DATE DUE	DATE DUE	DATE DUE
[illegible]		
JAN 15 2001		
[illegible]		
[illegible]		

EXAMINING LOCAL ITEM DEPENDENCE EFFECTS IN A LARGE SCALE SCIENCE ASSESSMENT BY A RASCH PARTIAL CREDIT MODEL

by

Jean Weiqin Yan

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology,
and Special Education

1996

ABSTRACT

EXAMINING LOCAL ITEM DEPENDENCE EFFECTS
IN A LARGE SCALE SCIENCE ASSESSMENT
BY A RASCH PARTIAL CREDIT MODEL

by

Jean Weiqin Yan

Frequently in a science assessment, several items are generated from the same scenario. These context-dependent items are traditionally analyzed as independent items. However, the potential local item dependence effects among these items may cause a biased estimation of the examinees' abilities in science literacy.

The purpose of this study was to investigate the local item dependence effects on testlets in the tryout version of the *Michigan High School Proficiency Test in Science* by the Rasch partial credit model.

Cluster sampling combined with stratified sampling was used in the tryout, in which school was the cluster unit and population density was the stratum unit. Data were analyzed in five different configurations to study the relationships between context-dependent items at the individual item level and at the testlet level.

The major findings of the study were:

1. Context-dependent items correlated more closely within-context than across-context for most original testlets.

2. Local dependence effects can be controlled and a better fit for item calibration can be obtained by employing the Rasch partial credit model for some, but not all original testlets.

3. There is no significant difference between the partial credit model and the dichotomous model in average person measures.

4. It seems that an implicit factor other than the local item dependence affects the misfit original testlets.

5. Truly statistically independent items should be analyzed independently, whether they belong to a context or not.

Additional costs will occur if one treats context-dependent items as testlets in a large-scale assesment because the partial credit model is more complex than the dichotomous model. More money, time, technology and human resources will be involved.

ACKNOWLEDGMENTS

I once joked that this list of acknowledgments would be longer than those in the Oscar Academy Awards, for so many people have contributed to the completion of this dissertation. To me, the experience of doctoral study and dissertation writing was invaluable and unforgettable for my professional development.

Today's accomplishment is primarily due to the unfailing support, guidance, and encouragement of my respectful advisor, Dr. William Mehrens. Throughout his busy schedule, Dr. Mehrens carefully scrutinized my manuscript many times and provided immediate and insightful suggestions, comments, and advice. His wisdom, open-mindedness, rich experiences of teaching, psychometrics, and education policy have been very precious to me throughout my doctoral study and will be, I believe, in the years to come.

I am deeply indebted to Dr. Benjamin Wright for his profound interest and substantial help in my study. Dr. Wright is not on my dissertation committee, but he has done as much as, if not more than, the committee members. Not only did he "rescue" me from the dead-end of the research, but also led me to the new direction and guided me step by step in the process of research through long distance

communication. Without his expertise in rating scale analysis and his encouragement, this study simply could not have been completed.

I wish to express my gratitude to the members of the dissertation committee: Dr. Steve Raudenbush, who recommended Dr. Wright to me, for his expertise in educational statistics and incisive criticism; Dr. Edward Smith for his thorough understanding of the Michigan science curriculum and the science assessment framework, and for his constructive and detailed comments and suggestions on the design of the study and the writing of this work; and Dr. Frederick Ignotavich for his expertise in education administration.

Special thanks go to my employer, Dr. Diane Smolen, at the Michigan Educational Assessment Program in Michigan Department of Education for her permission to use the *Michigan High School Proficiency Test* tryout data and for her consideration of adjusting my workload so that I had time to finish this project.

I sincerely appreciate Dr. Leonard Bianchi, Dr. Lindson Feun, Dr. Richard Houang, Dr. Mike Linacre, Dr. Robert Sykes, Dr. Richard Smith, and Ms. Wen-Ling Yang for their professional judgment in educational measurement and statistics, and valuable suggestions and comments to improve the quality of the study. All of them helped me without any reservation during the process of this study.

As for my colleagues, mentors, and dear friends Ms. Jan Hunt-Kost and Dr. Catherine Smith I can never say enough

"Thank you." Armed with their professional knowledge, both of them showed great interest in this study and contribute their precious time to edit my dissertation meticulously. Their continuous support, encouragement, and advice motivated me in my study and work.

Last but not least, I would like to thank my family, my relatives, and all my friends in China, the United States, and other parts of the world. Their unselfish love, deep faith, high expectation, and true understanding of my life pursuit inspired me to overcome countless obstacles in the past years to reach this milestone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The Problem

Traditional educational measurement theories assume that multiple-choice (MC) test items are not correlated to each other when examinees' abilities are controlled; each item is analyzed independently and dichotomously. Consequently, the unit of analysis is the item itself. However, in many testing situations, such as a short story in a reading comprehension test, a table in a mathematics test, or an investigation in a science test, a context is established and students are often asked a series of questions related to that context. Wainer and Kiely (1987) called a set of these context-dependent items a "testlet" and defined it as:

> "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow (p.190)."

For example, on the *Michigan High School Proficiency Test in Science* (tryout version, 1995), one testlet on life science had six context-dependent items, four of which were multiple-choice items and the remaining two were constructed-response questions. In this example, a genetic disease was described and students were asked to identify the information about the gene presented in the pedigree and draw conclusions about it.

Then the students identified the scientist who contributed to the explanation of the disease and the probability of an unborn baby getting the disease given the parents' health condition. Finally, a hypothetical situation was given and the students had to answer questions based on the pedigree the students had drawn and provide scientific reasons for the answers. These items were scored independently, even though they were related to the same context.

The immediate problem with conventional scoring methods under these circumstances is that the item response theory (IRT) assumption of local independence may be violated. In IRT, the assumption is that for a subpopulation of examinees at a given ability level, β_n, on a latent trait scale, the items are statistically independent of each other (i.e., $P(x_1=1, x_2=1 \mid \beta_n)=P(x_1=1 \mid \beta_n)P(x_2=1 \mid \beta_n)$, where x_1 is item 1 score and x_2 is item 2 score). Thus, the probability of answering one item correctly ($P(x_1=1 \mid \beta_n)$) does not affect the probability of the examinee's answering the other item correctly ($P(x_2=1 \mid \beta_n)$). When the items are statistically dependent, i.e., the probability of answering one item correctly depends on how one performs on the other, the equation does not hold (i.e., $P(x_1=1, x_2=1 \mid \beta_n) \neq P(x_1=1 \mid \beta_n) P(x_2=1 \mid \beta_n)$). The rationale for the assumption of local independence is that the trait value should provide all the related information about the examinee's knowledge and that the contribution of each item to the test can be evaluated independently of all other items.

One of the measurement implications of local item dependence is that there would be an effect on the test information obtained, because the test information function ($I(\beta_n)$) has an inverse relationship with the standard error of measurement (SEM) of the ability estimates at level β_n ($I(\beta_n)=1/\sqrt{SEM(Y)}$, Y is the examinee's total score). The estimate of information of a test is the sum of all the individual item information estimates, $I(\beta_n)=\sum_{i=1}^{L} I_i(\beta_n)$, i=1, ..., L, the number of items. The point is that this additive relationship is based on the assumption of local independence. When items are interdependent, the standard error of measurement of the test changes, depending on the direction of the correlation between items. Consequently, the test information calculated by $I_i(\beta_n)$, assuming local independence, will be an over- or underestimate of the true information (Thissen, Steinberg, & Mooney, 1989, Yen, 1993).

As to the direction of bias, Anastasi (1961) stated that:

> "Were the items in such a group to be placed in different halves of the test, the similarity of the half scores would be spuriously inflated, since any single error in understanding of the problem might affect items in both halves (p. 121)."

Guilford (1936) made a similar point:

> "Interdependent items tend to reduce the reliability. Such items are passed or failed together and this has the equivalent result of reducing the length of the test (p. 147)."

Theoretically large correlations between residuals may imply a second trait in the ability estimation. Rosenbaum

(1988) compared item response distributions when local independence was conditional between, but not within, item "bundles" (testlets) with two sets of IRT assumptions. One set was traditional IRT and the other was less restrictive on local independence, allowing dependence among pairs of items that shared the same context. He proved a theorem that at every level of ability, the standard error of measurement under a positively correlated bundle was at least as large as that from a conventional IRT model having the same item characteristic curves (ICCs). He also found that positive dependence within bundles increased the SEM along the ability continuum. He suggested that, other things being equal, it is preferable not to use bundles of positively dependent items since it may cause a larger SEM.

Thissen, Steinberg, and Mooney (1989) used a multivariate logistic latent trait model (Bock, 1972) to examine the violation of the local independence assumption with computerized adaptive test (CAT) data. They compared the results of a 4-testlet, 22-item test when the items were analyzed first as independent items and then as testlets. The results showed that, when testlet items were analyzed independently, the test information obtained was deceptively high. When those items were analyzed as testlets, the concurrent validity was slightly but significantly higher than that of the independently analyzed items. They concluded that the outcome of more information was "fooled" by the excess correlation within the testlet among items and that

the testlet scores appeared to be at least as valid as the individual item scores.

Yen (1993) used 3PL and 2PPCL models to study multiple-choice tests of the *Comprehensive Test of Basic Skills, Fourth Edition* (CTBS/4; CTB Macmillan/McGraw-Hill, 1989) and the performance assessment data of a state education assessment program. Item information and discrimination estimates obtained by testlet scale and by item scale on reading and math tests were compared. It was found that testlet analysis did result in a larger SEM, but it could be seen as a reflection of reality. However, in many cases, there was not much difference in parameter estimates when items were scaled as testlets or as independent items.

It seems that, for context-dependent items, using item as the unit of analysis may cause different erroneous results because some items may be more strongly correlated within a context than between contexts. These high correlations, which are context-specific rather than test-specific, result in biased measurement of the common factor between contexts (Thissen et al., 1989). The information curve in IRT and the high reliability index in classical test theory were misled by the excess item correlations within a testlet because context-dependent items may be themselves statistically dependent. An alternative is to analyze these items together as a unit.

Purpose of the Study

The purpose of the study was to explore the local item dependence effect when context-dependent items in the *Michigan High School Proficiency Test in Science* were analyzed as independent items and as testlets. In addition, originally independent items in the same test were randomly formed into testlets to conduct a concurrent validity analysis for the testlet effect. Both the traditional dichotomous rating scale and the partial credit scale in IRT Rasch models (Wright and Masters, 1982) were used. The computer software BIGSTEPS (Linacre and Wright, 1995; version 2.6) used here was designed to conduct Rasch measurement from the responses of a set of persons to a set of items.

If the results of the testlet-based analysis are not significantly different from the item-based analysis, it means that there is not enough evidence to reject the statement that context-dependent items within the testlets can be analyzed as individual items. The assumption of local independence will still hold. Consequently, it will not make a difference whether these context-dependent items are analyzed independently or as testlets. In general, the item-based analysis is easier to conduct and less expensive because the dichotomous scoring is a conventional approach and the scoring process has been established in the industry. Higher costs would occur for the testlet-based analysis because the scoring process and the scoring model is more complex and, therefore, more time, coding, computer

programming, technical support, and human resources would be involved. In addition, educating the education community and the public about the concepts of the testlet scoring would take a considerable amount of time and effort if one wants to use the testlet scale under this circumstance. In terms of the consequences of person or item estimation, there is little discussion in the literature on the impact of using the testlet-based analysis when context-dependent items are statistically independent. Practically speaking, one should choose the scale that is simpler and easier to analyze and interpret when there is no significant difference in item/person estimation between the two models.

However, if the results are significantly different, it indicates that local item dependence may exist, that context-dependent items are correlated either positively or negatively to each other within a testlet, and that actual measurement error is either overestimated or underestimated. As a result, these items should be analyzed as testlets with partial credit models. It is expected that the testlet analysis approach would provide an alternative in data analysis to control or alleviate the effect of the violation of the local independence assumption when local item dependence is indeed present.

Significance of the Study

Few studies have paid attention to the measurement characteristics of testlets, even though they have existed as

an item format almost as long as tests themselves. In the last decade, there has been growing interest in treating a set of context-dependent items as the unit of analysis in educational measurement research. One main reason that test developers are using larger tasks as the fundamental units of tests and further shifting their focus to this field is that, besides the testlet characteristics to be described later, modern tests serve more purposes than before. A test result may now be used not only for achievement assessment, diagnosis, placement, or admission purposes, but also as an important reference to policy making and education budgeting practices. The same amount of testing time and information are used to achieve more goals than before. Furthermore, researchers have experimentally projected that testlets as units of analysis can solve some of the measurement problems that could not be overcome by item-based analysis (Ebel, 1951; Wainer & Kiely, 1987; Rosenbaum, 1988; Thissen et al, 1988, 1989; Haladyna, 1992; Yen, 1984a, 1993).

Studies and discussions about testlets so far have been limited to applications of testlet concepts (Szeberényi & Tigyi, 1987; Wainer et al, 1990, 1991, 1992), construction and development of testlets (Engelhart, 1942; Gerberich, 1956; Gronlund, 1965; Biggs & Collis, 1982; Mehrens & Lehmann, 1984; Collis et al, 1986; Haladyna, 1991) and measurement precision (Cureton, 1965; Cattell & Burdsal, 1975; Wainer et al, 1990; Sireci et al, 1991; Ercikan, 1993). Studies on the effect of loss of local independence mostly

used IRT two-parameter (2PL) or three-parameter (3PL) polytomous models (Rosenbaum, 1988; Thissen et al, 1989, Donoghue, 1993, Yen, 1993).

A hidden problem in using a 2PL or 3PL model is that these models are sample dependent and results can vary from sample to sample because they do not have sufficient statistics and thus their mathematical formulas cannot converge. Consequently, the models cannot separate person parameter from item parameters. (Wright, 1992). An outstanding property of the Rasch model is that it has sufficient and necessary statistics that can separate person parameter from item parameter, and make it possible to construct the linear and objective measurement. More discussion about sufficient statistics for the IRT models will be presented later in Chapter 3.

Wilson (1988) used the family of Rasch models (dichotomous, partial credit, and rating scale) to study the local item dependence effect with an example of "superitems" (testlets) in the Structure of the Learning Outcome program. The results showed that the rating scale model calibration provided no evidence of the violation of the local item dependence assumption. Dependencies between items were adequately summarized by the dichotomous model item difficulties. On the other hand, the partial credit model calibration showed that one of the five testlets studied demonstrated a local item dependence effect. However, the sample size was very small in Wilson's study (1988). The data

were collected from only 30 students in the 9th and 10th grades, which is not comparable with a large scale assessment program.

Masters's (1982) Rasch partial credit model was originally developed to analyze multiple-category items and it has remained this way for most studies of this model. For multiple-choice item analysis, it was used for foil analysis to gain more information. Other uses have included theoretical exploration such as the multi-dimensionality issue (De Ayala, 1991) and necessary and sufficient conditions to equate the estimates from dichotomous and partial credit models (Huynh, 1994). However, most comparisons were on the item level, not on the testlet level. Wilson and Iventosch (1988) conducted a study at the testlet level, but the items were performance-based and the research was experimental with small samples. So far, studies have found that the partial credit model added more detailed information to the dichotomous model and provided the opportunity to observe the local dependence between items within a testlet when the situation occurred.

A review of the literature on this topic indicates that there have been no studies examining the local dependence due to the testlet effect in any large-scale, high-stake state assessment programs using Masters' partial credit model for MC items. This study attempts to do so. (Studies done with 2PL or 3PL partial credit models are not the focus of the discussion here, which does not mean that they are not

important. Rather, the intent is to concentrate on the main models of interest under study and to avoid complexity and issues inherent in 2PL and 3PL partial credit models.) In addition, the study will explore the curriculum impact on item analysis study. Sometimes it is possible that the context of constructing a testlet makes perfect sense in curriculum, but it does not affect the analysis of scoring scales psychometrically. The study results of the testlets in the newly developed *Michigan High School Proficiency Test in Science* will provide evidence of a real life example in applying an alternative item analysis method to a large scale, high-stake assessment program. It will also explore other techniques that people can use in item analysis so that the methods and results of this study can contribute to the item analysis field.

Research Hypotheses

Based on the purpose and rationale of the study, the following research hypotheses are proposed to study the local item dependence effect.

