.64	0.64	0.64	0.64	0.64

Note: The data sets U60U_Part, U60M_Part, M60M_Part and M40M_Part are not listed as they included the same items as intact tests or existing entities in the table, which, therefore, yielded the same results by design. Specifically, U60U_Part and U60M_Part are the same as U_Intact, M60M_Part the same as the M_Intact, and U40M_Part the same as U40U_Part.

Table A-7. Average Reliability Estimates for Data Sets Representing Intact and Customized Tests

	Across Correlations	r=.00	r=.30	r=.60	r=.90
U_Intact	0.90				
U60U_All	0.93				
U40U_All	0.90				
U40U_Part	0.81				
U60M_All	0.90	0.86	0.89	0.91	0.92
U40M_All	0.84	0.79	0.83	0.87	0.89
M_Intact	0.85	0.79	0.84	0.87	0.89
M60M_All	0.87	0.79	0.86	0.90	0.92
M40M_Prop_All	0.83	0.74	0.81	0.86	0.89
M40M_Prop_Part	0.73	0.65	0.72	0.77	0.80
M40M_Disp_All	0.83	0.75	0.82	0.86	0.89
M40M_Disp_Part	0.76	0.71	0.75	0.78	0.80

Table A-8. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Sets Using Different Estimation Items under Unidimensionality

Percentile		10	20	30	40	50	60	70	80	90
Unidimensional_All	Mean	-1.25	-0.79	-0.48	-0.21	0.03	0.26	0.52	0.82	1.26
	Std. Err.	(0.05	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04)
Unidimensional_Part	Mean	-1.26	-0.78	-0.47	-0.20	0.03	0.27	0.52	0.83	1.26
	Std. Err.	(0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06)
U_Intact	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
Percentile Difference										
Unidimensional_All	Mean	-0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00
	Std. Err.	(0.05	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.05)
Unidimensional_Part	Mean	-0.02	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.00
	Std. Err.	(0.06	0.04	0.03	0.04	0.04	0.03	0.03	0.03	0.04)

Note:

1. Unidimensional_All refers to the average of U60U_All and U40U_All, and Unidimensional_Part refers to the average of U60U_Part and U40U_Part.

2. The intact test, U_Intact, is listed for comparison in the top part of the table. In the bottom part of the table, differences between percentiles for U_Intact and each of the other two data sets are given. The same procedure is used in the following tables describing estimated ability distributions.

3. The standard errors of statistics are computed as the empirical standard deviations of the statistics based on the simulated data sets. The same notation is used in all tables.

Table A-9. Comparing Ability Estimates and Percentile Estimates Derived from Unidimensional Data Sets Using Different Items in Estimation

	Customized Data Sets	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	Unidimensional_All	0.99	(0.00)	0.00	(0.01)	0.20	(0.01)
	Unidimensional_Part	0.99	(0.00)	0.00	(0.01)	0.15	(0.01)
Percentile	Unidimensional_All	0.99	(0.00)	0.04	(0.26)	5.00	(0.36)
	Unidimensional_Part	0.99	(0.00)	0.22	(0.31)	3.69	(0.41)

Table A-10. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Sets Representing Different Customizing Strategies under Unidimensionality

Percentile		10	20	30	40	50	60	70	80	90
U60U	Mean	-1.25	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.27
	Std. Err.	(0.04	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.04)
U40U	Mean	-1.26	-0.78	-0.46	-0.20	0.04	0.27	0.53	0.83	1.26
	Std. Err.	(0.08	0.05	0.06	0.06	0.06	0.05	0.04	0.05	0.07)
U_Intact	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
Percentile Difference										
U60U	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Std. Err.	(0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.02)
U40U	Mean	-0.02	0.01	0.02	0.03	0.02	0.02	0.01	0.01	0.00
	Std. Err.	(0.09	0.06	0.05	0.06	0.06	0.05	0.04	0.04	0.06)

Note: U60U is the average of U60U_Part and U60U_All, and U40U is the average of U40U_Part and U40U_All.

Table A-11. Comparing Ability Estimates and Percentile Estimates Derived from Data Sets Representing Different Customizing Strategies under Unidimensionality

	Customized Data Sets	Correlation	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U60U	1.00	(0.00)	0.00	(0.00)	0.08	(0.01)
	U40U	0.98	(0.00)	0.00	(0.01)	0.25	(0.02)
Percentile	U60U	1.00	(0.00)	-0.03	(0.10)	1.92	(0.23)
	U40U	0.99	(0.00)	0.29	(0.49)	6.23	(0.57)

Table A-12. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Sets Representing Tests of Different Lengths under Unidimensionality

Percentile			10	20	30	40	50	60	70	80	90
Data Sets	Items										
U40U_Part	20	Mean	-1.27	-0.78	-0.45	-0.19	0.05	0.28	0.53	0.83	1.27
		Std. Err.	(0.12	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.09)
U_Intact	40	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
		Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
U60U_All	60	Mean	-1.25	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.82	1.27
		Std. Err.	(0.05	0.03	0.03	0.03	0.02	0.02	0.03	0.03	0.04)
Percentile Difference											
U40U_Part	20	Mean	-0.03	0.02	0.03	0.04	0.03	0.02	0.02	0.01	0.00
		Std. Err.	(0.12	0.07	0.07	0.07	0.07	0.07	0.05	0.06	0.09)
U60U_All	60	Mean	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
		Std. Err.	(0.05	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.05)

Table A-13. Comparing Ability Estimates and Percentile Estimates Derived from Data Sets Representing Tests of Different Lengths under Unidimensionality

	Cust. Data Sets	Item No.	Correlation (Cus.- Intact)	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus.–Intact|	Std. Err.
Ability	U40U_Part	20	0.97	(0.01)	0.00	(0.01)	0.30	(0.03)
	U60U_All	60	0.99	(0.00)	0.00	(0.01)	0.15	(0.01)
Percentile	U40U_Part	20	0.98	(0.01)	0.44	(0.63)	7.38	(0.82)
	U60U_All	60	0.99	(0.00)	-0.07	(0.20)	3.84	(0.46)

Note. The statistics for a given data set are related / compared to those for the 40-item intact tests.

Table A-14. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Sets Representing the Combined Effects of Customizing Strategies and Estimation Items under Unidimensionality

Percentile		10	20	30	40	50	60	70	80	90
U40U_All	Mean	-1.24	-0.78	-0.47	-0.20	0.03	0.27	0.52	0.82	1.26
	Std. Err.	(0.07	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.05)
U40U_Part	Mean	-1.27	-0.78	-0.45	-0.19	0.05	0.28	0.53	0.83	1.27
	Std. Err.	(0.12	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.09)
U60U_All	Mean	-1.25	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.82	1.27
	Std. Err.	(0.05	0.03	0.03	0.03	0.02	0.02	0.03	0.03	0.04)
U_Intact /U60U_Part	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
Percentile Difference										
U40U_All	Mean	0.00	0.01	0.01	0.02	0.01	0.01	0.00	0.00	-0.01
	Std. Err.	(0.07	0.05	0.04	0.05	0.05	0.04	0.04	0.04	0.06)
U40U_Part	Mean	-0.03	0.02	0.03	0.04	0.03	0.02	0.02	0.01	0.00
	Std. Err.	(0.12	0.07	0.07	0.07	0.07	0.07	0.05	0.06	0.09)
U60U_All	Mean	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
	Std. Err.	(0.05	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.05)
U60U_Part	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Std. Err.	(.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00)

Note: The data set U60U_Part includes the same set of items as the unidimensional intact test, thus yielding the same results as U_Intact by design.