1. For context-dependent items,

   (a) the average item correlations within an original testlet are larger than the average correlations with items from other testlet configurations;

   (b) when they are analyzed as a testlet by the Rasch partial credit model, they produce a better testlet fit statistic than when they are analyzed as individual items by the Rasch dichotomous model;

   (c) when they are analyzed as a testlet by the Rasch partial credit model, they produce better person fit

statistics than when they are analyzed as individual items by the Rasch dichotomous model;

(d) when they are analyzed as a testlet, the measurement errors are smaller than when they are analyzed as individual items. In other words, the person separation reliability is higher for testlet-based analysis than for item-based analysis.

2. For independent items,

   (a) when they are analyzed as a testlet by the Rasch partial credit model, the testlet fit statistics are the same as the item fit statistics when they are analyzed as individual items by the Rasch dichotomous model;

   (b) person fit statistics stay the same regardless of whether the items are analyzed as random testlets or as individual items;

   (c) the reliability of the person separation ratio is the same for both testlet-based analysis and item-based analysis.

3. When context-dependent items in the original testlets of the same tryout form are decomposed and reformed into the same number of new testlets, each with an item from each original testlet, as if they were in different contexts,

   (a) the average correlations between items within a reformed testlet are smaller than the average correlations between items within an original testlet;

   (b) person fit estimated by the reformed testlets are not as good as those estimated by the original testlets.

Two Scoring Scales of IRT Rasch Models

The purpose of any test theory is to describe how inferences from examinees' test scores or item responses can be made about unobservable characteristics that are measured by tests. These characteristics are referred to as *traits* or *abilities*. Since they are not directly measurable, they are called latent traits or abilities. With item response theory,

test developers usually assume that a single latent trait is considered to be responsible for item responses on a test if the test is designed to measure that trait. An item response model specifies a relationship between the observable examinee test performance and the unobservable trait or ability assumed to underlie performance on the test. The relationship is described by a mathematical formula which explains how examinees at different ability levels on the trait scale should respond to an item. Graphically, this relationship is reflected by the item characteristic curve (ICC), the key concept of IRT. Basically, an ICC plots the probability of responding correctly to an item as a function of the latent trait underlying performance on the test items. This knowledge allows one to compare the performance of examinees who have taken different tests. It also permits one to apply the results of an item analysis to groups with different ability levels.

Different item response models are constructed through specified assumptions that one is willing to make about the test data set under study. For this study, two models in the family of Rasch models (i.e., one parameter models) were used: the dichotomous model (DM) and the partial credit model (PCM). The family of models was named after Georg Rasch, a Danish mathematician, who formulated this approach in the 1950s and 1960s. It is a method for obtaining objective, fundamental measures from stochastic observations of ordered category responses (Linacre and Wright, 1995). The family of

Rasch models is suitable for testlet analysis as it has well-developed and interpretable polytomous extensions that embody the assumed item/category dependence and that make inter-model comparisons relatively easy by having identical sufficient statistics for the person ability parameters.

The dichotomous model assumes that there are only two levels or categories of performance such as right/wrong, yes/no, or pass/fail for an item. It provides a way to place persons and items on a scale with a clear probabilistic interpretation of distance on the scale. Items scored in this way can be considered as "one-step" items. If an examinee completes the step, 1 point is awarded, otherwise, 0. That is, responding to an item correctly means completing a step. This scoring method is widely used in the multiple-choice item tests. The model was used here whenever items in the data were analyzed independently.

The partial credit model (PCM) is an extension of the DM and handles data that scale more than one step in an item. For example, writing assessment frequently scores examinees with different writing levels. The PCM's basic observation is the number of steps that an examinee accomplishes in an item. If, for example, an item has 3 steps, an examinee can get a score of x = 0, 1, 2, or 3 points. More examples of partial credit scoring are provided in Appendix A. It can be seen that the basic measure in the PCM is the step difficulties within an item. The assumption for the PCM is that the step difficulties are not equally distanced among the performance

levels. For example, in Example 1 of Appendix A ($\sqrt{(9.0/0.3)-5}=?$), step 2, (30-5=25), is much easier than step 1, (9.0/0.3=30). In addition, the number of steps across items for a test does not have to be the same. Theoretically, steps in an item of the PCM should be ordered and are answered accordingly. One needs to complete step 1 before moving on to step 2. In this study, the "steps" were the number of items in a testlet. The mechanism of the partial credit to an item was borrowed here to award partial credit to a testlet in that the items in a testlet were analogous to the steps in an item and the testlet was analogous to a conventional MC item. The total number of the raw score for a testlet would be treated as the testlet score and was used for testlet analysis. Details are presented in Chapter 3.

Structure of The Study

In the first chapter, the problem of local item dependence, the measurement issues in testlet analysis, the purpose of the study, the significance of the study, the research hypotheses, and two scoring models in the family of Rasch models have been introduced. In Chapter 2 the author reviews the literature on the concepts of testlets, characteristics of testlets, construction and development of testlets, application of testlet concepts, and research on the local independence assumption in IRT. Chapter 3 is the methodology chapter in which the testing materials, the data, the sampling procedures, the research hypotheses, item

scoring, testlet categories, calibration models, estimation measures, the data analyses, and the computer program of this study are the foci. In Chapter 4 the results of different measures described in Chapter 3 are reported and discussed. In the final chapter a summary of the study and the results by hypothesis are furnished. Also presented are the conclusions, limitations, generalizability of the study, and recommendations for future research.

CHAPTER 2

LITERATURE REVIEW

There are six sections in this chapter. The first two sections cover concepts and characteristics of testlets. In the third section, testlet construction and development are discussed. The fourth and fifth sections are devoted to the application of testlet concepts and measurement precision, especially when the assumption of local independence is violated. The focus is on theoretical development, assumptions, and characteristics. Finally, the literature reviewed to the present study is summarized.

Concepts of Testlets

The problem of violating local independence with context-dependent items and consequential estimation bias invited a review of the structure of context-dependent items, which was discussed extensively a few decades ago (Ebel, 1951; Anastasi, 1961; Gronlund, 1965; Mehrens & Lehmann, 1984). Ebel named the context-dependent items as the "interpretive test exercises" and predicted that this format would be highly promising. In his *Writing the Test Items*, Ebel (1951) defined the interpretive test exercise as follows:

> "The interpretive test exercise consists of an introductory selection of material followed by a series of questions calling for various interpretations. The material to be interpreted may be a selection of almost any type of writing (news, fiction, science, poetry, etc.), a table, map, chart, diagram, or illustration; the description of an experiment or a legal problem; even a baseball box score or a portion of a music configuration. The questions on this material may be based on explicit statements in the material, on inferences, explanations, generalizations, conclusions, criticisms, and on many other interpretations (p. 241)."

Gronlund (1965), following Ebel, used the same name but a less specific definition:

> "An interpretive exercise consists of a series of objective items based on a common set of data. The data may be in the form of written materials, tables, charts, graphs, maps, or pictures. The series of related test items may also take various forms but are most commonly of the multiple-choice or alternative-response variety (p. 161)."

Nevertheless, Gronlund demonstrated extensively the forms and uses of the interpretive exercise to measure complex achievement of an examinee, such as the ability to recognize assumptions, inferences, and relevance of information, to apply principles, and to interpret experimental findings.

Mehrens and Lehmann's (1984) definition of the interpretive exercise was similar to Gronlund's but emphasized that the introductory material should be identical for all students:

> "The interpretive exercise consists of either an introductory statement, pictorial material, or a combination of the two, followed by a series of questions that measure in part the student's ability to interpret the material. All test items are based on a set of materials that is identical for all students (p. 295)"

What was different was that Mehrens and Lehmann presented *interlinear exercise* as a format in the context-dependent

literature. In their definition, an interlinear exercise was "somewhat of a cross between the essay question (the student is given some latitude of free expression in that he decides what is to be corrected and how it is to be corrected) and the objective item (the answer can be objectively scored) (p. 295)." For example,

Example 1.

"Harry was ~~alright~~ *all right* at ~~grammer~~ *grammar*, but he didn't excel at ~~speling~~ *spelling*."

"The researchers ~~are of the opinion~~ *believe* ~~that this~~ *the* test *often* produces biased results ~~a great number of times owing to the fact that~~ *because* subjects ~~exhibit a~~ tend~~ency~~ to misinterpret the questions."

It should be pointed out that all definitions above include the pictorial form as a medium to be used to present the material to examinees. It is considered that the pictorial form fit very well for younger children and for children with some reading deficiencies. It is a unique tool for directly measuring an examinee's ability to interpret graphs, maps, tables, and even cartoons. In some cases, pictorial material presents and explains far more precisely, simply, and effectively than does text material.

Other terms that have been used for the content-dependent items included "superitems" (Cureton, 1965), "application test" (Szeberënyi and Tigyi, 1987), "item bundle" (Rosenbaum, 1988), and "item set" (Haladyna, 1992). Szeberënyi and Tigyi defined an "application test" as follows:

> "The test consists of a description of an experiment, including data presented in tables or figures interspersed with built-in multiple-choice questions (p.73)."

Rosenbaum's definition of "item bundle" was that:

> "An item bundle is a small group of multiple-choice items that share a common reading passage or graph, or a small group of matching items that shares distractors (p.349)."

Haladyna's definition for a testlet was the simplest one:

> "A context-dependent item set consists of an introductory stimulus and a set of related test items (p.21)."

The term "testlet" was first introduced by Wainer and Kiely (1987) as:

> "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow (p.190)."

This definition was different from the previous ones in that it clearly spelled out the nature of the information selection as "a single content area" and emphasized its development "as a unit." This implied that the items generated from that content area should be analyzed together as a unit. Secondly, it identified the logical relationship between items. It may also be inferred that the testlet concept has covered several different forms of context-dependent items. This more inclusive definition has been widely accepted and therefore will be used hereafter in this study. Wainer and Kiely (1987) expected that using the testlet as the unit of analysis could ease some of the observed and prospective difficulties associated with most of the current algorithmic methods of test construction, specifically, for computerized adaptive tests.

There are two ways of classifying testlets: by content form and by logical relationship (see Figure 1). The content form consists of four categories of testlets. The "pictorial form" bases its stimulus for questioning on pictures, maps, graphs, figures of data, photographs, art works, and the like. The "interlinear form" consists of a single passage with a number of denotations that provide an opportunity for questioning such as grammar error analysis in writing tests. The "interpretive exercise" uses a stimulus to set the stage for interpreting questions. The "problem-solving scenario" contains a problem and questions aimed at various steps in the solution of the problem (Haladyna, 1992).

Insert Figure 1 here

The logical method classifies testlets into two categories, linear and hierarchical. By Wainer and Kiely's (1987) definition, each item is embedded in a pre-developed testlet, carrying its own context with it. If the paths through a testlet lead examinees to successive items of greater or less difficulty, depending on their previous responses, and culminate in a series of ordered score categories, it is called a hierarchical testlet (Figure 2).

Insert Figure 2 Here

In Figure 2, Item 2 is supposed to be an item of medium difficulty. If it is answered correctly, the student will be presented with a more difficult item (Item 3); otherwise, Item 1 follows. At level II, the final outcome for answering Item 3 correctly is outcome A; while an incorrect answer results in outcome B. The same process is true for item 1. If the examinee answers the item correctly, outcome C will be the result, otherwise, outcome D will be the measurement score.

If a testlet contains a single path of several items that is administered to all examinees, it is called a linear testlet (see Figure 3).

Insert Figure 3 Here

In this case, all examinees are exposed to the same items without discrimination. Depending on the purpose of the test, the two forms may be combined to construct mixed formats of testlets. Nevertheless, in most cases, testlets are constructed in the linear form. Hierarchical forms are more often used in adaptive tests.

Characteristics of Testlets

One major characteristic of a testlet is that it can be adapted to all types of tests, such as mathematical problem solving, scientific problem-solving, statistical reasoning, essay, performance-type activities, and higher-order

thinking. Because of this compatibility, testlets provide an effective setting that allows the test developer to present relatively complex topics and to ask meaning-construction questions. Usually in the one-item or independent question format, a test developer can ask only simple and straight forward questions and the essence of the item is in the stem. One has very limited room to provide necessary background information or "raw material" with which an examinee can show his or her abilities to interpret, synthesize, organize, and evaluate in solving a problem. Various item forms and modes of presentation make the testlet a popular format because it is not only effective and flexible in providing a whole picture of a problem, but also in assessing different aspects of an examinee's knowledge of a topic. Thus, this format provides a more coherent measure of a larger set of skills than is ordinarily possible with an item-base format. Frequently, it is found that test developers and test takers have different perceptions of a problem, which makes many examinees perform unsatisfactorily. Testlets reduce ambiguity by providing a common ground of information more detailed than that of independent items, and by controlling the amount of factual information given to the examinees. Further, it allows the test builder to provide guidance through a complex problem by suggesting, with the judicious use of subproblems, a path toward the solution of a larger question. These suggestions and subproblems can provide both instructional help and an explicit framework for awarding partial credits

through polytomous scoring procedures (Wainer, Kaplan, & Lewis, 1992).

However, despite its wide application, the testlet has its own special problems. First, it is very difficult and time-consuming to develop testlets of high quality, especially those dealing with complex topics. It is not uncommon for original passages to be revised numerous times to satisfy the specifications of content, level of difficulty, and the outcomes of assessment required for use in real tests. Secondly, it takes considerably longer to administer testlets than to administer independent multiple-choice items because testlets require comprehensive interpretation ability. Since a testlet usually tests multiple abilities of an examinee, understanding the problem becomes essential. Thirdly, it may require that an examinee possess comprehensive reading ability. Often a testlet of moderate length is at least as long as a lengthy independent multiple-choice item. Lastly, because of the time factor, the number of items for a given testlet is restricted to a certain degree, which may cause a reduction in the reliability of the test (Mehrens & Lehmann, 1984).

Testlet Construction and Development

Structures of testlets have changed considerably with the development of testing and measurement. Two frequently used forms in the early development of testlets are option-sharing and alternative response items. The following

examples show their formulations.

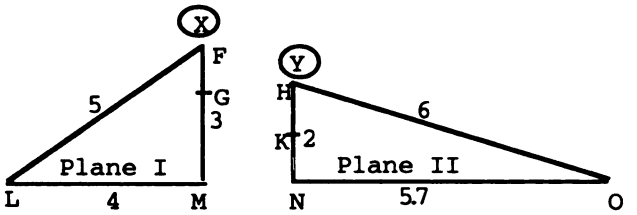
Example 2: Energy Conservation Exercise in Science

Directions: The numbers preceding the paired items in the exercise below refer to the corresponding numbers on the answer sheet. Considering each pair from the standpoint of quantity, blacken space

A, if the item at the left is greater than that at the right.

B, if the item at the right is greater than that at the left.

C, if the two items are of essentially the same magnitude.



Two spheres, X and Y, of equal masses and radii are placed on two inclined planes, as shown in the diagram. Neglect friction and air resistance, and assume that potential energy is measured from the level of points L, M, N, and O.

70. Potential energy of X at F - Potential energy of Y at H.
71. Potential energy of X at M - Potential energy of Y at N.
72. Potential energy of X at M - Potential energy of X at L.
73. Kinetic energy of X on rolling to L - Kinetic energy of X on falling to M.
74. Kinetic energy of X on rolling to L - Kinetic energy of Y on falling to O.
75. Work done on X in raising it from M to F - Work done on X in moving it from L to F.
76. Work done on X in raising it from M to F - Work done on Y in raising it from N to H.*

* Other items of the series involved comparisons with respect to acceleration, time, loss or gain in potential or kinetic energy, power, force, mechanical advantage, and mechanical efficiency. The exercise as a whole requires the application of numerous principles of mechanics. (Engelhart, 1942, p. 110)

In the next example, the item stem is followed by several sentences the pupil is expected to classify according to their degree of causal relationship to the common stem.

Example 3: Sample of Test Exercise in Science

Directions: In the following examples, the first part is followed by *several* OTHER parts. Your job is to find out if the first part is a direct cause or an indirect cause or if it is not a cause of the other parts that follow it.

If the first part *directly* causes the second (numbered) part, draw a circle around the letter D.
If the first part *indirectly* causes the second (numbered) part, draw a circle around the letter I.
If the first part is in *no way* a cause of the second (numbered) part, draw a circle around the letter N.

A girl chews a cracker.
D I N 64. The cracker is broken into smaller pieces.
D I N 65. The starch in the cracker changes into sugar.
D I N 66. The girl gains energy from the cracker.
D I N 67. The cracker is salty.
(Gerich, 1956, Excerpt 106, p. 112)

Example 4 is taken from the *GRE Educational Test Sample Test* (1989),

Example 4: GRE Education Test Sample Question:
The following people have been involved in educational innovations and/or research that have aided curriculum planning and learning. Select the person who is associated with the accomplishments in each of the questions below.

(A) Jean Piagét
(B) Robert J. Havighurst
(C) B. F. Skinner
(D) Jane Mercer
(E) Ned Flanders

66. Established the basis for teaching machines and other programmed learning.
67. Emphasized the importance of concrete objects as instructional materials in the education of young children.
68. Developed a system for analyzing the interaction of students and teacher.

Example 2 is in pictorial form. The graph and description of conditions to solve the problem are presented at the beginning of the problem. The examinee is supposed to match each of the following seven items to any one of the earlier mentioned conditions. Example 3 tests an examinee's ability to understand cause-effect relationships. The stem is very short, one simple sentence, but the directions are relatively long. The alternative responses in this testlet

were "direct," "indirect," or "no relationship." Example 4 starts with the options and is followed by three questions sharing the same options.

It can be seen that the alternative response form requires directions for each testlet, which is not efficient in the test construction. While MC items, however, do not need directions to set up conditions, they do require more space and more higher-order thinking skills to solve the problems (see Example 5 on the next page).

The main differences between constructing testlets and traditional MC item writing reside in the selection of appropriate introduction material and construction of items relating to that material. Strategically, the two parts should be developed simultaneously, since selecting the introduction material is similar to selecting the topics for individual items and the introductory material is crucial to the quality control of the testlet.

Example 5: Presidential Electoral Votes in the United States
By Political Parties
1904 - 1944

Year	Republican	Democratic	Progressive
1904	336	140	
1908	321	162	
1912	8	435	88
1916	254	277	
1920	404	127	
1924	382	136	13
1928	444	87	
1932	59	472	
1936	8	523	
1940	82	449	
1944	99	432	

1. Which party held the presidency during 1926?
1) Republican
2) Democratic
3) Progressive
4) The table does not tell

2. In what year was the Republican victory the most decisive?
1) 1904
2) 1924
3) 1928
4) 1936

3. Which of these statements about Democratic party strength is supported by the table?
1) The Democrats won easy victories in both 1912 and 1916.
2) The Democrats have been by far the strongest political party since 1904.
3) Democratic party strength has been slowly increasing since 1932.
4) Democratic party strength has been slowly decreasing since 1936.

4. Between which two consecutive elections was there the greatest increase in the number of Democratic electoral votes?
1) 1908 and 1912
2) 1912 and 1916
3) 1928 and 1932
4) 1932 and 1936

5. The percentage of the electoral votes received by the Democrats was the *largest* in what year?
1) 1944
2) 1936
3) 1928
4) 1912

(Ebel, 1951, p.243).

Evaluation of Applications of Testlet Assessment

Discussion of the testlet was mostly limited to its form and construction in the early literature. Issues of its application have emerged in recent studies. Szeberënyi and Tigyi (1987) described their employment of the testlet (they called it an "application test") as a problem-solving exercise tool for teaching and assessment of competence in a medical biology class. The typical structure of their testlet was somewhat similar to that of a scientific paper. The objectives of the experiments presented in the testlet were summarized in a short introduction with a brief description of methods. Experimental data were presented in the text, in a table or in pictorial form. A typical test contained 4-6 testlets, each with 10-15 MC items, and was concluded by a discussion of the results. An important feature of their testlet test was that it was an open-book examination. Students were allowed to use any source of information (textbook, lecture notes, research papers, etc.) to eliminate assessing sheer factual knowledge from the test and to guarantee testing problem-solving skills to some extent. As a result, a test usually took three hours to finish. Szeberënyi and Tigyi (1987) stated that their experience of 12 years in using testlets was very successful. They thought that testlets were valuable tools to assess higher levels of the cognitive domain at different levels of difficulty and could be used for teaching. Factual knowledge in a testlet was necessary but not sufficient to solve the problems. As for

students' feedback, the majority of students liked testlets as learning aids and accepted them as a form of examination.

Wainer and Lewis (1990) investigated three different applications of testlet assessment and described psychometric models that they considered to be most suitable for each application.

One application was drawn from *Using Baysian Decision Theory to Design a Computerized Mastery Test* (Lewis and Sheehan, 1988), which employed the *Test of Seismic Knowledge* developed by ETS for architectural certification. Since it was a "pass-fail" test, the study focused on testlet difficulty in the region around the decision point.

The item pool consisted of 110 items. Sixty percent of the items dealt with physical and technical aspects of seismic knowledge (Type 1 items), and 40% covered economic, legal, and perceptual concepts (Type 2 items). The goal of the study was to create testlets that could be interchanged randomly while retaining unbiasness and measurement accuracy (the degree to which the selected testlets varied with respect to the average likelihood of a particular number-right score). The item pool was divided into 10-item testlets, with each testlet balanced for content and equal in average difficulty and discrimination. The testlets were constructed by cross-classifying the item pool by item type and estimated item difficulty. After testlet selection, the experts in the subject field edited the final version. The validity of the testlet interchangeability assumption was

evaluated by determining the degree to which the six selected testlets varied with respect to the average likelihood of a particular number-right score. Likelihoods were evaluated at five different points on the latent proficiency scale which corresponded to five important decision points surrounding the anticipated cutscore. This validity check shows that, for examinees near the cutscore, the average number-right score has about the same probability regardless of which testlet was administered.

After completion of a testlet presented to an examinee, a pass or fail decision was made by a statistical determination. It was expected that the number-right score approach carried all the information necessary to implement the Baysian decision process that was employed in the application. The tests allowed test developers to simultaneously maximize the probability of classifying individuals and minimize the amount of testing.

The second application, conducted by Thissen, Steinberg, and Mooney (1989), used traditional reading comprehension items as linear testlets and applied an adapted IRT model in a testlet-level analysis. Items were from IRT scored computerized adaptive tests and were used to study possible violation of the local independence assumption when several items shared the same stem. In the formulation, Thissen et al. (1989) considered the examinees' responses to m questions relating to the same passage as a polytomous response and then scored it either 0, 1, 2, ..., or m, depending upon how

many of *m* questions an examinee answered correctly. They compared the results of a 22-item test where the items were first treated as independent items with the results from four testlets grouped by four passages by these items. The reading passages varied from one to six paragraphs and were followed by three to eight questions about the content. In addition, the authors evaluated the concurrent validity of these four testlets' scores with that of 54 other independently scored items in the same test.

The Thissen et al. study used a testlet response model proposed by Bock (1972) for responses of two or more nominal categories for each passage. The model required conditional independence between testlets only, not within them. The testlets were formed linearly and administered linearly. The traditional 3-PL IRT model was used to score the passage items as if they were independent. The results showed that the 3-PL scoring appeared to provide substantially more information over most values of the latent trait, especially at the positive side of its continuum. However, the concurrent validity study with the statistical program LISREL (Jöreskog and Sörbom, v. 7, 1984) showed that the four testlets' scores were slightly but significantly superior to the 3-PL scores ($\chi^2_{(1)}=8.8$, $p<.003$) with an external criterion, the raw score on a simultaneously administered 54-item test of verbal proficiency. Thissen et al. (1989) found that the information curve computed from the 3-PL model when its assumption of local independence was violated was deceptively

high. They considered that this phenomenon was "fooled" by the excess intra-passage correlation among the items and that the 22-item test was estimated to be more precise than it actually was. The testlet scores appeared to be at least as valid, if not slightly more so, as the 3-PL model scores.

The third example of testlet application, called *Validity-Based Scoring* (Lewis, 1989), was an alternative approach to IRT. The method was based on the assumption that it was possible to obtain information on some criterion measure(s), at least for a calibration sample of students. *Validity-Based Scoring* assigned the predicted values on the criterion as the scores for each possible outcome for the testlet. These scores were simply the mean criterion values for the group of students with each given testlet result. The group standard deviations on the criterion variables may be interpreted as conditional standard errors of prediction for these scores. Two hierarchical testlets related to elementary algebra were constructed by ETS and were administered linearly. An adaptive approach was employed for working on a testlet, in which a more difficult item followed a correct response, while an easier item followed an incorrect response. The students' responses to the items in the testlets were used to group students and were treated as indicator variables that were then used as the predictors for criterion measures in the sample of students. If the scores for the groups did not reflect theoretical ordering of the response groups, or if differences between scores for

adjacent groups were small relative to the standard errors, follow-up diagnostics was explored. Compared with IRT, the advantage of this approach was that it gave information directly relevant to the test used in the prediction of a relevant criterion. However, it was strictly data-driven without any theoretical basis.

From the studies described above, Wainer and Lewis (1990) concluded that a testlet formulation could provide a more precise estimation of test quality to allow the use of powerful statistical sequential decision-making and to help develop more efficient tests. They emphasized that the testlet scoring must be fully integrated with a validity criterion since this was the most important characteristic of a test. Specifically, IRT and testlets were two notions that were somewhat independent. One could use the testlet approach, even in an adaptive mode, without recourse to IRT at all (the sesmic knowledge test). Or one could tie the testlet's construction and scoring intimately to IRT (the paragraph comprehension test). Or one could choose between the two and use IRT to construct the testlets, but not use IRT in the scoring (the *Validity-Based Scoring* example).

Wainer, Lewis, Kaplan, and Braswell (1991) employed both hierarchical and linear models to construct two 15-item testlet-based tests on basic algebra skills and factoring skills. They focused on the amount of information that could be obtained from a testlet of moderate length, as well as on the gains and losses associated with making the internal

structure of the testlet adaptive. The two tests were administered to 2,080 ninth and tenth graders. The test results were evenly and randomly divided into two sets, with one set serving as the exploratory sample and the other as the confirmatory sample, later used for cross-validation. The data were fitted with a 3-PL model using marginal maximum likelihood. A value of each examinee's proficiency (β) was estimated for the entire 30-item test. Items to form a testlet were chosen in two ways. The first was the stepwise, optimal tree with replacement, in which the hierarchy was formed first by selecting the item that yielded the minimum posterior variance of the two groups. The second item was chosen when its addition to the first one minimized the variance. The process continued until a four-item testlet was reached. Choosing the best 4-item testlet (fixed format) was the second procedure, in which all combinations of 4-item testlets, 1,365 of them (15-choose-4 combinations = 15!/11!4!) were examined and the one that performed best on the same criterion used for constructing the tree was selected. The criterion was to predict the examinees' proficiency estimated on all 30 items from a 4-item testlet. That is, the authors of the study estimated proficiency (β) on the entire pool and then tried to predict it as precisely as possible from various 4-item testlets. This procedure produced a proper subset of optimal trees of method one, yet it allowed much simpler technology (paper-and-pencil) than any other adaptive test.

In the Wainer et al. (1991) study, hierarchical and linear formats were also compared with each other. It was found that, although a hierarchical testlet was superior to a linear testlet, the increased information was modest in most places along the proficiency continuum, except when β=-.5 or β=.25, where adaptive testlets provided considerably more information than the fixed testlets, but at a high cost. It was concluded that, in situations similar to those described in the study, the fixed format (choosing the best testlet) could produce as good a testlet as the optimal adaptive testlet of equal length from the same pool. In addition, the authors recognized that although no major decisions could be made on a 4-item testlet performance, many small decisions were possible. The study emphasized the posterior variance of the items without indicating whether the items configurating the testlet derived from the same content or paragraph. This did not match Wainer and Kiely's (1987) definition of testlet, where a group of items had to be related to a single content area.

In summary, this section described studies that evaluated testlet assessment effects in the classroom setting and in educational measurement experiments. Researchers in those studies found that testlets as an item format, if analyzed as a unit, can provide more information about the examinees. Testlet scores can provide as valid person estimates as a dichotomous IRT 3-parameter model did. In addition, when the method was applied to a classroom setting,

it enhanced student learning. Although most of these studies did not directly relate to the local item dependence issue, they provide background information as to what kinds of experiments have been done with applications of the testlet assessment and to direct people's interests to study other issues related to the testlet assessment such as local item dependence.

Local Item Dependence Effects

One of the psychometric characteristics that researchers have discussed extensively is the loss of local independence when items are related to the same topic but are scored individually. In IRT, the assumption of local independence implies that an examinee's responses to different items in a test are statistically independent for a given ability. For this assumption to be true, an examinee's performance on one item must not affect, either for better or for worse, his or her responses to any other items in the test. When local independence exists, the probability of any pattern of item scores occurring for an examinee is simply the product of the probability of the occurrence of the scores on *each* test item. For example, the probability of the occurrence of the five-item response pattern $\mathbf{v}$ = (1 0 1 1 0), where 1 denotes a correct response and 0 an incorrect response, is equal to $P_1*(1-P_2)*P_3*P_4*(1-P_5)$, where P_i is the probability that the examinee will respond correctly to item i and $1-P_i$ is the probability that the examinee will respond incorrectly,

usually represented by Q_i. In general terms, local independence can be expressed symbolically as the following:

$$P(V_1=v_1,\ V_2=v_2,\ \ldots\ ,\ V_n=v_n|\beta)$$
$$= \prod_{n} P_1(\beta)^{v_1}Q_1(\beta)^{1-v_1}P_2(\beta)^{v_2}Q_2(\beta)^{1-v_2}\ \ldots\ P_n(\beta)^{v_n}Q_n(\beta)^{1-v_n}$$
$$= \prod_{i=1} P_i(\beta)^{v_i}Q_i(\beta)^{1-v_i}, \text{ where} \qquad (1)$$

v_i represents the binary responses,

β is a student's latent ability,

P_i is the probability of an examinee answering the *i*th item correctly,

$Q_i=1-P_i$, is the probability of an examinee answering the *i*th item incorrectly, and

$i=1, 2, \ldots, n$, is the item.

In other words, the assumption of local independence applies when the probability of the response pattern for each examinee is equal to the product of the probability associated with the examinee's response to each item.

It should be mentioned that the assumption of local independence for the case when β is unidimensional and the assumption of a unidimensional latent space are equivalent. Suppose a set of test items measures a common ability for examinees at a fixed ability level (β). If items are not statistically independent, it would imply that some examinees have higher expected test scores than other examinees of the same ability level. Consequently, more than one ability would be necessary to account for examinee test performance. As a result, the test becomes multidimentional. Since local independence assumes that the item responses are

statistically independent for examinees at a fixed ability level, only one ability should be accountable for the relationship among a group of test items (Hambleton & Swaminathan, 1985). It is also important to note that the assumption of local independence does not imply that test items are uncorrelated over the total group of examinees (Lord & Novick, 1968, p. 361). Positive correlation between pairs of items results whenever there is variation among the examinees on the ability continuum measured by the test items, but item scores are uncorrelated at a fixed ability level.

Cattell and Burdsal (1975) asserted that the individual item responses were not very reliable for measuring human behavior because of their poor repeat reliability (i.e., low dependability coefficient) and vulnerability to cultural localism (i.e., low transferability coefficient). They thought that parcels (i.e., testlets) composed by apparent content or by actual correlations within a personality sphere of items were also defective because of their subjectivity or the ambiguity of real correlations between the pairs of items. They introduced the concept of "radial parceling" in the context of personality measurement and rating scales. The essential difference of the radial parcel method from the usual clustering of items was that the number of items in a parcel or testlet was not predetermined. Instead, it required two factor analyses, first at the item level and then at the parcel level. The first factoring *yielded* the parcels. That

is, the factor analysis was conducted on the items without considering their contents, just to get a general grouping of items into parcels. The second factor analysis was conducted on the parcels to make precise factors extracted from the first analysis. The goal of this method was to obtain an invariant and maximally homogeneous solution of a common factor space. However, the method was rather complicated and no other study has ever used it.