Table A-15. Comparing Ability Estimates and Percentile Estimates Derived from Data Sets Representing the Combined Effects of Customizing Strategies and Estimation Items under Unidimensionality

	Customized Data Sets	Correlation	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U40U_All	0.98	(0.00)	0.00	(0.01)	0.27	(0.01)
	U40U_Part	0.97	(0.01)	0.00	(0.01)	0.30	(0.03)
	U60U_All	0.99	(0.00)	0.00	(0.01)	0.15	(0.01)
	U60U_Part	1.00	(0.00)	0.00	(0.00)	0.00	(0.00)
Percentile	U40U_All	0.98	(0.00)	0.15	(0.42)	6.66	(0.40)
	U40U_Part	0.98	(0.01)	0.44	(0.63)	7.38	(0.82)
	U60U_All	0.99	(0.00)	-0.07	(0.20)	3.84	(0.46)
	U60U_Part	1.00	(0.00)	0.00	(0.00)	0.00	(0.00)

Table A-16. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Multidimensional Data Sets Using Different Items to Estimate Norms

Percentile		10	20	30	40	50	60	70	80	90
U_All	Mean	-1.25	-0.79	-0.48	-0.22	0.03	0.27	0.52	0.83	1.26
	Std. Err.	(0.05	0.03	0.04	0.03	0.04	0.03	0.03	0.03	0.05)
U_Part	Mean	-1.26	-0.78	-0.47	-0.20	0.03	0.27	0.52	0.83	1.26
	Std. Err.	(0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06)
U_Intact	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
M_All	Mean	-1.25	-0.79	-0.48	-0.21	0.03	0.26	0.52	0.82	1.26
	Std. Err.	(0.05	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04)
M_Part	Mean	-1.26	-0.78	-0.46	-0.20	0.03	0.27	0.52	0.82	1.26
	Std. Err.	(0.06	0.04	0.04	0.04	0.04	0.04	0.05	0.04	0.06)
M_Intact	Mean	-1.26	-0.80	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06)
Percentile Difference										
U_All	Mean	-0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.01	-0.01
	Std. Err.	(0.06	0.04	0.04	0.03	0.04	0.03	0.03	0.03	0.05)
U_Part	Mean	-0.02	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.00
	Std. Err.	(0.06	0.04	0.03	0.04	0.04	0.03	0.03	0.03	.04)
M_All	Mean	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	Std. Err.	(0.06	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.05)
M_Part	Mean	-0.01	0.02	0.02	0.02	0.01	0.00	0.00	0.00	0.00
	Std. Err.	(0.05	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.05)

Note:

U_All is the average of U60M_All and U40M_All, and U_Part is the average of U60M_Part and U40M_Part.

M_All is the average of M60M_All, M40M_Prop_All, M40M_Disp_All, and M_Part is the average of M60M_Part, M40M_Prop_ Part, M40M_Disp_ Part.

Table A-17. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Using Different Items to Estimate Normative Information

	Customized Data Sets	Correlation	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U_All	0.95	(0.01)	0.00	(0.01)	0.26	(0.03)
	U_Part	0.99	(0.00)	0.00	(0.01)	0.15	(0.01)
	M_All	0.82	(0.05)	0.00	(0.01)	0.49	(0.07)
	M_Part	0.98	(0.01)	-0.01	(0.01)	0.22	(0.03)
Percentile	U_All	0.97	(0.01)	0.09	(0.34)	6.43	(0.83)
	U_Part	0.99	(0.00)	0.22	(0.31)	3.69	(0.41)
	M_All	0.82	(0.06)	0.06	(0.26)	13.33	(2.04)
	M_Part	0.98	(0.01)	0.04	(0.27)	5.55	(1.01)

Table A-18. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Multidimensional Data Sets Representing Different Customizing Strategies

Percentile		10	20	30	40	50	60	70	80	90
U60M	Mean	-1.25	-0.8	-0.48	-0.22	0.02	0.26	0.52	0.82	1.27
	Std. Err.	(0.04	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.04)
U40M	Mean	-1.26	-0.78	-0.46	-0.2	0.04	0.27	0.53	0.83	1.26
	Std. Err.	(0.09	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.07)
U_Intact	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
M60M	Mean	-1.26	-0.8	-0.48	-0.22	0.03	0.27	0.52	0.82	1.26
	Std. Err.	(0.04	0.04	0.03	0.03	0.03	0.03	0.04	0.03	0.05)
M40M_Prop	Mean	-1.26	-0.78	-0.46	-0.2	0.04	0.27	0.52	0.81	1.25
	Std. Err.	(0.08	0.06	0.05	0.05	0.05	0.05	0.05	0.04	0.06)
M40M_Disp	Mean	-1.25	-0.78	-0.47	-0.21	0.03	0.26	0.51	0.82	1.26
	Std. Err.	(0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06)
M_Intact	Mean	-1.26	-0.8	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06)
Percentile Difference										
U60M	Mean	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	Std. Err.	(0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.02
U40M	Mean	-0.02	0.01	0.02	0.03	0.02	0.02	0.01	0.01	-0.01)
	Std. Err.	(0.10	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.07
M60M	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00)
	Std. Err.	(0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.02	0.02
M40M_Prop	Mean	-0.01	0.02	0.02	0.02	0.01	0.00	0.00	-0.01	-0.01)
	Std. Err.	(0.08	0.06	0.05	0.05	0.05	0.05	0.05	0.04	0.07
M40M_Disp	Mean	0.01	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.00
	Std. Err.	(0.06	0.04	0.04	0.03	0.03	0.03	0.03	0.04	0.06)

Table A-19. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing Different Customizing Strategies

	Customized Data Sets	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U60M	1.00	(0.00)	0.00	(0.00)	0.08	(0.01)
	U40M	0.96	(0.01)	0.00	(0.01)	0.31	(0.03)
	M60M	0.98	(0.01)	0.00	(0.00)	0.14	(0.03)
	M40M_Prop	0.89	(0.03)	0.00	(0.01)	0.41	(0.04)
	M40M_Disp	0.88	(0.05)	0.00	(0.01)	0.42	(0.08)
Percentile	U60M	1.00	(0.00)	0.00	(0.11)	1.80	(0.24)
	U40M	0.97	(0.01)	0.31	(0.57)	7.77	(0.86)
	M60M	0.99	(0.01)	0.07	(0.10)	3.61	(0.74)
	M40M_Prop	0.90	(0.03)	0.10	(0.43)	11.20	(1.28)
	M40M_Disp	0.87	(0.06)	-0.03	(0.34)	11.28	(2.55)

Table A-20. Selected Percentiles of Estimated Ability Distributions Derived from Selected Multidimensional Data Sets Representing Tests of Different Length

Percentile		10	20	30	40	50	60	70	80	90
Data Sets	Item #									
M40M_Prop_Part	20	-1.28	-0.77	-0.45	-0.19	0.05	0.28	0.52	0.81	1.25
		(0.10	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.10)
M_Intact	40	-1.26	-0.80	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
		(0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06)
M60M_All	60	-1.26	-0.80	-0.48	-0.22	0.03	0.27	0.53	0.83	1.27
		(0.05	0.04	0.04	0.04	0.03	0.03	0.04	0.04	0.05)
Percentile Difference										
M40M_Prop_Part	20	-0.02	0.03	0.03	0.03	0.02	0.01	0.00	-0.01	-0.01
		(0.11	0.08	0.07	0.07	0.07	0.06	0.06	0.06	0.10)
M60M_All	60	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
		(0.06	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.05)

Note: The three data sets M40M_Prop_Part, M_Intact, M60M_All were selected to represent the 20-item, 40-item and 60-item conditions for the multidimensional data sets derived from two-dimensional intact tests.