Rosenbaum (1988) used unidimensional IRT to describe observable item response distributions when there was conditional independence between, but not within, the testlets. He contrasted the behavior of the population distribution of item responses, $P(X=x)$, under two sets of assumptions. One set was based on the conventional IRT assumptions, which were: (1) for a test containing J dichotomously scored items, $\mathbf{X} = (X_1, X_2, \ldots, X_J)$ was the response variable, $P(\mathbf{X}=\mathbf{x}|B=\beta)$ was assumed to have a simple structure, where β was a latent variable, (2) item responses are conditionally independent given β, and (3) correct responses are more common among examinees with higher values of β. Expressed symbolically, they were:

$$P(X=x) = \int P(X=x|B=\beta)\,dF(\beta), \qquad (2)$$

$dF(\beta)$ is the population distribution of B, which is normal,

$$X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_J \,|B, \text{ and} \qquad (3)$$

$$P(X_j=1|B=\beta) \text{ is nondecreasing for } j=1,\ldots, J \qquad (4)$$

The other set of assumptions was similar to the first set except that a weaker version of the second assumption was

replaced. It allowed dependence among items that shared a common prompt such as a reading passage or a graph. In other words, responses to items in the same testlet may demonstrate dependence even among examinees with the same level of latent variable. By presenting an example of the results of applying the Mentel-Haenszel statistics to all $\binom{40}{2}$=780 pairs of MC items in the 40-item biology subscore of the College Board's 1982 *Advanced Placement Examination in Biology*, Rosenbaum (1988) delineated theoretically conditional independence and monotonicity for "bundle items" (testlets) and observation distributions. It was found that each negative partial association in the subscore violated traditional IRT assumptions. Alternatively, the weak assumption of conditional independence was explored, in which items sharing the same material were bundled together and the original assumption of $X_1 \coprod X_2 \ldots \coprod X_J | \beta$ was replaced by $V_1 \coprod V_2 \ldots \coprod V_I | \beta$, where $V_i = (X_1, X_3, \ldots, X_i)$ represented a group of items in a bundle. This meant that responses to items in the same bundle may exhibit dependence even among examinees with the same values of β, possibly because some examinees had more difficulty understanding a particular reading passage or a graph, and, therefore, they may have had more difficulty with all items relating to that passage or graph. Rosenbaum (1988) further proved that, with the nondecreasing assumption, every pair of items in the same bundle had a non-negative (population) correlation at a given β. However, he proved a theorem mathematically that at every level of β, the

standard error of measurement (SEM) under a positive bundle model was at least as large as a conventional IRT model having the same item characteristic curves. Informally, positive dependence within bundles increased the SEM at every level β of B. The theorem suggested that other things being equal, it would be preferable not to use bundles of positively dependent items when designing a test since doing so may cause a larger SEM. Similarly, using a conventional IRT model for a test with bundled items may lead to an undue underestimate of SEM. The principal finding in this paper was that dependence within such testlets has predictable and testable consequences for the population distribution of item responses.

Does Rosenbaum's (1988) conclusion mean that results based on conventional IRT models with smaller SEM's are more reliable, and therefore more highly correlated with other measures than the testlet scores? No. Sireci, Thissen, and Wainer (1991) compared two pieces of research in their study of reliability estimation calculated on two reading comprehension tests constructed by testlets with both traditional true score and IRT methods. In the first study, they found that, when items were used as the unit of analysis, reliability values ranged from .86 to .88 for both methods. When testlets were used as the unit of analysis, reliability values estimated ranged only from .75 to .80. They concluded that the item responses within passages were more highly correlated than were item responses between

passages. Failing to take into account the dependencies caused by having four sets of items, each set referring to a common passage, yielded a 10-15% over-estimation of reliability.

The second relevant study was the one that Thissen et al. conducted in 1989, discussed earlier in this chapter. Again, the item level reliability (.70 for traditional α and .74 for IRT for 22-items) was 0.08 higher than the testlet reliability (.62 for traditional α and .66 for IRT). The results implied that in IRT, only when the items were locally independent did the product of the item trace lines provide a precise description of the posterior density for examinees with that response pattern. Item-based marginal reliability provides a precise estimation of the average variance of these posterior densities. If local independence only held between some larger units of the test (e.g., testlets), then trace lines for those units were multiplied to produce the posterior densities, and the correct estimate of (marginal) reliability was based on those trace lines. They concluded that if a test was constructed of testlets, the within-testlet structure must be taken into account in the calculation of test statistics. Failing to do so may yield serious biases in estimating some statistics such as reliability. Their study showed that traditional reliability calculated on two reading comprehension tests composed of four testlets was substantially over-estimated.

Studies so far have been using the number-right score of testlets as the testlet value to evaluate the relationship between local independence and test statistics such as test information. Ercikan (1993) brought up the issue of information loss when using the lump sum score for a testlet without considering the response pattern and examined the change in measurement precision when the sum of item raw scores was used in the testlet methodology. The study focused on the effect of the number of testlets and the number of items within a testlet on test information when the sum of item raw scores was used as the testlet response. Data were drawn from two 5th and 8th grade constructed-response mathematics tests and one 10th grade MC mathematics test. The test responses were calibrated with random sample sizes ranging from 3,000 to 7,000. In particular, only the locally independent items were grouped to form testlets to avoid the confounding effect of locally dependent items on test information for each type of test. Information values for tests with different numbers of testlets and numbers of items within testlets were compared to those from the original tests without testlets. For the MC item test there were 19 versions. Version A lacked any testlet and the rest had one to four testlets. The number of items within a testlet ranged from two to eight. For constructed-response tests, test 1 had 17 versions and test 2 had 19 versions. Again, Version A was the non-testlet version. The maximum number of testlets in a version was five for test 1 and nine for test 2. Items within

a testlet ranged from two to eight for all versions of both tests. Regarding calibration models, the 3-PL IRT model was used in MC test versions and the 2-PL partial credit model (Yen, 1993) was used for constructed-response versions.

Most testlet versions based on constructed-response tests resulted in reductions in test information. There was, however, not a clear trend for change in test information when greater numbers of testlets were formed with an equal number of items. For most of the MC versions, test information was increased. The mean scale score difference between testlet versions and the original test version (Version A) was small. The correlations of scale scores between the testlet versions and Version A were all very high. There were a few cases where a large difference in scale scores between different testlet versions and Version A was observed. However, the results did not provide information about what kind of changes in test information should be expected if all testlet response patterns were used as different indicators of ability instead of the sum of item raw scores within testlets.

According to Wainer and Kiely (1987), the main purposes of using the testlet were two fold: control and fairness. Control meant that by defining the exchangeable amount of test construction as something larger than the item, the test developer could recover some of the control over the structure of the finished test that was relinquished when it was decided to use an automatic test construction algorithm.

Fairness meant that all examinees were administered the same sets of items and therefore, the comparison was made on scores derived from tests of very similar contents. The real reason that testlets were developed was because of the internal relationship among those items in a testlet. If testlets were formed with locally independent items instead of locally dependent items, the original goal of adapting testlets may not be met and the results from the study may lack validity.

Yen (1984a) examined the effects of local item dependence on the fit and equating performance of the 3-PL model in the analysis of unidimensional and two-dimensional simulated data and in the analysis of real data of three mathematics achievement tests at grade 3 and grade 6. The simulated data used item parameters from three different configurations to design the multi-dimensional tests. In the real data, items were grouped into sets that appear most likely to show local dependence. The fit measures of local dependence were Q_2 and Q_3 for both simulated data and real data. Q_2 was proposed by van de Wollenberg (1982) for the Rasch model. It takes the form of a Pearson chi-square statistic to examine local dependence for pairs of items and it is sensitive to multi-dimensionality. Q_3 fit statistic (Yen, 1984a) calculates the correlation between two items by removing the nonlinear effects of person ability from the item scores. As a result, the statistic examines local dependence with correlation of examinees' random error scores

of these two items. Local dependence is suspected if the correlation is significantly different from zero. Yen (1984a) pointed out that local dependence had direction when a test is multidimensional. Positive local dependence occurred when two or more items measure special traits that did not appear in the rest of the test, while negative local dependence could appear between two sets of items that measured different traits. Results from two-dimensional simulated data showed that

> "... If a combination of two underlying traits is used as the unidimensional trait, then items that are influenced by both underlying traits will show negative local dependence and items that are influenced by only one underlying trait will show positive local dependence. If only one of the underlying traits is used as the unidimensional trait, then items that are influenced only by that underlying trait will show slight negative local dependence due to part-whole contamination and items that are influenced by both underlying traits will show positive local dependence (p.142)."

For real data, the items with high Q_2 and Q_3 values tended to have similar item parameters, but this was not necessarily true vise versa. Locally dependent item sets of *Mathematical Computation* seemed to be slightly more difficult and discriminating if the items that accumulated in skills (e.g., be able to calculate addition before computing multiplication or division) involved easier items. However, most *Mathematics Concepts and Application* items of high local dependence were relatively moderate in both difficulty and discrimination parameters. In addition, substantial unsystematic errors of equating were found from the test of multi-dimensions. Systematic errors of equating were only found when two tests

measured different dimensions but were taught sequentially.

Yen (1993) later pointed out that the basic principle in producing local item dependence was that there was an additional factor that consistently affected the performance of *some students* on some items to a greater extent than others. Factors such as external assistance or interference, test speediness, fatigue, practice, item or response format, passage dependence, item chaining, explanation of previous answer, scoring rubrics or raters, content, and knowledge and abilities, all were possible causes of item dependence. She further discussed some measurement implications when items were locally dependent. One implication was for performance assessment. In measurement of educational achievement, while MC tests usually focused more on developing discrete items that were closely tied to objective structures and separating performance into pieces, performance assessment tests embraced measuring a behavior as a whole. If measuring a behavior as a whole was the goal of the assessment, then one item may be sufficient to achieve that purpose; otherwise, independent items should be used. Another measurement implication that was cited frequently was test information and standard error of measurement due to local dependence. The third important measurement implication was test validity. Since the validity of a test score impacted the appropriateness of decisions made, it was desired that decisions be broadly based and that the conclusions cover a variety of situations. In order to generalize these results

to different real life behaviors of typical interests, samples of observations should be as independent as possible. If items were locally dependent, it meant that individual observations covered a range of behaviors that was smaller than was attempted.

Yen's (1993) empirical study compared MC tests of CTBS/4 and performance assessment data of the *Maryland Performance Assessment Program* in grades 3, 5, and 8 in 1991 with a two-parameter partial credit model, a special case of Bock's (1972) model. Item information and discrimination estimates obtained by a testlet-base scale and by an item-base scale in reading and math test items were compared. It was found that for both reading and math, locally dependent testlets were about one third lower than non-dependent testlets in relative efficiency and in ratio of mean item discrimination, and were only about 60% of the non-testlet items in those two values. As for SEM, testlets do result in larger SEMs, but it could be seen as a reflection of reality. However, in many cases, there was not much difference in parameter estimates when items were scaled independently or as a testlet. This implied that small discrepancies between different scalings would not affect test score precision practically. In addition, testlet trace lines were strongly affected by local item dependence for locally dependent testlets in both directions, while local item dependence had almost no effect on item characteristic functions. In order to better manage locally dependent test items, Yen (1993), in addition to other

strategies, suggested using testlets as an alternative to minimize the local item dependence effects:

> "One of the major advantages of testlets is that they do not interfere with the design of authentic tests that are intended to involve dependent items. The testlets provided a more accurate description of the item trace lines and the information provided by the items in a test (p. 212)."

Summary

Testlet concepts have been applied widely in regular classroom testing, computerized adaptive testing, and non-traditional, non-IRT scoring for a long time. Its main forms include option-sharing, picture or table, alternative response, and stem or passage sharing. The most evident advantage to using a testlet in a test is that it can provide a more authentic situation in which to examine and assess more complex abilities of examinees.

Evaluation of testlet assessment shows that testlets as an item format, if analyzed as a unit, can provide more information about examinees. Its scores can be as valid as a conventional IRT 3-parameter model in person ability estimation. It also enhances student learning.

In IRT, local independence means that an examinee's responses to different items in a test are statistically independent. When the assumption does not hold, the examinee's performance on one item may affect his or her responses to other items in the test. Factors such as item format, passage dependence, fatigue, knowledge and ability, are possible causes of item dependence.

It was found that although item-based parameter estimation appeared to provide more information over most levels of the latent trait continuum, this extra gain in information may be "fooled" by the excess intra-passage correlation among the context-dependent items. In other words, the test information value is over-estimated. This situation was especially true when the assumption of local independence was violated.

It was also found that local item dependence has direction. Positive local dependence occurs when two or more items measure special traits that do not appear in the rest of the test, while negative local dependence can appear between two sets of items that measure different traits. Studies reviewed previously showed that if a test is constructed of testlets, the within-testlet structure must be taken into account when calculating test statistics. Failing to do so may yield serious biases in estimating statistics such as reliability. As for SEM, testlets do result in a larger SEM because of the within-testlet structure, but it can be seen as reflection of reality.

De Ayala et al. (1988) and Yen (1993) pointed out in their studies that, though some research had been done on polytomous scoring methods, these models needed to be studied more in large-scale assessment programs. One strategy that was suggested was using testlets to manage the local item dependence situation.

The present study tried to apply the family of the Rasch models to testlet cases in a large-scale assessment program, and to estimate the person measure, item calibration, and testlet fitness when the violation of the assumption of local independence was controlled by the testlet. It is hoped that the study results will help explain whether local item dependence has any effect in the person and item/testlet parameter estimation to the tests that are similar to the one under study.

CHAPTER 3

METHODOLOGY

Overview

The purpose of this study is to use context-dependent testlets as the unit of analysis to detect local item dependence effects. The Rasch dichotomous scoring model (DM), where items within a testlet are analyzed as independent items, is compared with the Rasch partial credit scoring model (PCM), where these items are analyzed as a holistic unit. The question of whether potential local item dependence in an 11th grade science proficiency test exists and can be controlled by using the testlet as the unit of analysis is discussed. Further, different estimation measures on person ability, person and item/testlet fit, measurement/calibration errors, and test reliability for the analyses are described. The testing materials, the data, the sampling procedures, the research hypotheses, and the calibration models are presented. Also described in this chapter are the analysis plan and the computer software program used for testing the hypotheses.

Testing Materials

Science Assessment Framework

The *Michigan High School Proficiency Test in Science* (MHSPT) was constructed within the framework of the *Assessment Frameworks for the MHSPT in Science* (Michigan State Board of Education, 1994), which was developed by the Michigan Science Teachers Association under contract with the Michigan Department of Education. The panel of the framework development consisted of science teachers, a special education teacher, school administrators, assessment specialists, and university scientists. A broad representation of Michigan's educational community was involved in the project.

In 1991, the Michigan State Board of Education adopted the *Michigan Essential Goals and Objectives for Science Education (K-12)* (Michigan State Board of Education, 1991). The Michigan Legislature Public Act 25 (1990) required that the above document and the *Model Core Curriculum Outcomes* (Michigan State Board of Education, 1991) serve as the curriculum foundation for science education. The items/exercises of the science proficiency test should then be generated to measure those outcomes and objectives. These two documents were designed to identify what scientifically literate persons should know and be able to do. The science curriculum objectives are organized in three subject matter areas: life science, earth and space science, and physical

science. The activities based on the outcomes and objectives are categorized as *using scientific knowledge*, *constructing scientific knowledge*, and *reflecting on scientific knowledge*. Each of the categories is briefly described below.

Using scientific knowledge means students can use their knowledge of life, space and earth, and physical sciences as reflected in the essential goals and objectives to understand the world around them and to guide their actions. They can describe and explain real world objects, systems, or events, predict future events or observations, and design systems or courses of action that enable people to adapt to and modify the world around them.

Constructing scientific knowledge means that students can develop solutions to problems they encounter and learn by interpreting text, graphics, tables, pictures or other representations of scientific knowledge.

Reflecting scientific knowledge means students can "step back" and analyze or reflect upon their own knowledge and justify personal knowledge using either theoretically or empirically based arguments and describe the limitations of their own knowledge and scientific knowledge in general.

One major purpose of the framework is to give clear direction to persons developing the MHSPT in Science and to provide detailed information on both the core outcomes and all of the essential goals and objectives under each topic (*Assessment Frameworks for the Michigan High School*

Proficiency Test in Science, 1994) (See Figure 4).

Insert Figure 4 Here

The Test

The purpose of the *MHSPT in Science* is to determine the extent to which, at the end of 10th grade, a student has achieved scientific literacy in *using*, *constructing*, and *reflecting scientific knowledge*. The test is written to require application of theoretical concepts to real world contexts. The assessment entails a great deal of reading and writing. Although students answer many multiple choice questions, they are also required to write their responses to eight questions. Written responses to questions require students to evaluate and critically analyze scientific investigations and scientific text.

The tryout version of the *MHSPT in Science* was a test of 54 items, which lasted 120 minutes without a break. It consisted of four parts, each part used a specific kind of item format. The configuration is briefly described below:

A. Thirty (30) independent items. Each item poses a single task or question about a specific real world context. It usually assesses one core objective outcome. The purpose is to test a wide designated sample of outcomes.

B. Four (4) cluster problems (i.e., testlets). A cluster problem, according to the HSPT Science Assessment Framework, "presents a real world context (an event, a situation or an object) and asks a series of questions about it" (p. 54). Each cluster problem includes four MC questions and one constructed-response question. There is a cluster from each of the three scientific content areas and one integrated cluster covering two or more science areas. The item distribution for each cluster problem includes at least three items on *using* objectives, at least one item on *constructing* objectives and one on *reflecting* objectives.

C. One (1) investigation. The investigation requires the students to read a report of an experiment conducted by the tenth grade students and to respond to two or three constructed-response questions about the report that will cover *constructing* science outcomes only.

D. One (1) text criticism. The text criticism presents students with a passage to read from the popular press (newspaper or periodical). Students respond to two or three constructed-response questions covering only *reflecting* core outcomes.

Sixty percent of the test items assesses *using* objectives which are distributed equally among life, earth and space, and physical science objectives. Twenty percent

assesses *constructing* objectives and twenty percent *reflecting* objectives. However, *constructing* and *reflecting* objectives do not have to be distributed equally across all three content areas (see Table 1).

Table 1. Michigan Science Proficiency Test Tryout Form Configuration

Science Subject Area	Life Science			Physical Science			Earth Science			Integrated Science			Number of Items
Objective Category	U	C	R	U	C	R	U	C	R	U	C	R	
Testlet Problems (4 multiple-choice and 1 constructed-response)	3	1	1	3	1	1	3	1	1	3	1	1	20
Independent Items	7	2	1	7	2	1	7	2	1				30

Science Subject Area	Life, Physical, or Earth Science		Number of Problems
Objective Category	C	R	
Text Criticism Problem		1	1
Investigation Problem	1		1

A sample testlet is attached in Appendix B.

Tryout Design

The tryout for the *MHSPT in Science* was administered during the week of Nov. 14 - 18, 1994. There were 10 forms (Forms 20-29) in total and no items in common between forms. The forms were organized into four triplets and two quadruplets. The following table displays how the forms were grouped in the data collection design:

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Form 20	Form 22	Form 24	Form 26	Form 29	Form 23
Form 21	Form 23	Form 25	Form 27	Form 20	Form 25
Form 22	Form 24	Form 26	Form 28	Form 21	Form 27
			Form 29		Form 28

The forms within each group were spiraled (e.g., in group 1, forms were ordered repeatedly in Forms 20, 21, and 22 fashion.) and were administered to students within classrooms. By doing so, no two forms were the same for the students sitting next to each other. Each tryout school received only one group of forms. Students taking different forms were considered to form randomly equivalent groups. In addition, each form was administered to two different groups of students. In other words, there were forms in common between groups. This design allowed the equating of forms by the assumption of randomly equivalent groups. (An alternative design of spiraling all forms within schools was not used due to security concerns.)

Data

The data for this study came from the first tryout of the new items for the *MHSPT in Science*. The information from the tryout was used to discard or revise items/exercises as necessary. All ten forms in the tryout attained the full length of the real test. All items were written by the Exercise Development Team (EDT) which was composed of experienced science teachers in Michigan. The items were scrutinized by the Content Advisory Committee (CAC) and the Bias Review Committee (BRC). CAC members consisted of

Michigan science teachers, school principals, local district science personnel, and university science professors. Michigan teachers of different disciplines, university faculties, and Michigan Department of Education staff formed the BRC panel. There were no reliability and validity data or item statistics available at this stage for these items, since this was the first tryout.

Because the focus of this study was on the testlet issues, only the context-dependent MC items within testlets and some independent MC items were studied. The constructed-response questions within the testlets or in other parts of the test were not included in the research because they were hand scored by different scoring rubrics, which may introduce interrater and other kinds of errors that would make the study too complex to be handled. Other item formats such as investigation or text critique questions were not addressed because those items required constructed-responses also.

Sampling Procedures

Cluster sampling in combination with stratified sampling was used in the tryout. By Michigan Legislative act PA 335, 1993, all the 11th grade students in Michigan public schools are required to take the MHSPT in Communication Arts (including Reading and Writing), Mathematics, and Science. Therefore, the target population and the sampling frame was all the 11th grade students in public schools. The enrollment

of the 11th grade students in the fall of 1994 was 106,642. In Michigan, schools are classified into seven strata by the resident population size of the community where the school is located (See Appendix C). Schools participating in the science tryout were randomly sampled from each stratum roughly proportional to the population by the stratum school weight. There were eighty schools with 12,632 students in the total sampled for the science test. When a school was chosen to become part of the sample, all the 11th graders within that school were included. Eight schools declined to participate in the tryout test. Finally there were 10,074 students in total from 72 schools who actually took the science tryout. Table 2 below displays the distribution.

Table 2. Number of Schools and Students Sampled in Science Tryout for Each Stratum

Stratum	Total # Schools	Total # Students	Schools Selected	Schools Part.'d	Students Sampled	Student Weight
1	49	9,935	5	5	1,400	11.1%
2	64	11,465	7	6	1,427	11.3
3	106	23,616	12	12	281	22.6
4	62	10,350	8	6	1,112	8.8
5	7	1,666	1	1	339	2.7
6	232	32,524	26	22	3,372	26.7
7	218	17,086	21	20	2,121	16.8
Total	738	106,642	80	72	12,632	100.0

Item Scoring

All the independent MC items and testlet MC items were scored dichotomously. That is, one point was awarded if a

student answered the item correctly, zero otherwise. For context-dependent MC items, raw scores of each testlet were summed to obtain a testlet score. For instance, each tryout form had 4 testlets, each testlet had 4 MC items, totaling 16 context-dependent MC items for a form. A testlet can have scores x = 0, 1, 2, 3, or 4, depending on how many items a student answered correctly. The maximum testlet score in a form for a student is 16.

Original Testlets vs. Random Testlets and Reformed Testlets

For research purposes, there are three types of testlet configurations in this study: original testlets, random testlets, and reformed testlets. The testlets developed as a result of the Michigan science objectives and outcomes are called *original testlets*. To verify local dependence effects for items within a context-dependent testlet, 16 additional independent MC items in the same tryout form were randomly selected to form 4 new testlets. These testlets are called *random testlets* and were scored the same way as the original testlets. Results from these two kinds of testlets were compared for the local dependence effect. In addition, the original testlets of the same tryout form were broken up and recomposed into four other new testlets, each with an item from an original testlet as if they were in different contexts. These testlets are called *reformed testlets* to be distinguished from the other two kinds of testlets. The

intention of doing this is to see how those context-dependent items perform when they were detached from their original context and were analyzed as if they were in the new contexts. The tryout form and the items themselves will not change, just the item configuration does. Their comparisons with the original testlets were expected to provide more information about the local dependence within a testlet.

Research Hypotheses

As stated in Chapter 1, the research hypotheses are:

1. For context-dependent items,

   (a) the average item correlations within an original testlet are larger than the average correlations with items from other testlet configurations;

   (b) when they are analyzed as a testlet by the Rasch partial credit model, they produce a better testlet fit statistic than when they are analyzed as individual items by the Rasch dichotomous model;

   (c) when they are analyzed as a testlet by the Rasch partial credit model, they produce better person fit statistics than when they are analyzed as individual items by the Rasch dichotomous model;

   (d) when they are analyzed as a testlet, the measurement errors are smaller than when they are analyzed as individual items. In other words, the person separation reliability is higher for testlet-based analysis than for item-based analysis.

2. For independent items,

   (a) when they are analyzed as a testlet by the Rasch partial credit model, the testlet fit statistics are the same as the item fit statistics when they are analyzed as individual items by the Rasch dichotomous model;

   (b) person fit statistics stay the same regardless of whether the items are analyzed as random testlets or as individual items.

(c) The reliability of the person separation ratio is the same for testlet-based analysis and for item-based analysis.

3. When context-dependent items in the original testlets of the same tryout form are decomposed and reformed into the same number of new testlets, each with an item from each original testlet as if they were in different contexts,

   (a) the average correlations between items within a reformed testlet are smaller than the average correlations between items within an original testlet;

   (b) person fit estimated by the reformed testlets are not as good as those estimated by the original testlets.

Calibration Models

The Dichotomous Model (DM)

The dichotomous model (Rasch, 1960) is the simplest form in the family of Rasch models. It is used to estimate person and item parameters when items are scored dichotomously. For a dichotomously scored item i, the model specifies the probability of a correct response to the item as an exponential function of the difference between person ability β_n and item difficulty δ_i:

$$\phi_{nij} = \frac{\pi_{ni(j=1)}}{\pi_{ni(j=0)}+\pi_{ni(j=1)}} = \frac{\exp(\beta_n-\delta_{ij})}{1+\exp(\beta_n-\delta_{ij})}, \text{where} \qquad (5)$$

ϕ_{nij} is the person n's probability of scoring 1 rather than 0 on item i,

β_n is the ability of person n, n=1, 2, ..., N,

δ_{ij} is the difficulty of item i, i=1,2, ..., L,

$\pi_{nij} = \phi_{nij}$ is the person n's probability of scoring 1 on item i, and

$\pi_{ni(j=0)} = 1 - \phi_{ni(j=1)}$ is person n's probability of answering

item *i* incorrectly.

j=0, 1, is the score of item *i*.

Parameters to be estimated in this model are person ability (β_n) and item difficulty (δ_{ij}).

Number of paramters=N+L-1, in which N is the number of students and L is the number of items. For example, for a 16-item test, the total number of parameters = N+16-1=N+15.

According to Masters (1982), the model can separate the person parameter, β_n, from the estimation equation for the items so as to make it possible to estimate item parameters sample free in the calibration. Consequently, the item and person parameters can be estimated on the basis of the existence of sufficient statistics. That is, the model establishes the parameter separability by conditioning the person parameters out of the calibration procedures entirely. Specifically, a test score of an examinee contains all the information for estimating a student's ability, and the item difficulties can be estimated from a simple count of persons completing each level or "step" (if PCM) of an item. The concept is explained mathematically in Eqs. (10) to (14) later. The model is used in this study whenever items are analyzed independently.

The Partial Credit Model (PCM)

The partial credit model (Wright and Masters, 1982) is an extension of the DM in that it provides a direct

expression of the probability of an examinee with ability β_n responding at a particular performance level (e.g., 1, 2, ..., m). For items with more than two performance levels (i.e., 0, 1), additional probability expressions are needed to describe the probability of getting score 2, rather than 1, score 3, rather than 2, and so on, in terms of item step difficulty parameters $\delta_{i2}, \delta_{i3}, \ldots, \delta_{im}$. The general form for the PCM to score k rather than k-1 is,

$$\phi_{nik} = \frac{\pi_{nik}}{\pi_{nik-1}+\pi_{nik}} = \frac{\exp(\beta_n-\delta_{ik})}{1+\exp(\beta_n-\delta_{ik})}, \quad k=1, \ldots, j, \ldots, m, \qquad (6)$$

and $\sum_{k=0}^{m} \pi_{nik} = 1$. In Eq. (6),

ϕ_{nik} is the probability of person n answering step k, rather than step k-1, of item i correctly,

π_{nik} is the probability of person n answering step k of item i correctly,

β_n is the person latent ability, and

δ_{ik} is the difficulty of the kth step in item i.

In the PCM, the probability of person n scoring x or completing any number of steps on item i is,

$$\pi_{nix} = \frac{\exp\sum_{j=0}^{x}(\beta_n - \delta_{ij})}{\sum_{k=0}^{mi}\exp\sum_{j=0}^{k}(\beta_n - \delta_{ij})}, \quad x=0, 1, 2, \ldots, m_i. \qquad (7)$$

In Eq. (7), δ_{ij} is the difficulty parameter for the jth step in the item. $\delta_{i0} \equiv 0$, so that $\sum_{j=0}^{0}(\beta_n - \delta_{ij})=0$, and $\exp\sum_{j=0}^{0}(\beta_n - \delta_{ij})=1$. Consequently, the probability of scoring 0 would be

$$\pi_{ni0} = \frac{1}{\sum_{k=0}^{mi}\exp\sum_{j=0}^{k}(\beta_n - \delta_{ij})}. \qquad (8)$$

The observation x in Eq. (7) is the count of the completed steps for item i. The numerator contains only the difficulties of these x completed steps $\delta_{i1}, \delta_{i2}, \ldots, \delta_{ix}$. The denominator is the sum of all m_i+1 possible numerators (Wright and Masters, 1982). In other words, the formula is the ratio of x-step difficulties over the total possible m-step difficulties.

Parameters to be estimated in this model are person ability (β_n), step difficulty (δ_{ij}), and testlet difficulty, which is the average of all possible step measures for that testlet.

The number of parameters equals N+M+L-1, in which N is the person parameter, M= $\sum_{i=1}^{L} m_i$, the total number of steps in all the testlets, and L is the number of testlets.

For a 4-testlet test with each testlet having 4 items (steps), the number of parameters equals N+4*4+4-1 = N+16+4-1=N+19.

Although the PCM requires that the steps within an item be completed in sequence, the steps need not be equally difficult nor be ordered by step difficulties. If an item has only two performance levels (i.e., 0, 1), then the PCM reduces to the DM.

In the present study, the items within a testlet become "steps" and each "step" (i.e., item) is scored 0 or 1. A testlet replaces the position of an item. The order of the items is the number of steps to be completed by an examinee.

Estimation Measures

The unconditional maximum likelihood estimation procedure was used here. The method involves a set of equations in which the item difficulty and latent trait score estimates are unknowns. Implementation of the procedure begins by calculating initial values for the difficulty and latent trait score estimates. These values are essentially guesses about the unconditional maximum likelihood estimates. The computer program BIGSTEPS uses these estimates in a procedure to produce a second set of difficulty and latent trait estimates. The second set is then used to produce the third set, and so on. This iterative procedure continues until further cycles through the procedure produces only minimal changes in the estimates. This final set comprises the unconditional maximum likelihood estimates. Since the calibration models used in this study were proposed by Wright and Masters (1982), all the estimation formulas and notations used here follow theirs.

Phi Coefficient

Phi correlation coefficient is used to describe the relationship between responses of two dichotomously scored items. Since the items within a testlet are equivalent to the steps of a multi-level item, subscript j is used whenever items are testlet "steps." Its formula is

$$\rho_{phi} = \frac{p_{jj'} - p_j p_{j'}}{\sqrt{p_j q_j p_{j'} q_{j'}}}, \text{ where} \qquad (9)$$

$p_{jj'}$ is the joint proportion of students answering both items correctly,

p_j is the proportion of students answering item j correctly,

$p_{j'}$ is the proportion of students answering item j' correctly,

q_j is the proportion of students answering item j incorrectly,

$q_{j'}$ is the proportion of students answering item j' incorrectly,

$p_j q_j$ is the variance for item j, and

$p_{j'} q_{j'}$ is the variance for item j'.

It is hypothesized that the context-dependent items may be correlated more closely within a testlet than correlations with items of other testlets. Therefore, phi correlation coefficients between pairs of items are calculated here to examine the hypothesis.