Table A-21. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing Tests of Different Length

	Customized Data Sets	Item No.	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	M40M_Prop_Part	20	0.95	(0.02)	-0.01	(0.02)	0.36	(0.04)
	M60M_All	60	0.94	(0.03)	0.01	(0.01)	0.28	(0.05)
Percentile	M40M_Prop_Part	20	0.96	(0.02)	0.14	(0.53)	9.16	(1.31)
	M60M_All	60	0.95	(0.03)	0.13	(0.19)	7.22	(1.48)

Note: The values in this table represent comparisons with the data set, M_Intact.

Table A-22. Selected Percentiles of the Estimated Ability Distributions Derived from Multidimensional Data Sets Representing the Combined Effects of Customizing Strategies and Estimation Items

Percentile		10	20	30	40	50	60	70	80	90
U40M_All	Mean	-1.25	-0.78	-0.47	-0.21	0.03	0.27	0.52	0.83	1.25
	Std. Err.	(0.08	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.07)
U40M_Part	Mean	-1.27	-0.78	-0.45	-0.19	0.05	0.28	0.53	0.83	1.27
	Std. Err.	(0.12	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.09)
U60M_All	Mean	-1.25	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.83	1.27
	Std. Err.	(0.04	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.04)
U_Intact	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
/U60M_Part	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
M40M_Prop_All	Mean	-1.28	-0.77	-0.45	-0.19	0.05	0.28	0.52	0.81	1.25
	Std. Err.	(0.07	0.05	0.05	0.04	0.05	0.04	0.05	0.04	0.06)
M40M_Prop_Part	Mean	-1.25	-0.79	-0.47	-0.20	0.03	0.26	0.51	0.81	1.25
	Std. Err.	(0.10	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.10)
M40M_Disp_All	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.51	0.82	1.26
	Std. Err.	(0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05)
M40M_Disp_Part	Mean	-1.25	-0.77	-0.46	-0.20	0.03	0.26	0.51	0.82	1.27
	Std. Err.	(0.09	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.09)
M60M_All	Mean	-1.26	-0.80	-0.48	-0.22	0.03	0.27	0.53	0.83	1.27
	Std. Err.	(0.05	0.04	0.04	0.04	0.03	0.03	0.04	0.04	0.05)
M_Intact	Mean	-1.26	-0.80	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
/ M60M_Part	Std. Err.	(0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06)
Percentile Differences										
U40M_All	Mean	0.02	0.01	0.01	0.02	0.01	0.01	0.01	0.01	-0.01
	Std. Err.	(0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.07)
U40M_Part	Mean	0.04	0.03	0.03	0.04	0.03	0.02	0.02	0.01	0.00
	Std. Err.	(0.07	0.07	0.07	0.07	0.07	0.07	0.05	0.06	0.09)
U60M_All	Mean	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00
	Std. Err.	(0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.05)
U60M_Part	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Std. Err.	(0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00)
M40M_Prop_All	Mean	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Std. Err.	(0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06)
M40M_Prop_Part	Mean	0.02	0.03	0.03	0.02	0.01	0.00	0.00	0.00	0.01
	Std. Err.	(0.04	0.05	0.05	0.04	0.04	0.04	0.05	0.06	0.09)
M40M_Disp_All	Mean	0.02	0.02	0.02	0.02	0.01	0.00	0.00	-0.01	-0.01
	Std. Err.	(0.05	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.07)
M40M_Disp_Part	Mean	0.03	0.03	0.03	0.03	0.02	0.01	0.00	-0.01	-0.01
	Std. Err.	(0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.10)
M60M_All	Mean	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
	Std. Err.	(0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.05)
M60M_Part	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Std. Err.	(0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00)

Table A-23. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing the Combined Effects of Customizing Strategies and Estimation Items

	Customized Data Sets	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U60M_All	0.98	(0.00)	0.01	(0.01)	0.16	(0.02)
	U60M_Part	1.00	(0.00)	0.00	(0.00)	0.00	(0.00)
	U40M_All	0.91	(0.02)	0.00	(0.01)	0.39	(0.04)
	U40M_Part	0.97	(0.01)	0.00	(0.01)	0.30	(0.03)
	M60M_All	0.94	(0.03)	0.01	(0.01)	0.28	(0.05)
	M60M_Part	1.00	(0.00)	0.00	(0.00)	0.00	(0.00)
	M40M_Prop_All	0.72	(0.07)	0.00	(0.01)	0.62	(0.07)
	M40M_Prop_Part	0.95	(0.02)	-0.01	(0.02)	0.36	(0.04)
	M40M_Disp_All	0.72	(0.11)	0.00	(0.01)	0.60	(0.12)
	M40M_Disp_Part	0.92	(0.03)	-0.01	(0.02)	0.38	(0.08)
Percentile	U60M_All	0.99	(0.00)	0.00	(0.21)	3.60	(0.48)
	U60M_Part	1.00	(0.00)	0.00	(0.00)	0.00	(0.00)
	U40M_All	0.93	(0.03)	0.18	(0.57)	9.84	(1.31)
	U40M_Part	0.98	(0.01)	0.44	(0.63)	7.38	(0.82)
	M60M_All	0.95	(0.03)	0.13	(0.19)	7.22	(1.48)
	M60M_Part	1.00	(0.00)	0.00	(0.00)	0.00	(0.00)
	M40M_Prop_All	0.72	(0.08)	0.06	(0.45)	17.19	(2.19)
	M40M_Prop_Part	0.96	(0.02)	0.14	(0.53)	9.16	(1.31)
	M40M_Disp_All	0.72	(0.13)	-0.03	(0.32)	16.49	(3.82)
	M40M_Disp_Part	0.92	(0.04)	-0.03	(0.52)	9.86	(2.25)

Note: Because new items were not used in estimation, U60M_Part yields the same results as U_Intact and M60M_Part the same as M_Intact Similarly, U40M_Part yields the same results as U40U_Part.