Person Ability Measure

In the unconditional maximum likelihood estimation procedure, the likelihood of the data matrix ((x_{ni})) is the continued product of the unconditional probabilities π_{nix} over n and i,

$$\Lambda = \prod_{n}^{N} \prod_{i}^{L} \pi_{nix} = \frac{\exp \sum_{n}^{N} \sum_{i}^{L} \sum_{j=0}^{x_{ni}} (\beta_n - \delta_{ij})}{\prod_{n}^{N} \prod_{i}^{L} [\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})]}. \qquad (10)$$

In Eq. (10),

π_{nix} is the probability of person n answering x steps in item i correctly,

x_{ni} is the observed score for person n on item i,

β_n is the person latent ability,

δ_{ij} is the step difficulty for item i, and

i=1, 2, ..., L, the number of items,

j=1, 2, ,.., m_i the item step, and

n=1, 2, ..., N, the person.

The logarithm of Eq. (10) is

$$\lambda = \log \Lambda = \sum_{n}^{N} \sum_{i}^{L} x_{ni} \beta_n - \sum_{n}^{N} \sum_{i}^{L} \sum_{j=1}^{x_{ni}} \delta_{ij} - \sum_{n}^{N} \sum_{i}^{L} \log[\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})], \qquad (11)$$

in which $\sum_{j=0}^{x_{ni}} \delta_{ij} = \sum_{j=1}^{x_{ni}} \delta_{ij}$ because $\delta_{i0} \equiv 0$. Taking the first derivative of Eq.(11) with respect of β_n, one gets

$$\frac{\partial \lambda}{\partial \beta_n} = r_n - \sum_{i}^{L} \sum_{k=1}^{m_i} k \pi_{nik}, \qquad \text{i=1,L} \qquad (12)$$

where $r_n = \sum_{i}^{L} x_{ni}$ is the test score for person n,

π_{nik} is the probability of person n completing k steps in testlet i,

k=1, 2, ..., m_i is the number of steps (i.e., items here) in testlet i,

$\sum_{k=1}^{m_i} k \pi_{nik}$ is the number of steps person n is expected to complete in testlet i, and

$\sum_{i}^{L}\sum_{k=1}^{m_i} k\pi_{nik}$ is the number of steps person n is expected to complete on the L-testlet test, or the expected score of r_n, the test score for person n. Symbolically,

$$E(r_n) = \sum_{i}^{L}\sum_{k=1}^{m_i} k\pi_{nik}. \qquad (13)$$

Setting Eq. (12) to 0, and solving for β_n, we will get an estimate of person ability, b_r.

The standard error of the estimate can be calculated by

$$SE(b_r) = [\sum_{i}^{L}(\sum_{k=1}^{m_i} k^2 P_{rik} - (\sum_{k=1}^{m_i} k P_{rik})^2)]^{-1/2}, \qquad (14)$$

where P_{rik} is the estimated probability of a person with a score of r responding in step k to testlet i of the last iteration.

The person fit statistic is

$$t_n = (v_n^{1/3} - 1)(3/q_n) + (q_n/3), \qquad (15)$$

where v_n is weighted mean square, q_n is the standard deviation of the weighted mean square, and t_n is the standardized weighted mean square for person n.

Testlet Measure

Taking the first derivative of Eq.(11) above with respect to δ_{ij}, one gets

$$\frac{\partial\lambda}{\partial\delta_{ij}} = -S_{ij} + \sum_{n}^{N}\sum_{k=j}^{m_i}\pi_{nik}, \qquad n=1,N;\ j=1, \ldots, k, \ldots, m_i, \qquad (16)$$

where $S_{ij} = \sum_{n=1}^{N}\sum_{j=1}^{x_{ni}}\delta_{ij}$ is the number of persons completing step j in testlet i. $\sum_{k=j}^{m_i}\pi_{nik}$ is the probability of person n completing at least j steps in testlet i, and

$\sum_{n}^{N}\sum_{k=j}^{m_i}\pi_{nik}$ is the number of persons expected to complete at

least j steps in testlet i. In other words, it is the expected value of S_{ij}. Symbolically, the expected value for step difficulty (d_{ij}) in testlet i is

$$E(d_{ij}) = \sum_{n}^{N} \sum_{k=j}^{m_i} \pi_{nik} . \quad (17)$$

Setting Eq. (16) to 0, and solving for δ_{ij}, we will get the estimate of testlet step parameter, d_{ij}.

The standard error of d_{ij} is

$$SE(d_{ij}) = [\sum_{r}^{M-1} N_r (\sum_{k=j}^{m_i} P_{rik} - (\sum_{k=j}^{m_i} P_{rik})^2)]^{-1/2} \quad (18)$$

where N_r is the number of persons with score r, $M = \sum_{i=1}^{L} m_i$.

The formula for testlet fit is

$$t_i = (v_i^{1/3} - 1)(3/q_i) + (q_i/3), \quad (19)$$

where v_i is the weighted mean square, q_i is the standard deviation of the weighted mean square, and t_i is the standardized weighted mean square for testlet i. Detailed derivation of Eq. (19) is done subsequently in the Local Dependent Item Measure section.

For the simplicity of this study, the testlets do not take response patterns into consideration, and students' raw scores on the items within a testlet are summed up to a single number-right score.

Local Item Dependence Measure

To assess the local dependence effect, dichotomously scored items are first calibrated with the Rasch dichotomous model as individual items and then by the partial credit model as testlets. The difficulties obtained from both

calibrations are compared for their estimated values, calibration errors, and item/testlet fits. The item fit statistics are calculated as follows (Wright & Masters, 1982):

observed response: x_{ni},

expected value of x_{ni}: $E_{ni} = \sum_{k=0}^{mi} k\pi_{nik}$, (20)

where $\pi_{nik} = \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij}) / \Psi_{ni}$, (21)

and $\Psi_{ni} = \sum_{k=0}^{mi} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})$, (22)

variance of x_{ni}: $W_{ni} = \sum_{k=0}^{mi} (k - E_{ni})^2 \pi_{nik}$, (23)

kurtosis of x_{ni}: $C_{ni} = \sum_{k=0}^{mi} (k - E_{ni})^4 \pi_{nik}$, (24)

score residual: $y_{ni} = x_{ni} - E_{ni}$, (25)

standardized residual: $z_{ni} = y_{ni} / W_{ni}^{1/2}$, (26)

standardized residual squared: z_{ni}^2, (27)

score residual squared: $y_{ni}^2 = W_{ni} z_{ni}^2$, (28)

unweighted mean square: $u_i = \sum_{n=1}^{N} z_{ni}^2 / N$, the outfit statistics, where N is the number of persons in the sample, (29)

weighted mean square: $v_i = \sum_{n}^{N} W_{ni} z_{ni}^2 / \sum_{n}^{N} W_{ni} = \sum_{n}^{N} y_{ni}^2 / \sum_{n}^{N} W_{ni}$, (30)

and finally,

standardized weighted mean square: $t_i = (v_i^{1/3} - 1)(3/q_i) + (q_i/3)$, the infit statistic, has a mean of 0 and variance 1. (31)

q_i is the SD of the weighted mean square, v_i. In the formula, it is

$$q_i = [\sum_{n}^{N} (C_{ni} - W_{ni}^2) / (\sum_{n}^{N} W_{ni})]^{1/2}. \quad (32)$$

Similarly, the person fit statistic can be obtained in this manner also.

The information-weighted fit statistic (v_i) obtained from the computer program BIGSTEPS would have an expected value of 1. Values substantially less than 1 indicate dependence in the data; values substantially greater than 1 indicate noise. More about the fit statistics will be discussed in Chapter 4.

Person Separation Ratio Indices

In classical testing theory, an observed variance is composed of two components. That is:
Observed variance (σ_x^2)=True variance(σ_T^2)+Error variance (σ_E^2), and the reliability is obtained by the following,

$$\text{Reliability } (\rho) = \frac{\sigma_T^2}{\sigma_x^2} \qquad (33)$$

One example of this kind of reliability is the coefficient α. A problem with classical reliability is that it depends on the population measured and on the measuring instrument. One has to specify the instrument and the population it applies to whenever he or she speaks of reliability because of population dependence.

In IRT Rasch models, "true" variance is the "adjusted" variance (i.e., observed variance adjusted for measurement error). Error variance is a mean-square error (derived from the model) inflated by misfit to the model encountered in the data (Wright, 1996). Because the intention of most tests is to identify individual differences, indices of separation of persons on the ability continuum have been developed to see how well a particular test separates the persons in a

particular sample. One such index is the person separation index (G_p), which is the number of statistically different performance strata that the test can identify in the sample. The index is the ratio of the adjusted SD (SA_p=(obs. SD^2_p - $MSE_p)^{1/2}$) to the root mean square error ($RMSE_p$). In the formula,

$$G_p = \frac{SA_p}{RMSE_p}, \tag{34}$$

where SA_p is the sample SD adjusted for measurement error, and $RMSE_p$ is the root mean square measurement error, p is the person, which equals 1, ..., N.

For example, a separation index of 3.5 means that if repeatedly tested, the ability estimates on the ability continuum can be consistently separated into roughly 3 strata by the test for samples like the one tested. In other words, G_p gives a sample standard deviation in standard error units. Person separation index provides an alternative way to examine the internal consistence of a test. Some consider it easier to interpret than the reliability coefficient.

When Eq. (34) is squared, it becomes the ratio of sample variance adjusted for measurement error to the mean of sample measurement error variance.

$$G_p^2 = \frac{SA_p^2}{MSE_p} \tag{35}$$

Eqs. (34) and (35) imply that the larger the person separation, the smaller the measurement error and the more precise an estimate is. The reliability of person separation

then is the ratio of the adjusted sample variance to the observed variance. Mathematically, it is

$$R_p = \frac{SA_p^2}{SD_p^2} = 1 - \frac{MSE_p}{SD_p^2}. \quad (36)$$

This reliability is analogous to KR-20, Cronbach's α, and the generalizability coefficient in the sense of classical testing theory. The relationship between the reliability of person separation and the classical reliability (ρ) is,

$$reliability(\rho) = \frac{G_p^2}{1+G_p^2}, \quad (37)$$

or, $$G_p = \sqrt{\frac{\rho}{1+\rho}}. \quad (38)$$

The indices are used here to examine the hypotheses 1(d) and 2(c).

Data Analysis

To test different hypotheses in this study, three things are done with the data. First, items within each original testlet are scored twice, once as independent items and once as a testlet. For all the science tryout forms, the testlet items are located in the same positions. They are:

Original Testlet 1: 11, 12, 13, 14;

Original Testlet 2: 28, 29, 30, 31;

Original Testlet 3: 45, 46, 47, 48;

Original Testlet 4: 50, 51, 52, 53.

Second, additional sixteen independent items in the same test form are randomly selected from the 30 independent MC

items to randomly form another four hypothetical testlets. The rationale for these random testlets is to see if there is a local dependence effect on the truly independent items when they are analyzed as a testlet. It is equivalent to running a concurrent validity study. One set of context-dependent items are analyzed at their original configuration, the other set of independent items from the same tryout form are analyzed at a hypothetical configuration, and results of these two sets are compared in terms of testlet statistics and person estimates to see whether there is a dependence effect in the original testlets. If there is no significant difference between the two sets of estimates in person and/or in item or testlets parameters, then one may infer that the null hypothesis of no local item dependence effect among context-dependent items within a testlet holds. These random testlets are first scored as individual items and then scored as testlets. The items composing the random testlets are truly randomly selected from the context-independent items in the same form. Since there are no items in common in any two forms, the same number of items are chosen for the simplicity of the analysis. The random testlets for Forms 20-29 are:

Random Testlet 1: 1, 8, 24, 38;

Random Testlet 2: 2, 9, 25, 40;

Random Testlet 3: 3, 18, 21, 41;

Random Testlet 4: 4, 20, 37, 43.

Third, the original testlets are broken up and reformed into 4 new testlets (similar to Latin Square design). The purpose was similar to the random testlets. That is, examining local dependence effects with items from different original testlets. The items in these reformed testlets are scored twice as in the original testlets. The reformed testlets for Forms 20-29 were:

Reformed Testlet 1: 11, 28, 45, 50;

Reformed Testlet 2: 12, 29, 46, 51;

Reformed Testlet 3: 13, 30, 47, 52;

Reformed Testlet 4: 14, 31, 48, 53.

According to the design, each kind of testlet configuration is analyzed twice. The first time the items are analyzed as individual items by the dichotomous model regardless of whether they are context-dependent or independent. The second time testlet scores are calculated for each testlet and then they are analyzed with the partial credit model. The configurations of different testlets and other forms of items are demonstrated below in Table 3.

By the unidimensionality property of the IRT testing theory, testlets are expected to correlate to each other as little as possible at a given level on the ability continuum. Therefore, it is assumed that when a testlet is used as the unit of analysis, the correlations between testlets at a given ability level should be small.

Table 3. Data Configurations of Science Items

	Data Configuration				
	Original Testlets (Testlets consist of context-dependent items as designed)	Random Testlets (Independent items from the same tryout form)	Reformed Testlets (Testlets consist of items from different original testlets)	Context-dependent Items (Items used to form the original/reformed testlets)	Independent Items (Items used to form the random testlets)
# of Testlets	4	4	4		
# of Items in a Testlet	4	4	4		
Total # of Items in the data set	16	16	16	16	16
Dichotomous Model Analysis				Yes	Yes
Partial Credit Model Analysis	Yes	Yes	Yes		

BIGSTEPS Computer Software

The computer program used in the parameter estimation and data analysis is BIGSTEPS (Linacre & Wright, 1995, version 2.6). The program is specifically designed to facilitate item analysis and scoring of psychological tests within the framework of IRT Rasch models. The program can analyze scores of both dichotomous and polytomous scales. Items may be grouped together or divided into subsets of one or more items that use the same scoring scale.

According to the program's user's guide, person measure and item calibration are reported in logits. "A logit (log-odds unit) is a unit of interval measurement which is well-defined within the context of a single homogeneous test" (Linacre & Wright, 1995, p.89). Mathematically, the logit $\lambda = \log\left[\frac{\pi}{1-\pi}\right]$ is the probability unit for λ defined by the

modeled process, where π is

$$P\{x;\beta,\delta\} = \pi = \frac{\exp(\beta-\delta)}{1+\exp(\beta-\delta)}. \quad (39)$$

This is the unit with which the Rasch measures can be compared as a uniformed standard unit.

Summary

Research data and the experimental methodology were described in this chapter. The first tryout data from the newly-developed *Michigan High School Proficiency Test in Science* was used. The test was designed to test students' abilities in *using*, *reflecting*, and *constructing* scientific knowledge. For each tryout form, only the context-dependent MC items within the testlets and an additional 16 randomly selected independent items were used in this study since testlet effect is the focus of the study. The constructed-response questions were not included in the research because they were hand-scored by different scoring rubrics, which may introduce interrater errors and over time errors for the same rater, and make the study too complex to be handled.

Cluster sampling in combination with stratified sampling was used in the tryout. Schools were used as the sampling unit. The sampling frame included all Michigan 11th graders in public schools. There were 10,074 students from 72 schools who actually took the tryout.

Ten tryout forms were spirally bundled into 6 groups, 3 or 4 forms in each group. Each tryout school received only

one group of forms. No items overlapped between forms but each form was administered to two randomly equivalent groups of 11th grade students.

All the MC items were scored dichotomously. Testlet were scored as number-right items within a testlet. The Rasch dichotomous model was used when items were analyzed independently and the Rasch partial credit model was used for the testlet analysis. All context-dependent items were first analyzed independently and then as testlets. Sixteen additional randomly-selected MC items were formed into 4 random testlets and were analyzed the same way as the original testlets. The original testlets were also reconfigured into 4 reformed testlets and were analyzed accordingly.

Different statistics to measure item correlation, test reliability, person and item/testlet fit statistics, and measurement/calibration errors were described for these analyses. It is expected that the results of the estimates would provide information about whether the local item dependence has any impact on the parameter estimation.

The computer program BIGSTEPS was used to run the analyses. The software was designed specifically for the data analysis of the Rasch models. The estimates are reported in logits. A logit is a unit of interval measurement that can make the comparison of measures on a uniformed standard unit.

CHAPTER 4

RESULTS AND DISCUSSIONS

As described in Chapter 3, to carry out the data analysis plan, data were organized and analyzed in five different ways. Four original testlets, 4 random testlets, 4 reformed testlets, 16 context-dependent items, and 16 independent MC items that formed the random testlets, were treated as though they were five different tests in each form. In essence, each form has 32 items (16 context-dependent items and 16 independent items) in total used in the analyses.

According to the plan, different statistics were computed for the data. Phi correlation coefficients (ϕ),testlet measures, person separation indices, and person ability measures, were all computed. In addition, a one-way ANOVA and average category measures were also calculated to provide an overall data description for each step in a testlet. Table 4 below summarizes these analyses and relates them to their hypotheses respectively. Results of these analyses are presented in the following sections. Discussions are often mixed with the results reporting in order not to lose the continuity. The chapter concludes with a summary.

Table 4. Match-up of the Analyses with Their Corresponding Hypotheses.

Analysis	Research Hypothesis								
	H1(a)	H1(b)	H1(c)	H1(d)	H2(a)	H2(b)	H2(c)	H3(a)	H3(b)
ϕ Coefficient	√							√	
Testlet Measure		√			√				
One-way ANOVA	Verification of fit statistics obtained from the partial credit model for local dependence.								
Person Ability Measure			√			√			√
Person Separation Indices				√			√		
Average Category Measure	Overall data description for each category (i.e., step) in a testlet.								

Phi Correlation Coefficient Results

The phi correlation coefficient (ϕ) is usually used to examine the linear relationship between two distinct dichotomously scored variables (e.g., male/female, smoking/non-smoking). The multiple-choice items in this study are dichotomous so that ϕ coefficient is appropriate. By Hypotheses 1(a) and 3(a) if the context-dependent items are generated from the same context, the average within-context items correlations should be larger than the average across-context item correlations. To test these hypotheses, ϕ coefficients were calculated for all the original, the random, and the reformed testlets for all tryout forms. The mean coefficients for each testlet for overall forms are listed in Table 5.

Table 5. Mean ϕ Coefficients for Items within Different Testlets by Form

Form	Type	Tlet.1 Items	Tlet.2 Items	Tlet.3 Items	Tlet.4 Items	Form mean
20	Original	.1203	.1169	.1731	.1180	.1321
	Random	.1761	.0790	.1298	.0725	.1144
	Reformed	.1799	.0681	.1258	.0644	.1096
21	Original	.1473	.0500	.3433	.1944	.1838
	Random	.0437	.0364	.0250	.0526	.0394
	Reformed	.1451	.1168	.1243	.0733	.1149
22	Original	.1838	.1292	.0747	.0934	.1203
	Random	.1287	.0300	.0492	.1484	.0891
	Reformed	.1378	.0673	.0500	.1890	.1110
23	Original	.1440	.1778	.0656	.2758	.1658
	Random	.0900	.0814	.1192	.0433	.0835
	Reformed	.1074	.0975	.1513	.0991	.1138
24	Original	.1126	.1376	.0620	.1714	.1209
	Random	.0866	.1175	.1170	.1474	.1171
	Reformed	.0431	.0835	.1247	.2429	.1236
25	Original	.1579	.1049	.1155	.0423	.1052
	Random	.1269	.0772	.1027	.0674	.0936
	Reformed	.0356	.1038	.0793	.0622	.0702
26	Original	.1421	.1243	.1824	.0980	.1367
	Random	.1455	.0972	.0941	.1209	.1144
	Reformed	.0758	.0422	.1516	.1809	.1126
27	Original	.1314	.0760	.2954	.1043	.1518
	Random	.0771	.0960	.1296	.0987	.1004
	Reformed	.0133	.0285	.1636	.2496	.1138
28	Original	.2306	.0318	.2487	.0970	.1520
	Random	.1510	.0585	.1215	.0730	.1010
	Reformed	.0366	.1230	.1044	.1275	.0979
29	Original	.2059	.1589	.0914	.1859	.1605
	Random	.1099	.0654	.0689	.1616	.1015
	Reformed	.1162	.1498	.1340	.0929	.1232
Mean By Testlet	Original	.1576	.1107	.1652	.1381	.1429
	Random	.1136	.0739	.0957	.0986	.0955
	Reformed	.0891	.0881	.1209	.1382	.1091

As is shown in the table out of the 40 original testlets, only one testlet (Testlet 3, Form 21) had an average ϕ coefficient above .30, which is relatively high for item correlation. Five testlets had mean coefficients between .20 and .30, more than half of the testlets (23) obtained moderate mean coefficients between .10 and .20, and the remaining 11 testlets had mean coefficients less than .10. For random testlets, twenty-three of them had mean ϕ coefficients less than .10, seventeen had mean coefficients between .10 and .20, but no testlets had mean coefficients greater than .20. For the reformed testlets, only Testlets 4 in Forms 24 and 27 had mean ϕ coefficients above .20 (ϕ=.2429 and .2496 respectively). Half of them (20) were between .10 and .20, and the remaining eighteen were under .10. As these data in Table 5 indicate, thirty-one of the original testlets and 27 of the reformed testlets had mean ϕ coefficients larger than those of the random testlets. The summary is in Table 6 below.

Table 6. Summary of Mean Item Correlations for the Testlets

ϕ Coef.	Original Testlets	Random Testlets	Reformed Testlets
> .30	1	0	0
.21 - .30	5	0	2
.11 - .20	23	17	20
.00 - .10	11	23	18
Total	40	40	40

Marginal mean coefficients for all forms by testlet (column mean) and for all testlets by form (row mean) were calculated also. For each marginal value, mean coefficients for the original testlets are higher than either random testlets or reformed testlets, except Form 24, where the reformed testlet mean is slightly, but not significantly, higher than the original testlet mean. Between the reformed and random testlet means, coefficient values vary irregularly. In some cases, random testlets have higher mean coefficients. Other times, vise versa. This outcome is not surprising, however, because the contents of the reformed testlets are not related to the same context any more, and they are almost equivalent to the random testlets in the sense of testlet construction. Overall, the results strongly suggest that context-dependent items do have higher correlations within-context than across-context or independent items do, which implies that local dependence may exist in some original testlets.

In summary, for the original testlets (ref. Hypothesis 1(a)) the results showed that, if the context-dependent items were generated from the same context, the average within-context item correlations were larger than the average across-context item correlation for a majority (29) of the original testlets. On the other hand, eleven reformed testlets (ref. Hypothesis 3(a)) had average within-context phi correlation larger than those of their corresponding original testlets. The remaining reformed testlets obtained

smaller average within-context correlations than their corresponding original testlets.

Testlet Measures Results

This section discusses the results for Hypotheses 1(b) and 2(a). Hypothesis 1(b) states that when context-dependent items are analyzed as a testlet by the Rasch partial credit model, the testlet calibration produces a better fit statistic than when these items are analyzed individually by the Rasch dichotomous model. Hypothesis 2(a) states that if the items are independent, then testlet fit statistics should be the same as the item fit statistics.

One rationale for using testlets as the unit of analysis is to determine whether the calibration errors are smaller when treating the context-dependent items in a testlet as a whole than when treating these items individually (i.e., ignoring the context effect), as well as determining whether such scaling produces better fits of testlet and/or person estimates.

The *User's Guide to BIGSTEPS* (Linacre & Wright, 1995) states that "INFIT is an information-weighted fit statistic, which is more sensitive to unexpected behavior affecting responses to items near the person's ability." And "MNSQ is the mean-square infit statistic with expectation 1. Values substantially below 1 indicate dependence in your data; values substantially above 1 indicate noise" (p. 82).

In the same manual, it is explained that, when values of infit mean square (MNSQ) statistic are, say, less than .8 or the standardized MNSQ is less than -2 SDs, it means there are redundant items and the test developers need to investigate the items to see if the test has similar items, one item answers another, or an item correlates with other variables, that is, there are local dependence effects. When the infit MNSQ is larger than, say, 1.2, or its standardized MNSQ is greater than +2 SDs, it may mean different things, such as biased items, qualitatively different items, or curriculum interaction. In these cases, one needs to investigate areas related to the problems (Linacre & Wright, 1995, p. 95).

By Eq. (30) the infit MNSQ is the sum of squares of the difference between the observed score and the expected score divided by the sum of variances on item *i* over N persons. In the formula,

$$v_i = \sum_n^N W_{ni} z_{ni}^2 / \sum_n^N W_{ni} = \sum_n^N y_{ni}^2 / \sum_n^N W_{ni} = \sum_n^N (x_{ni} - E_{ni})^2 / \sum_n^N W_{ni} ,$$

v_i is the weighted mean square,

i = 1, 2, ..., L, is the item,

n = 1, 2, ..., N, is the person,

$W_{ni} = \sum_{k=0}^{mi} (k - E_{ni})^2 \pi_{nik}$, is the variance for observed score x_{ni},

k = 1, 2, ..., m, is the item step, and

$y_{ni} = x_{ni} - E_{ni}$, is the residual,

$E_{ni} = \sum_{k=0}^{mi} k \pi_{nik}$ is the expected value of x_{ni},

$z_{ni} = y_{ni} / W_{ni}^{1/2}$, is the standardized residual, and

$\pi_{nik} = \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij}) / \Psi_{ni}$, is the expected probability of person n answering item i, kth step.

With the Rasch partial credit model, the smaller the discrepancy between the observed score and expected score, the larger the variance of x_{ni}. In the infit MNSQ formula, this means smaller residuals ($y_{ni} = x_{ni} - E_{ni}$). In other words, the formula will have a smaller numerator and a bigger denominator. As a result, v_i will be less than 1 when the numerator is smaller than the denominator.

Usually we expect an orderly pattern of responses. In other words, we want to see that the observed value is close to the expected value. However, when responses to an item are excessively orderly, that is, the observed scores are almost identical or identical to the expected scores, we may begin to suspect potential local dependence effects (Wright & Masters, 1982, p. 104). This would happen when problems like those mentioned earlier occur. An example of possible dependence is presented later in this section.

Table 7 (see Appendix E) displays the results of testlet fit statistics for the original testlets and item fit statistics for the context-dependent items that configure these testlets.

Insert Table 7 Here

In Table 7, seventeen out of 40 original testlets have 1 to 4 misfit items within a context when they are analyzed individually, but when they are analyzed as testlets, they produce a very good testlet fit. Considering Original Testlet 3 in Form 22 and Original Testlet 4 in Form 27 for example, when the items in those testlets are analyzed as individual items, all of the context-dependent items have misfit values beyond ±2 SDs (all 4 items have the "**" sign in col. 6). However, the items produce a proper testlet fit when they are analyzed as testlets (infit=1.03 for Testlet 3 in Form 22 and infit=.95 for Testlet 4 in Form 27). In addition, the standard errors of the estimates for the original testlets are uniformly .04, while the standard errors for the context-dependent items are larger, between .07 and .09 logit. These results mean that, for those context-dependent items, the testlet-based analyses are more appropriate statistically than the item-based analyses to examine students' abilities in the areas of interest.

For another 20 testlets, each also has 1 to 4 misfit context-dependent items when they were analyzed individually, but the testlet-based analysis still results in misfit calibrations (indicated by "**" sign in the table). Thirteen of these testlets have infit values substantially less than 1 (i.e., infit MNSQ < -2 SDs), implying that there may be local dependence effects in both the items of those testlets or the testlets themselves. This finding is a little surprising because these testlets are supposed to be independent to each

other by design or by model control. It seems that there are some other factors other than local dependence affecting the item and testlet calibration. Another 7 testlets have infit values substantially greater than 1 (i.e., infit MNSQ > +2 SDs). For instance, Original Testlet 3 in Form 23 has misfit values for all its context-dependent items and the resulting infit MNSQ (1.22) for the testlet shows noise in the data this time. This means students may have unexpected performance away from their expected scores. This outcome suggests that test developers need to look at the testlet construction, content or quality of the items.

By the definition of fit statistics, Testlet 3 in Form 23 demonstrates one extreme (i.e., v_i greater than 1). The testlet is an *earth science* problem which requires students to know the relationships between the ocean, coastal plateau, and mountain range. It is a relatively difficult testlet (difficulty measure=.98 logit). If a student were not clear about their relationships, the person would have a small probability of answering an item correctly. The items themselves are well written, with no signs of bias or trick, but for the two more difficult items (item #46's b=1.39 and item #48's b=1.19 logits), the percentages of students choosing a wrong option are larger than the percents of students choosing the right one (see Table 8 for detailed percentages). For item #46, the correct answer is option A. The percentage of students choosing A was 28% only, compared with 35% who chose the wrong option, D. The situation is

similar for item #48. The percentage of students choosing the right answer, C, was 31%, while the percent choosing the wrong answer, D, was 35%. In addition, the average correlation among all 4 items is very small (r=.0656).

Table 8. Students Responses to Testlet 3, Form 23.

Item #	Option A	Option B	Option C	Option D
45	9.6%	13.6%	55.0%√	18.6%
46	27.9%√	22.5%	11.3%	35.0%
47	10.4%	15.5%	34.2%	36.6%√
48	9.8%	20.8%	30.8%√	35.4%

√ means the correct answer.

The results of large infit MNSQs (values substantially above 1.0) indicate large discrepancies between the observed scores and expected scores, implying students did not perform at their ability levels. These large discrepancies are considered "noise" in the item analysis. Usually one would suspect the item quality in this kind of situation. In this case, however, one may have to examine if there is an interaction of science dimensions within the testlet to seek possible reasons for poor performance. Nevertheless, "noise" in the item analysis does not have any relationship to local dependence. It is presented here to demonstrate another side of the infit statistic (i.e., values greater than 1.0). It also shows that large discrepancies between observed scores

and expected scores do happen even though items are from the same context.

Testlet 4 in Form 23 provides an example of possible dependence. The testlet presented a diagram of the movement of carbon in the atmosphere and on the surface of Earth, and asked students to answer 4 questions based on the diagram. It was a relatively easy testlet (difficulty measure=-.82 logit) and most students chose the right answers of the items (see Table 9 for detail percentages). Looking at the item statistics, it seems that distractors for three of the four items were not very effective because they attracted few students. By examining the item contents closely, we can see that if a student can answer item #52 (a concept item) correctly, he or she can answer the items #50, #51 and #53 fairly easily. Consequently, the observed and expected score differences will be very small.

Table 9. Students Responses to Testlet 4, Form 23.

Item #	Option A	Option B	Option C	Option D
50	6.5%	79.5%√	7.0%	3.1%
51	11.1%	11.0%	28.0%	47.3%√
52	10.8%	66.0%√	8.6%	11.1%
53	7.3%	79.9%√	4.5%	4.7%

√ means the correct answer.

As described in this section, small residuals imply possible local dependence. The average item correlation of

this testlet (r=.2750) helps support the suspicion. This correlation is very high in this test, compared with the grand average correlation (r=.1429). When a situation like this is true, the infit statistic, v_i, will be very small (because the residual, y_{ni}, will be very small). For this testlet in particular, the infit MNSQ is .76, which indicates that possible local dependence may exist among the items.

In summary, the statistics in Table 7 show that, for 17 of the 40 original testlets, some of the context-dependent items were problematic when they were analyzed individually but produced good fit when they were analyzed as testlets. This provides strong evidence that the partial credit model is more appropriate for these items. However, for another 20 original testlets, each also had 1 to 4 misfit items when they were analyzed individually, but the final testlet fit statistics were still misfit. Thirteen of these 20 testlets indicate possible local dependence, which suggests further investigation of individual items in these testlets regarding their contents, item construction, or item quality.

Across the forms, there are only 2 original testlets (Testlet 1 in Form 21 and Testlet 2 in Form 23) where the fit statistics are within the normal range regardless of which scoring model is used. Therefore, it would not matter if items in these testlets are analyzed independently or as testlets.

The strangest case is Testlet 3 in Form 26. All its 4 items are perfectly fit when analyzed individually, but the

testlet fit is not acceptable (infit MNSQ=.88, less than -2 SDs). The reason of this outcome is unknown to the author. The only inference that can be made is that these items many be truly independent and should be analyzed independently, even though they are from the same context.

An analysis was also run for the random testlets and the independent items that form the random testlets (see Table 10 in Appendix E). The results are similar to those of the original testlets.

Out of 40 random testlets, 15 of them had from 1 to 4 misfit items when these items were analyzed as individual items, but they obtained very proper fit when they were analyzed as testlets. Another 54 items that were distributed in 24 random testlets obtained misfit results no matter which model was used. Out of these 24 misfit testlets, 16 show local dependence and 7 indicate noise in their data. Again 2 random testlets (Testlet 4 in Form 21 and Testlet 3 in Form 29) obtained misfit when they were analyzed as testlets but had very good fit for each item when they were analyzed as independent items. In addition, there is no random testlet that shows proper fit for both scoring models, which ideally should be the case for these developer-designed independent items.