Table A-24. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Multidimensional Data Sets Representing Different Correlation Levels (Based on Unidimensional Intact Tests)

Percentile		10	20	30	40	50	60	70	80	90
U_.0	Mean	-1.26	-0.79	-0.47	-0.21	0.03	0.27	0.53	0.83	1.27
	Std. Err.	(0.05	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.06)
U_.3	Mean	-1.26	-0.79	-0.48	-0.21	0.03	0.27	0.53	0.83	1.25
	Std. Err.	(0.05	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.04)
U_.6	Mean	-1.25	-0.79	-0.47	-0.21	0.03	0.27	0.52	0.82	1.26
	Std. Err.	(0.05	0.03	0.04	0.04	0.03	0.03	0.03	0.03	0.05)
U_.9	Mean	-1.25	-0.79	-0.47	-0.21	0.03	0.26	0.52	0.82	1.26
	Std. Err.	(0.05	0.03	0.03	0.04	0.03	0.03	0.03	0.03	0.05)
U_Intact	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
Percentile Difference										
U_.0	Mean	-0.02	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00
	Std. Err.	(0.06	0.03	0.03	0.03	0.04	0.04	0.03	0.03	0.04)
U_.3	Mean	-0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00
	Std. Err.	(0.06	0.04	0.03	0.04	0.03	0.03	0.02	0.03	0.04)
U_.6	Mean	-0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.01	0.00
	Std. Err.	(0.05	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.04)
U_.9	Mean	-0.02	0.01	0.01	0.01	0.01	0.01	0.00	0.01	-0.01
	Std. Err.	(0.05	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.04)

Note:

U_.0 is the average of U60U_Part, U60U_All, U40U_Part, U40U_All when the correlation =.00, U_.3 is the average when the correlation =.30, U_.6 is the average when the correlation =.60, U_.9 is the average when the correlation =.90

Similar notations (.0, .3, .6, .9) representing correlations of latent abilities are also used in later tables.

Table A-25. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing Different Correlation Levels (Based on Unidimensional Intact Tests)

	Customized Data Sets	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U_.0	0.97	(0.01)	0.00	(0.01)	0.17	(0.02)
	U_.3	0.97	(0.01)	0.00	(0.01)	0.19	(0.02)
	U_.6	0.97	(0.00)	0.00	(0.01)	0.19	(0.01)
	U_.9	0.98	(0.00)	0.00	(0.01)	0.16	(0.01)
Percentile	U_.0	0.98	(0.00)	0.15	(0.37)	4.00	(0.51)
	U_.3	0.98	(0.01)	0.19	(0.32)	4.58	(0.56)
	U_.6	0.98	(0.00)	0.12	(0.33)	4.70	(0.44)
	U_.9	0.98	(0.00)	0.12	(0.38)	4.08	(0.36)

Table A-26. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Multidimensional Data Sets Representing Different Correlation Levels (Based on Two-dimensional Intact Tests)

Percentile		10	20	30	40	50	60	70	80	90
M_.0	Mean	-1.26	-0.78	-0.46	-0.20	0.03	0.27	0.52	0.83	1.27
	Std. Err.	(0.06	0.03	0.04	0.04	0.04	0.05	0.05	0.04	0.06)
M_Intact _.0	Mean	-1.26	-0.80	-0.48	-0.21	0.03	0.28	0.53	0.83	1.27
	Std. Err.	(0.07	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.07)
M_.3	Mean	-1.26	-0.79	-0.47	-0.20	0.04	0.27	0.52	0.82	1.25
	Std. Err.	(0.05	0.04	0.03	0.04	0.04	0.04	0.05	0.03	0.04)
M_Intact _.3	Mean	-1.27	-0.81	-0.48	-0.22	0.03	0.27	0.52	0.82	1.25
	Std. Err.	(0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.05)
M_.6	Mean	-1.25	-0.79	-0.48	-0.21	0.03	0.26	0.51	0.81	1.26
	Std. Err.	(0.03	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.04)
M_Intact _.6	Mean	-1.25	-0.80	-0.49	-0.23	0.01	0.25	0.50	0.81	1.27
	Std. Err.	(0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.05)
M_.9	Mean	-1.24	-0.78	-0.47	-0.21	0.03	0.26	0.52	0.82	1.26
	Std. Err.	(0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.03)
M_Intact _.9	Mean	-1.24	-0.79	-0.49	-0.22	0.02	0.26	0.51	0.81	1.26
	Std. Err.	(0.04	0.04	0.03	0.03	0.02	0.02	0.03	0.03	0.04)
Percentile Difference										
M_.0	Mean	0.00	0.02	0.01	0.01	0.00	-0.01	-0.01	-0.01	0.00
	Std. Err.	(0.05	0.04	0.04	0.03	0.03	0.02	0.02	0.03	0.04)
M_.3	Mean	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	-0.01
	Std. Err.	(0.05	0.03	0.03	0.03	0.03	0.02	0.02	0.03	0.05)
M_.6	Mean	-0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.00	0.00
	Std. Err.	(0.04	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.04)
M_.9	Mean	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Std. Err.	(0.04	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.04)

Note: M_.0 is the average of M60M_Part, M60M_All, M40M_Prop_Part, M40M_Prop _All, M40M_Disp_Part, M40M_ Disp _All when the correlation =.00, M_.3 is the average when the correlation =.30, M_.6 is the average when the correlation =.60, M_.9 is the average when the correlation =.90

Table A-27. Comparing Ability Estimates and Percentile Estimates Derived from Multidimensional Data Sets Representing Different Correlation Levels (Based on Two-dimensional Intact Tests)

	Customized Data Sets	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	M_.0	0.90	(0.04)	0.00	(0.01)	0.34	(0.08)
	M_.3	0.92	(0.02)	0.00	(0.01)	0.31	(0.03)
	M_.6	0.96	(0.01)	0.00	(0.01)	0.23	(0.02)
	M_.9	0.97	(0.00)	0.00	(0.01)	0.17	(0.01)
Percentile	M_.0	0.91	(0.05)	-0.05	(0.38)	9.39	(2.42)
	M_.3	0.92	(0.02)	0.05	(0.33)	8.39	(1.05)
	M_.6	0.96	(0.01)	0.15	(0.20)	6.22	(0.52)
	M_.9	0.98	(0.00)	0.13	(0.21)	4.42	(0.31)

Table A-28. Selected Percentiles and Differences in Estimated Ability Distributions Derived from Data Set Representing the Combined Effects of Correlations of Latent Abilities and Customizing Strategies

Percentile		10	20	30	40	50	60	70	80	90
U60M_All_.0	Mean	-1.25	-0.8	-0.49	-0.22	0.02	0.26	0.53	0.83	1.27
	Std. Err.	0.04	0.03	0.03	0.02	0.02	0.03	0.02	0.03	0.05
U60M_All_.3	Mean	-1.25	-0.8	-0.49	-0.22	0.03	0.27	0.52	0.83	1.27
	Std. Err.	0.04	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.03
U60M_All_.6	Mean	-1.25	-0.8	-0.48	-0.23	0.02	0.26	0.52	0.83	1.26
	Std. Err.	0.05	0.03	0.03	0.02	0.03	0.03	0.03	0.03	0.04
U60M_All_.9	Mean	-1.25	-0.8	-0.49	-0.23	0.02	0.26	0.52	0.83	1.27
	Std. Err.	0.04	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.05
U40M_All_.0	Mean	-1.27	-0.79	-0.47	-0.2	0.04	0.27	0.53	0.83	1.26
	Std. Err.	(0.09	0.06	0.07	0.07	0.08	0.07	0.05	0.05	0.08)
U40M_All_.3	Mean	-1.25	-0.78	-0.47	-0.21	0.03	0.28	0.53	0.83	1.24
	Std. Err.	(0.08	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.07)
U40M_All_.6	Mean	-1.24	-0.78	-0.47	-0.21	0.03	0.26	0.52	0.82	1.25
	Std. Err.	(0.07	0.05	0.05	0.05	0.05	0.04	0.05	0.04	0.07)
U40M_All_.9	Mean	-1.25	-0.78	-0.47	-0.21	0.03	0.27	0.51	0.82	1.25
	Std. Err.	(0.07	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.05)
U_Intact	Mean	-1.24	-0.79	-0.48	-0.22	0.02	0.26	0.52	0.82	1.26
	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.05)
Percentile Difference										
U60M_All_.0	Mean	-0.01	-0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00
	Std. Err.	(0.04	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.04)
U60M_All_.3	Mean	0.00	-0.01	-0.01	0.00	0.01	0.01	0.01	0.02	0.02
	Std. Err.	(0.05	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.05)
U60M_All_.6	Mean	-0.01	-0.01	0.00	-0.01	0.00	0.00	0.00	0.01	0.00
	Std. Err.	(0.06	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.05)
U60M_All_.9	Mean	-0.02	-0.01	-0.01	0.00	-0.01	0.00	0.01	0.01	0.00
	Std. Err.	(0.05	0.04	0.03	0.03	0.02	0.02	0.02	0.03	0.06)
U40M_All_.0	Mean	-0.03	0	0.01	0.02	0.02	0.02	0.01	0.01	-0.01
	Std. Err.	(0.10	0.06	0.07	0.06	0.07	0.07	0.05	0.05	0.07)
U40M_All_.3	Mean	0	0.01	0.01	0.01	0.01	0.02	0.02	0.01	-0.01
	Std. Err.	(0.09	0.07	0.06	0.07	0.06	0.06	0.04	0.05	0.08)
U40M_All_.6	Mean	0	0.01	0.01	0.01	0.01	0.01	0	0	-0.01
	Std. Err.	(0.08	0.06	0.06	0.05	0.05	0.05	0.04	0.05	0.07)
U40M_All_.9	Mean	-0.01	0.01	0.01	0.02	0.01	0.01	0	0.01	-0.02
	Std. Err.	(0.08	0.06	0.05	0.05	0.05	0.04	0.04	0.05	0.05)

Note:

The data set U_Intact is not described based on correlations of latent abilities as its results are not subject to changes in the correlations.

Table A-29. Comparing Ability Estimates and Percentile Estimates for the Combined Effects of Correlations of Latent Abilities and Customizing Strategies

	Customized Data Sets	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U60M_All_.0	0.99	(0.00)	0.01	(0.01)	0.10	(0.01)
	U60M_All_.3	0.97	(0.01)	0.01	(0.01)	0.16	(0.02)
	U60M_All_.6	0.96	(0.01)	0.01	(0.01)	0.21	(0.03)
	U60M_All_.9	0.97	(0.00)	0.00	(0.01)	0.18	(0.01)
	U40M_All_.0	0.86	(0.03)	0.00	(0.02)	0.35	(0.05)
	U40M_All_.3	0.83	(0.03)	0.00	(0.01)	0.42	(0.05)
	U40M_All_.6	0.86	(0.02)	0.01	(0.01)	0.40	(0.03)
	U40M_All_.9	0.89	(0.01)	0.00	(0.01)	0.32	(0.02)
Percentile	U60M_All_.0	0.99	(0.00)	-0.01	(0.27)	2.00	(0.26)
	U60M_All_.3	0.99	(0.00)	0.09	(0.25)	3.73	(0.49)
	U60M_All_.6	0.97	(0.01)	-0.07	(0.34)	5.08	(0.97)
	U60M_All_.9	0.98	(0.00)	-0.09	(0.35)	4.52	(0.41)
	U40M_All_.0	0.90	(0.02)	0.18	(0.74)	8.52	(1.29)
	U40M_All_.3	0.86	(0.04)	0.26	(0.63)	10.62	(1.74)
	U40M_All_.6	0.88	(0.02)	0.09	(0.55)	10.39	(1.11)
	U40M_All_.9	0.91	(0.01)	0.09	(0.60)	8.06	(0.61)

Table A-30. Selected Percentiles of Estimated Ability Distributions Derived from Data Sets Representing the Combined Effects of Correlations of Latent Abilities, Customizing Strategies and Estimation Items

Percentiles		10	20	30	40	50	60	70	80	90
M60M_All_.0	Mean	-1.26	-0.81	-0.48	-0.21	0.03	0.28	0.54	0.84	1.27
	Std. Err.	(0.05	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.06)
M60M_All_.3	Mean	-1.26	-0.8	-0.48	-0.21	0.03	0.28	0.53	0.83	1.26
	Std. Err.	(0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04)
M60M_All_.6	Mean	-1.26	-0.8	-0.49	-0.22	0.02	0.26	0.52	0.83	1.28
	Std. Err.	(0.04	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.04)
M60M_All_.9	Mean	-1.26	-0.8	-0.49	-0.22	0.02	0.26	0.52	0.82	1.27
	Std. Err.	(0.04	0.03	0.03	0.03	0.02	0.02	0.03	0.03	0.05)
M40M_Prop_All_.0	Mean	-1.24	-0.78	-0.46	-0.2	0.03	0.25	0.5	0.8	1.25
	Std. Err.	(0.14	0.06	0.07	0.07	0.07	0.07	0.08	0.08	0.11)
M40M_Prop_All_.3	Mean	-1.25	-0.79	-0.47	-0.21	0.03	0.26	0.52	0.82	1.26
	Std. Err.	(0.09	0.05	0.05	0.06	0.05	0.06	0.06	0.05	0.08)
M40M_Prop_All_.6	Mean	-1.26	-0.79	-0.47	-0.21	0.03	0.27	0.51	0.82	1.25
	Std. Err.	(0.05	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.07)
M40M_Prop_All_.9	Mean	-1.24	-0.79	-0.47	-0.21	0.03	0.27	0.51	0.82	1.25
	Std. Err.	(0.04	0.04	0.03	0.03	0.03	0.03	0.04	0.06	0.08)
M40M_Prop_Part_.0	Mean	-1.29	-0.76	-0.44	-0.19	0.03	0.26	0.51	0.83	1.28
	Std. Err.	(0.09	0.06	0.06	0.05	0.06	0.06	0.06	0.05	0.08)
M40M_Prop_Part_.3	Mean	-1.29	-0.78	-0.45	-0.19	0.05	0.28	0.52	0.81	1.22
	Std. Err.	(0.06	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.05)
M40M_Prop_Part_.6	Mean	-1.27	-0.78	-0.45	-0.18	0.05	0.28	0.52	0.8	1.23
	Std. Err.	(0.07	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.05)
M40M_Prop_Part_.9	Mean	-1.26	-0.77	-0.45	-0.19	0.05	0.28	0.52	0.82	1.27
	Std. Err.	(0.05	0.05	0.05	0.04	0.04	0.03	0.03	0.03	0.04)
M40M_Disp_All_.0	Mean	-1.25	-0.8	-0.48	-0.21	0.03	0.27	0.52	0.82	1.26
	Std. Err.	(0.07	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.07)
M40M_Disp_All_.3	Mean	-1.25	-0.8	-0.48	-0.22	0.03	0.27	0.52	0.82	1.25
	Std. Err.	(0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.05)
M40M_Disp_All_.6	Mean	-1.25	-0.79	-0.48	-0.22	0.02	0.25	0.51	0.81	1.26
	Std. Err.	(0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.05)
M40M_Disp_All_.9	Mean	-1.23	-0.79	-0.47	-0.22	0.02	0.26	0.51	0.82	1.25
	Std. Err.	(0.04	0.04	0.03	0.03	0.02	0.02	0.03	0.03	0.04)
M40M_Disp_Part_.0	Mean	-1.28	-0.76	-0.45	-0.19	0.04	0.27	0.52	0.83	1.29
	Std. Err.	(0.08	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06)
M40M_Disp_Part_.3	Mean	-1.27	-0.78	-0.45	-0.19	0.04	0.27	0.52	0.8	1.24
	Std. Err.	(0.06	0.04	0.03	0.04	0.04	0.04	0.04	0.04	0.04)
M40M_Disp_Part_.6	Mean	-1.24	-0.78	-0.47	-0.22	0.02	0.25	0.5	0.81	1.28
	Std. Err.	(0.04	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.04)
M40M_Disp_Part_.9	Mean	-1.23	-0.77	-0.46	-0.21	0.03	0.26	0.52	0.83	1.28
	Std. Err.	(0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.03	0.05)

Table A-30 Cont'd

Percentiles		10	20	30	40	50	60	70	80	90
M_Intact _.0	Mean	-1.26	-0.8	-0.48	-0.21	0.03	0.28	0.53	0.83	1.27
	Std. Err.	(0.07	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.07)
M_Intact _.3	Mean	-1.27	-0.81	-0.48	-0.22	0.03	0.27	0.52	0.82	1.25
	Std. Err.	(0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.05)
M_Intact _.6	Mean	-1.25	-0.8	-0.49	-0.23	0.01	0.25	0.5	0.81	1.27
	Std. Err.	(0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.05)
M_Intact _.9	Mean	-1.24	-0.79	-0.49	-0.22	0.02	0.26	0.51	0.81	1.26
	Std. Err.	(0.04	0.04	0.03	0.03	0.02	0.02	0.03	0.03	0.04)

Note: The results of M60M_Part at different correlation levels are not listed in this table as they are the same as those of M_Intact by design.

Table A-31. Differences in Selected Percentiles Derived from Data Sets Representing the Combined Effects of Correlations of Latent Abilities, Customizing Strategies and Estimation Items

Percentile Differences	10	20	30	40	50	60	70	80	90
M60M_All_.0	0.00	-0.01	-0.01	0.00	0.00	0.01	0.01	0.01	0.00
	(0.06	0.04	0.04	0.03	0.03	0.02	0.03	0.03	0.04)
M60M_All_.3	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01
	(0.06	0.04	0.03	0.03	0.03	0.03	0.03	0.04	0.05)
M60M_All_.6	-0.01	-0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01
	(0.05	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.05)
M60M_All_.9	-0.01	-0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01
	(0.05	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.04)
M40M_Prop_All_.0	0.03	0.03	0.02	0.01	0.00	-0.02	-0.03	-0.03	-0.02
	(0.1	0.08	0.07	0.06	0.06	0.05	0.06	0.06	0.09)
M40M_Prop_All_.3	0.02	0.02	0.01	0.01	0.00	-0.01	0.00	0.00	0.00
	(0.08	0.06	0.06	0.05	0.06	0.05	0.04	0.05	0.07)
M40M_Prop_All_.6	-0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.00	-0.01
	(0.08	0.05	0.04	0.04	0.04	0.04	0.03	0.04	0.06)
M40M_Prop_All_.9	0.01	0.00	0.02	0.02	0.01	0.01	0.01	0.00	-0.01
	(0.07	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.05)
M40M_Prop_Part_.0	-0.03	0.05	0.04	0.03	0.00	-0.01	-0.02	0.00	0.02
	(0.13	0.09	0.08	0.09	0.08	0.07	0.06	0.08	0.11)
M40M_Prop_Part_.3	-0.01	0.03	0.03	0.03	0.02	0.01	0.00	-0.01	-0.03
	(0.13	0.08	0.07	0.07	0.07	0.06	0.07	0.05	0.08)
M40M_Prop_Part_.6	-0.02	0.02	0.03	0.04	0.03	0.03	0.02	-0.01	-0.03
	(0.09	0.08	0.06	0.05	0.05	0.05	0.05	0.04	0.1)
M40M_Prop_Part_.9	-0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.01	0.01
	(0.08	0.07	0.07	0.06	0.05	0.05	0.05	0.05	0.09)
M40M_Disp_All_.0	0.01	0.01	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01
	(0.08	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.08)
M40M_Disp_All_.3	0.03	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	0.00
	(0.07	0.05	0.04	0.04	0.04	0.04	0.03	0.04	0.06)
M40M_Disp_All_.6	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	(0.06	0.04	0.04	0.04	0.03	0.03	0.03	0.04	0.06)
M40M_Disp_All_.9	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00
	(0.05	0.04	0.03	0.03	0.03	0.02	0.03	0.04	0.04)
M40M_Disp_Part_.0	-0.02	0.05	0.03	0.02	0.01	-0.01	-0.01	0.00	0.02
	(0.14	0.08	0.06	0.06	0.05	0.05	0.06	0.07	0.1)
M40M_Disp_Part_.3	0.00	0.03	0.03	0.02	0.01	0.00	-0.01	-0.02	-0.01
	(0.09	0.06	0.06	0.05	0.05	0.05	0.04	0.05	0.09)
M40M_Disp_Part_.6	0.01	0.02	0.01	0.01	0.01	0.00	0.00	-0.01	0.01
	(0.06	0.05	0.03	0.03	0.03	0.04	0.03	0.04	0.08)
M40M_Disp_Part_.9	0.01	0.02	0.03	0.02	0.01	0.00	0.01	0.02	0.02
	(0.05	0.04	0.04	0.03	0.03	0.03	0.04	0.06	0.08)

Table A-32. Comparing Ability and Percentile Estimates for the Combined Effects of Latent Ability Correlations, Customizing Strategies and Estimation Items

Ability	Corr. Level	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
M60M_All	0.00	0.94	(0.07)	0.01	(0.02)	0.23	(0.10)
	0.30	0.91	(0.03)	0.01	(0.01)	0.34	(0.05)
	0.60	0.94	(0.01))	0.01	(0.01)	0.27	(0.02)
	0.90	0.97	(0.00)	0.00	(0.01)	0.19	(0.01)
M40M_Prop_All	0.00	0.49	(0.14)	0.01	(0.02)	0.80	(0.13)
	0.30	0.70	(0.05)	0.00	(0.02)	0.62	(0.06)
	0.60	0.84	(0.02)	0.00	(0.02)	0.45	(0.03)
	0.90	0.92	(0.01)	0.00	(0.01)	0.31	(0.01)
M40M_Prop_Part	0.00	0.86	(0.04)	-0.01	(0.03)	0.40	(0.07)
	0.30	0.89	(0.02)	-0.02	(0.03)	0.36	(0.04)
	0.60	0.91	(0.01)	-0.01	(0.02)	0.32	(0.03)
	0.90	0.92	(0.01)	-0.01	(0.02)	0.30	(0.02)
M40M_Disp_All	0.00	0.52	(0.26)	0.00	(0.02)	0.74	(0.26)
	0.30	0.70	(0.07)	0.00	(0.02)	0.62	(0.09)
	0.60	0.85	(0.02)	0.00	(0.01)	0.44	(0.03)
	0.90	0.92	(0.01)	0.00	(0.01)	0.30	(0.02)
M40M_Disp_Part	0.00	0.84	(0.09)	-0.02	(0.05)	0.42	(0.15)
	0.30	0.87	(0.04)	0.00	(0.02)	0.39	(0.07)
	0.60	0.91	(0.02)	0.00	(0.01)	0.33	(0.03)
	0.90	0.94	(0.01)	0.00	(0.01)	0.27	(0.02)
Percentile							
M60M_All	0.00	0.96	(0.06)	0.12	(0.38)	5.48	(2.73)
	0.30	0.92	(0.03)	0.19	(0.35)	8.87	(1.54)
	0.60	0.94	(0.01)	0.09	(0.22)	7.31	(0.79)
	0.90	0.98	(0.00)	-0.04	(0.22)	4.78	(0.43)
M40M_Prop_All	0.00	0.49	(0.16)	-0.06	(0.76)	22.46	(4.29)
	0.30	0.70	(0.06)	0.04	(0.61)	17.15	(1.83)
	0.60	0.85	(0.02)	0.22	(0.39)	11.97	(0.90)
	0.90	0.93	(0.01)	0.16	(0.41)	7.96	(0.53)
M40M_Prop_Part	0.00	0.88	(0.05)	-0.03	(0.85)	10.29	(2.19)
	0.30	0.91	(0.03)	0.04	(0.77)	9.17	(1.37)
	0.60	0.93	(0.01)	0.42	(0.56)	8.03	(0.82)
	0.90	0.94	(0.01)	0.43	(0.61)	7.38	(0.74)
M40M_Disp_All	0.00	0.53	(0.29)	-0.07	(0.61)	20.47	(8.40)
	0.30	0.69	(0.09)	-0.05	(0.45)	17.09	(2.77)
	0.60	0.85	(0.03)	0.04	(0.28)	11.91	(1.11)
	0.90	0.94	(0.01)	-0.01	(0.28)	7.71	(0.59)
M40M_Disp_Part	0.00	0.86	(0.09)	-0.25	(1.02)	10.69	(4.11)
	0.30	0.88	(0.04)	0.06	(0.62)	10.34	(2.03)
	0.60	0.92	(0.02)	0.12	(0.46)	8.54	(1.11)
	0.90	0.95	(0.01)	0.24	(0.38)	6.82	(0.62)

Table A-33. Selected Percentiles of Estimated Ability Distributions Derived from Data Sets Representing the Combined Effects of Correlations of Latent Abilities and Changing Test Length

Percentiles	Corr.	Item No.	10	20	30	40	50	60	70	80	90
M40M_Prop_Part	0.00	20	-1.29	-0.76	-0.44	-0.19	0.03	0.26	0.51	0.83	1.28
			(0.13	0.08	0.08	0.09	0.08	0.08	0.07	0.09	0.13)
M60M_All		60	-1.26	-0.81	-0.48	-0.21	0.03	0.28	0.54	0.84	1.27
			(0.05	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.06)
M_Intact		40	-1.26	-0.80	-0.48	-0.21	0.03	0.28	0.53	0.83	1.27
			(0.07	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.07)
M40M_Prop_Part	0.30	20	-1.29	-0.78	-0.45	-0.19	0.05	0.28	0.52	0.81	1.22
			(0.12	0.08	0.07	0.07	0.08	0.08	0.09	0.06	0.08)
M60M_All		60	-1.26	-0.80	-0.48	-0.21	0.03	0.28	0.53	0.83	1.26
			(0.12	0.08	0.07	0.07	0.08	0.08	0.09	0.06	0.08)
M_Intact		40	-1.27	-0.81	-0.48	-0.22	0.03	0.27	0.52	0.82	1.25
			(0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04)
M40M_Prop_Part	0.60	20	-1.27	-0.78	-0.45	-0.18	0.05	0.28	0.52	0.80	1.23
			(0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.05)
M60M_All		60	-1.26	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.83	1.28
			(0.08	0.07	0.06	0.06	0.05	0.05	0.05	0.05	0.09)
M_Intact		40	-1.25	-0.80	-0.49	-0.23	0.01	0.25	0.50	0.81	1.27
			(0.04	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.04)
M40M_Prop_Part	0.90	20	-1.26	-0.77	-0.45	-0.19	0.05	0.28	0.52	0.82	1.27
			(0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.05)
M60M_All		60	-1.26	-0.80	-0.49	-0.22	0.02	0.26	0.52	0.82	1.27
			(0.08	0.08	0.07	0.06	0.05	0.05	0.05	0.05	0.08)
M_Intact		40	-1.24	-0.79	-0.49	-0.22	0.02	0.26	0.51	0.81	1.26
			(0.04	0.04	0.03	0.03	0.02	0.02	0.03	0.03	0.04)
Percentile Difference											
M40M_Prop_Part	0.00	20	-0.03	0.05	0.04	0.03	0.00	-0.01	-0.02	0.00	0.02
			(0.13	0.09	0.08	0.09	0.08	0.07	0.06	0.08	0.11)
M60M_All		60	0.00	-0.01	-0.01	0.00	0.00	0.01	0.01	0.01	0.00
			(0.06	0.04	0.04	0.03	0.03	0.02	0.03	0.03	0.04)
M40M_Prop_Part	0.30	20	-0.01	0.03	0.03	0.03	0.02	0.01	0.00	-0.01	-0.03
			(0.13	0.08	0.07	0.07	0.07	0.06	0.07	0.05	0.08)
M60M_All		60	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01
			(0.06	0.04	0.03	0.03	0.03	0.03	0.03	0.04	0.05)
M40M_Prop_Part	0.60	20	-0.02	0.02	0.03	0.04	0.03	0.03	0.02	-0.01	-0.03
			(0.09	0.08	0.06	0.05	0.05	0.05	0.05	0.04	0.10)
M60M_All		60	-0.01	-0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01
			(0.05	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.05)
M40M_Prop_Part	0.90	20	-0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.01	0.01
			(0.08	0.07	0.07	0.06	0.05	0.05	0.05	0.05	0.09)
M60M_All		60	-0.01	-0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01
			(0.05	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.04)

Table A-34. Comparing Ability Estimates and Percentile Estimates for the Combined Effects of Correlations of Latent Abilities and Changing Test Length

	Data Set	Item No.	Corr. Level	Corr.	Std. Err.	Difference (Cus.-Intact)	Std. Err.	Absolute Difference |Cus.–Intact|	Std. Err.
Ability	M40M_Prop_Part	20	0.00	0.86	(0.04)	-0.01	(0.03)	0.40	(0.07)
	M60M_All	60		0.94	(0.07)	0.01	(0.02)	0.23	(0.10)
	M40M_Prop_Part	20	0.30	0.89	(0.02)	-0.02	(0.03)	0.36	(0.04)
	M60M_All	60		0.91	(0.03)	0.01	(0.01)	0.34	(0.05)
	M40M_Prop_Part	20	0.60	0.91	(0.01)	-0.01	(0.02)	0.32	(0.03)
	M60M_All	60		0.94	(0.01)	0.01	(0.01)	0.27	(0.02)
	M40M_Prop_Part	20	0.90	0.92	(0.01)	-0.01	(0.02)	0.30	(0.02)
	M60M_All	60		0.97	(0.00)	0.00	(0.01)	0.19	(0.01)
Percentile	M40M_Prop_Part	20	0.00	0.88	(0.05)	-0.03	(0.85)	10.29	(2.19)
	M60M_All	60		0.96	(0.06)	0.12	(0.38)	5.48	(2.73)
	M40M_Prop_Part	20	0.30	0.91	(0.03)	0.04	(0.77)	9.17	(1.37)
	M60M_All	60		0.92	(0.03)	0.19	(0.35)	8.87	(1.54)
	M40M_Prop_Part	20	0.60	0.93	(0.01)	0.42	(0.56)	8.03	(0.82)
	M60M_All	60		0.94	(0.01)	0.09	(0.22)	7.31	(0.79)
	M40M_Prop_Part	20	0.90	0.94	(0.01)	0.43	(0.61)	7.38	(0.74)
	M60M_All	60		0.98	(0.00)	-0.04	(0.22)	4.78	(0.43)

Table A-35. Comparing the Effect of Using Different Items for Norms Estimation for Unidimensional and Multidimensional Data Sets

	Customized Data Sets	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	Unidimensional_All	0.99	(0.00)	0.00	(0.01)	0.20	(0.01)
	Unidimensional_Part	0.99	(0.00)	0.00	(0.01)	0.15	(0.01)
	U_All	0.95	(0.01)	0.00	(0.01)	0.26	(0.03)
	U_Part	0.99	(0.00)	0.00	(0.01)	0.15	(0.01)
	M_All	0.82	(0.05)	0.00	(0.01)	0.49	(0.07)
	M_Part	0.98	(0.01)	-0.01	(0.01)	0.22	(0.03)
Percentile	Unidimensional_All	0.99	(0.00)	0.04	(0.26)	5.00	(0.36)
	Unidimensional_Part	0.99	(0.00)	0.22	(0.31)	3.69	(0.41)
	U_All	0.97	(0.01)	0.09	(0.34)	6.43	(0.83)
	U_Part	0.99	(0.00)	0.22	(0.31)	3.69	(0.41)
	M_All	0.82	(0.06)	0.06	(0.26)	13.33	(2.04)
	M_Part	0.98	(0.01)	0.04	(0.27)	5.55	(1.01)

Note:

Unidimensional_All and Unidimensional_Part represent the conditions when unidimensionality is satisfied. Specifically, Unidimensional_All is the average of U60U_All and U40U_All, and Unidimensional_Part is the average of U60U_ Part and U40U_ Part.

U_All and U_Part represent the conditions in which multidimensional data sets are derived from unidimensional intact tests. Specifically, U_All is the average of U60M_All and U40M_All, and U_Part is the average of U60M_ Part and U40M_ Part.

M_All and M_Part represent the conditions in which multidimensional data sets are derived from two-dimensional intact test. Specifically, M_All is the average of M60U_All, M40M_Prop_All, and M40M_Disp_All, and M_Part is the average of M60U_Part, M40M_Prop_Part, and M40M_Disp_Part

Table A-36. Comparing Customizing Strategies for Unidimensional and Multidimensional Data Sets

	Customized Data Sets	Correlation	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U60U	1.00	(0.00)	0.00	(0.00)	0.08	(0.01)
	U40U	0.98	(0.00)	0.00	(0.01)	0.25	(0.02)
	U60M	1.00	(0.00)	0.00	(0.00)	0.08	(0.01)
	U40M	0.96	(0.01)	0.00	(0.01)	0.31	(0.03)
	M60M	0.98	(0.01)	0.00	(0.00)	0.14	(0.03)
	M40M_Prop	0.89	(0.03)	0.00	(0.01)	0.41	(0.04)
	M40M_Disp	0.88	(0.05)	0.00	(0.01)	0.42	(0.08)
Percentile	U60U	1.00	(0.00)	-0.03	(0.10)	1.92	(0.23)
	U40U	0.99	(0.00)	0.29	(0.49)	6.23	(0.57)
	U60M	1.00	(0.00)	0.00	(0.11)	1.80	(0.24)
	U40M	0.97	(0.01)	0.31	(0.57)	7.77	(0.86)
	M60M	0.99	(0.01)	0.07	(0.10)	3.61	(0.74)
	M40M_Prop	0.90	(0.03)	0.10	(0.43)	11.20	(1.28)
	M40M_Disp	0.87	(0.06)	-0.03	(0.34)	11.28	(2.55)

Note:

U60U and U40U represent conditions when unidimensionality was satisfied. Specifically, U60U represents customizing full-length intact tests and U40U represents customizing reduced-length intact tests under unidimensionality.

U60M and U40M represent conditions of multidimensional data sets that were derived from unidimensional intact test. Specifically, U60M represents customizing full-length intact tests and U40M represents customizing reduced-length unidimensional intact tests by adding items measuring a different latent ability.

M60M, M40M_Prop, and M40M_Disp represent conditions of multidimensional data sets that were derived from two-dimensional intact test. Specifically, M60M represents customizing full-length intact tests, M40M_Prop represents customizing proportionally reduced-length intact tests, and M40M_Disp represents customizing disproportionally reduced-length intact tests.

Table A-37. Comparing the Effect of Changing Test Length for Unidimensional and Multidimensional Data Sets

	Customized Data Sets	Item No.	Corr.	Std. Err.	Difference (Cus.- Intact)	Std. Err.	Absolute Difference |Cus. – Intact|	Std. Err.
Ability	U40U_Part	20	0.97	(0.01)	0.00	(0.01)	0.30	(0.03)
	U60U_All	60	0.99	(0.00)	0.00	(0.01)	0.15	(0.01)
	M40M_Prop_Part	20	0.95	(0.02)	-0.01	(0.02)	0.36	(0.04)
	M60M_All	60	0.94	(0.03)	0.01	(0.01)	0.28	(0.05)
Percentile	U40U_Part	20	0.98	(0.01)	0.44	(0.63)	7.38	(0.82)
	U60U_All	60	0.99	(0.00)	-0.07	(0.20)	3.84	(0.46)
	M40M_Prop_Part	20	0.96	(0.02)	0.14	(0.53)	9.16	(1.31)
	M60M_All	60	0.95	(0.03)	0.13	(0.19)	7.22	(1.48)

Note.

U40U_Part and U60U_All represent changing tests to 20 items and 60 items based on 40 item unidimensional intact tests when unidimensionality is satisfied.

M40M_Prop_Part and M60M_All represent changing tests to 20 items and 60 items based on 40 item two-dimensional intact tests.