The outcome of misfit items converting into proper fit testlets that are related to no specific contexts is interesting, at the same time a little bit disturbing too. Theoretically, the developer-designed independent items

should behave as statistically independent. However, the results of these 15 misfit-items-to-fit-testlets here show that they are actually better off when they are analyzed as testlets. One needs to see if there is local dependence effects in these items or the results are just from random errors. The results for the random testlet analyses indicate that these labeled "independent" items may not be really statistically independent, even though they were designed to be so. Some items may be related to each other or to a common factor statistically, and more study is needed on these items.

One difference between the random testlets and the original testlets in fit statistic analyses is that the range of the independent item standard errors (.07-.14) is larger than those of the context-dependent items in the original testlets (.07-.09). This suggests student performance varies more for these independent items than for those context-dependent items, which further suggests that the context may have impact on student ability estimation as well as testlet calibration.

Insert Table 10 Here

Regarding the hypotheses tested in this section, it may be concluded that for the context-dependent items (ref. H1(b)), mixed results have been obtained. More than 40% (17) of the original testlets demonstrate a better fit when they

were analyzed as testlets. Half (20) of the original testlets have misfit by both models. Only 5% (2) of them obtain good fit as individual items and as testlets. For the independent items (ref. H2(a)), the testlet fit statistics are not the same as the items fit statistics. Sixty items in 15 random testlets have obtained a better fit when they were analyzed as (hypothetical) testlets. Another 34% of the items (54) show misfit with these items being analyzed as testlets and as items. The results are contradictary to the test development in that these items do not contain local independence with them. It is suspected that there may be an implicit factor affecting item calibration. For Hypothesis 3(b), the person fit statistics estimated by the reformed testlets are not significantly different from the person fit statistics estimated by the original testlets.

Verification of Local Dependence Effects

One way to verify whether the context-dependent items demonstrate dependence to each other when they are analyzed individually is to first check the variance homoscedasticity of the item fit statistics and then conduct a one-way ANOVA to compare the means of the fit statistics regressed on testlets.

The fit statistic discussed in the last section is a weighted mean square with degrees of freedom by the number of students responding to an item minus 1. In this study, the degrees of freedom are relatively large for all forms since

the test is large-scale. Consequently, the null hypothesis of local item independence within an original testlet would be easily rejected even though the dependence effect is very small. An alternative is to conduct a one-way ANOVA to verify whether the item fit statistics obtained by the Rasch partial credit model truly indicate local dependence between context-dependent items within a testlet.

In this ANOVA, the natural log of the infit statistic is the outcome variable and the testlet is the classification variable. If the confidence interval (CI) of its estimate includes 0 (because the expected value of infit is 1, so ln(E(infit) should be 0), it can be inferred that there is not enough evidence to show that items within a testlet are dependent.

Under normality and random sampling assumptions, the test statistic for a population variance equal to a pre-determined value is

$$\frac{vs^2}{\sigma^2} \sim \chi^2_v, \tag{40}$$

where v, equal to n-1, is the degree of freedom of the chi-square distribution, n is the number of examinees responding to the item, and s^2 is some mean square, equal to $\frac{ss}{v}$, ss is sum of squares. (In this study, s^2 is the weighted mean square of a context-dependent item.) Thus,

$$E(s^2) = \sigma^2, \text{ and} \tag{41}$$

$$\text{var}(s^2) = \frac{2\sigma^4}{\nu}. \quad (42)$$

Further, if we take the natural log of s^2, we get

$$E[\ln(s^2)] \approx \ln(\sigma^2), \text{ and} \quad (43)$$

$$\text{var}[\ln(s^2)] \approx \frac{2}{\nu}. \quad (44)$$

Consequently, because the term σ^2 is "logged out," if the degrees of freedom (*df*) for all the context-dependent items are the same, then the comparison between the infit statistics will not be biased. Otherwise, some adjustment may be needed. Table 11 in Appendix E lists *df*'s for all context-dependent items.

Insert Table 11 here

Values in Table 11 show that the majority of discrepancies between *df*'s from the highest to the lowest within a testlet are between 1 to 4 out of about 1,000 students. Two testlets (Testlet 4 in Forms 21 and 28) have somewhat larger differences in *df*'s, 9 for Form 21 and 14 for Form 28, respectively. Table 12 (see Appendix E) lists all the discrepancies in *df*'s.

Insert Table 12 here

We may assume that the small differences in *df* within a testlet are negligible because the infit statistic is a

weighted mean square (i.e., variance is considered) and the sample size is large (1000 or so). A one-way ANOVA has been conducted then for each form. The results are shown in Table 13 (see Appendix E). The graph of confidence intervals (CI) is displayed in Figure 5 (see Appendix E).

Insert Table 13, Figures 5-6 here

As stated earlier, the expected value for the infit statistic is 1 and its natural log is 0. It can be seen from Figure 5 that 35 out of 40 testlet statistics have included 0 in their CIs across the forms. Two testlets (Testlet 3 in Form 23 and Testlet 4 in Form 25) have values above 0 (indicating noise) and three testlets (Testlet 4 in Form 23, Testlet 3 in Form 27, and Testlet 1 in Form 28) have values below 0 point (indicating local dependence). The omnibus F statistics in Table 13 helps support the evidence. For all ten forms, 7 of them have nonsignificant F tests, indicating all testlets may include 0 and their infit statistics are within the normal range. Forms 21, 23, and 28 have significant F tests, implying that some of their testlets may have misfit statistics. The large SDs for some testlets in the table also show that these testlets would have a wide confident interval. Figure 5 explains the outcome graphically.

Figure 6 shows the point estimates of ln(infit MNSQ) for all testlets. The majority (31) of estimates fall between

-.05 and +.05, very close to 0, which provides the evidence to support that the testlet-based analysis produces appropriate fit statistics for the majority (30) of the original testlets in this test when a CI is built for each testlet.

Mean Person Ability Measures Results

It is acknowledged that the real purpose of any data analysis method in education is to try to measure person abilities as precisely as possible. Chapter 3 Hypothesis 1(c) stated that when the context-dependent items are analyzed as the original testlets, the person measure will have a better fit than when these items are analyzed individually. Hypothesis 2(b) proposed that since the independent items are not linked to a particular context, the person fit statistic will stay the same regardless of whether the items are analyzed individually or as testlets. For Hypothesis 3(b), because the reformed testlets are not context specific, it is hypothesized that the person fit statistics will not be as good as those of the original testlets.

Table 14 in Appendix E presents results for mean person ability measures for different data configurations. In the table, the first column is the data configuration. The second column is the mean of the estimated person measures for the examinees in different data configurations in each tryout form. The estimates are in logits. For most forms, the original testlets have slightly lower mean person measures

than the context-dependent items do, except Form 26. In addition, their values vary between -.50 and .50 logit values, right around the middle point of 0 on the ability continuum. Only the independent-item data configuration for Forms 24 and 26 and the random testlets in Forms 24, 26 and 27 have mean measures greater than .50 logit value. Most of the time, these measures do not differ much for most forms no matter how the context-dependent items are analyzed: individually or as testlets.

Column 3 is infit mean-square (MNSQ) for the mean person measure. It is the average of the infit mean-squares associated with responses of the sample and it has an expected value of 1.0. Values in Column 3 show that regardless of types of data configuration, no infit MNSQ statistic has a value substantially below 1.0. The lowest value is .92, and the highest is 1.0, which indicates that in average there is not enough evidence to prove unexpected behavior affecting responses to items or testlets near students ability levels.

Outfit in Column 4 is an outlier-sensitive fit statistic. Its MNSQ is the mean-square outfit statistic with an expectation of 1.0. As with the infit statistic, values substantially less than 1.0 indicate dependency, while values substantially greater than 1.0 indicate the presence of unexpected outliers. In this sample, the outfit MNSQ statistics ranges from .94 to 1.10, which indicates that the data fit the model relatively well.

Insert Table 14 Here

One thing that has to be explained here is the phrase "data fit the model." Usually in statistical analyses, researchers test whether a model fits data because the model is designed to imitate data, so it has to be faithful to the data as much as possible. Otherwise, another model is used.

The Rasch model used here, however, is not designed to fit any data. Instead it is developed to define measurement. As Wright (1992) pointed out: "The Rasch model is a statement, a specification of the requirements of measurement -- the kind of statement that appears in Edward Thorndike's work, in Thurstone's work, in Guttman's work (p. 197)." Therefore, ".... The Rasch model is theory centered: data must fit, else get better data (p. 200)." As a result, the phrase "data fit the model" is used in this study.

In summary, regarding the hypotheses discussed in this section, the conclusions will be the following. For the context-dependent items, there is no significant difference in person fit when the items were analyzed individually or as testlets (ref. H1(c)). For the independent items, the person fit statistics stay the same regardless of which model is used (ref. H2(b)). For the reformed testlets, even though the testlets are not context-specific, they nevertheless still

produce proper person fit as do those of the original testlets (ref. (H3(b)).

Person Separation Indices Results

It is hypothesized (Hypotheses 1(d) and 2(c)) in this study that, when items are context-dependent, they will produce smaller measurement errors when they are analyzed as testlets than when they are analyzed as individual items. Otherwise, if items are independent, it does not matter which scoring model is used. In this section person separation indices are examined to test the above hypotheses. In addition, the person separation ratio index will also provide an alternative for examining the reliabilities of different data configurations.

In Table 15 RMSE is the root mean square standard error computed over the persons or over the items. The computer program BIGSTEPS computes two kinds of RMSE: model RMSE and real RMSE. Model RMSE is computed on the assumption that the data fit the model, and that all misfit in the data is merely a reflection of the stochastic nature of the model. Real RMSE (col. 3) is computed over the persons or items on the basis that misfit in the data is due to departures in the data from model specifications (Linacre & Wright, 1995). Columns 4 (adjusted standard deviation) and 5 (separation ratio) are described earlier in Chapter 3. By Eq. (34), Column 5 is equal to Column 4 divided by Column 3.

Values in Table 15 show that, regardless of item configurations, all but 3 person separation ratios range from 1.00 to 1.60 logits. Recall that testlets are much larger units than the items are and, more importantly, they have taken any local dependence effect into account. When a test consisting of larger units such as testlets here obtains similar separation ratios as a test consisting of smaller units such as single items, one can infer that the testlet-based analysis produces better fit statistics for person estimation than the item-based analysis does because the former has relatively smaller measurement errors.

Insert Table 15 Here

Table 16 lists the reliabilities of person separation for different data configurations. It can be seen that for all tryout forms, the mean reliabilities of person separation ratio for the original testlets was .62, while results of the other types were .66 for the random testlets, .68 for the reformed testlets, .63 for the context-dependent items, and .60 for the independent items. The reliabilities of person separation for the original testlets was very competitive to those of the context-dependent items, considering that the later ignores the within-testlet structure and their real reliabilities may be a proportion to the values appearing in the table here. The results imply that for the items in these forms, using the original testlet configuration would have at

least a good, if not better, reliability estimate as analyzing the context-dependent items individually.

Therefore, for Hypothesis 1(d), it can be inferred that when items are context-dependent, the person separation ratios are not statistically different as to whether items are analyzed as testlets or as individual items. When the items are independent (ref. H2(c)), the relability of person separation ratio is the same for both the testlet-based analysis and the item-based analysis. Overall, the testlet-based analysis indicates an implicitly higher test reliability than the item-based analysis does because the former takes local item dependence effects into account when they are present in the data.

Table 16. Reliabilities of Person Separation for Different Data Configurations

Form	Original Testlets	Random Testlets	Reformed Testlets	Context-Dep. Items	Indep. Items
20	.60	.70	.65	.62	.62
21	.62	.53	.72	.67	.43
22	.65	.67	.68	.64	.61
23	.61	.68	.69	.63	.59
24	.65	.72	.66	.63	.66
25	.53	.63	.60	.53	.57
26	63	.66	.69	.65	.63
27	.62	.69	.71	.63	.61
28	.59	.67	.67	.61	.62
29	.65	.66	.70	.67	.61
Mean	.62	.66	.68	.63	.60

Average Category Measure Results

In partial credit models, when observations are ordinal, it is implicitly assumed that the higher the category level,

the greater the latent ability demonstrated. Consequently, the "more able" students would perform better in average and achieve higher scores than "less able" students. Average category measures presented in this section do not aim at a particular hypothesis, but rather provide some descriptive statistics for the sample under study rather than inferential information. The average category measure estimates the average ability for all students who reach a particular category of a testlet. The purpose of this index is to investigate whether each category is properly scored as it is intended. It is expected that the average category measure increases along the variable in the correct rank order. The higher the category number, the more latent ability is evidenced. In this study, the total number of categories in a testlet is the maximum number of score points of a testlet, including 0. For example, a score of 3 points means a student is in category 3 of this testlet.

Table 17 in Appendix E presents the results of average category measures (also called average measure for simplicity) for the original testlets, the random testlets, and their infit statistics for each category respectively. Values of average measures from Table 17 show that student average abilities of reaching different score categories for the original testlets are similar to those in the random testlets for all 10 tryout forms, most of them ranging between ±2.0 logits. The next column of the same table contains its infit MNSQ, the ratio of the observed residual

sum of squares due to ratings of a specific score (e.g., $X_{ni}=x$) over the expected residual sum of squares.

When the data fit the model, the modeled variance approximates the residual sum of squares. Differences are diagnostic of misfit. This infit MNSQ summarizes the agreement of responses for each category. It has an expectation of 1.0 and can range from 0 to ∞. Values substantially greater than 1.0 indicate improbable category use (e.g., some students obtain scores that do not match their abilities). Values substantially less than 1.0 indicate overly predicable category use (e.g., students choose the same options for all items).

Insert Table 17 Here

In Table 17, some testlet categories have infit MNSQ substantially larger than 1.0, implying abnormal observations for some students' performance. For example, in Form 23, Category 4 of Original Testlet 3 has an infit MNSQ of 1.76, which means some students who score 4 points for the testlet perform unexpectedly well. On the other hand, Category 3 of Original Testlet 4 in the same form shows an overwhelmingly low infit value (.67). This suggests that some students may make obvious choices (e.g., choosing eye-catching options as correct answers) or select the same options for all items in the testlet rather than using their higher-order thinking skills. Another finding in this table is that there is no

pattern within or between the original testlets and the random testlets regarding when over prediction or improbable observations would occur. For instance, in Form 22, Random Testlet 3 shows high infit values (e.g., 1.24 to 1.97) for 3 of its 5 categories, while in Random Testlet 4 of the same form, the category values are substantially low (.72 to .85). The same thing happens in the original testlets. In Form 23, Original Testlet 4, except for Category 0, where the infit measure is normal (.96), other categories manifest substantial low infit MNSQs (.67 to .76). Original Testlet 3 in Form 29 has the opposite situation, where the infit values range from 1.13 to 1.31 for its categories, suggesting some students who should have reached one category actually went to another category or vise versa.

Results of the average measures are also presented in terms of the range of categories. In Table 18 (in Appendix E), ranges of the random testlets are almost uniformly larger than those of the original testlets. The few exceptions are Testlet 2 in Form 21 and Testlet 3 in Form 22, where the ranges of the original testlets are slightly larger than that of the random testlets. One possible explanation for the narrower range of the original testlets may be that, although items within an original testlet are not closely correlated to each other, they are not as difficult when tested together as a whole unit as tnat of the random testlets, where items are tested in different places of the test.

Insert Table 18 Here

Summary

Different analyses were conducted to examine the differences between the testlet-based scale and the item-based scale. It was found that the context-dependent items overall correlate more closely within an original testlet than with items outside that testlet. There is obvious evidence that local item dependence may exist in some of the original testlets.

A good proportion (40%) of the context-dependent items demonstrate better fit for testlet calibration when they are analyzed as testlets. This suggests that these items have misfit either in local dependence or noise if analyzed individually. The Rasch partial credit model is the better model to control these errors for these items. However, another 50% of the original testlets (20) cannot reach proper fit by either model, which leads to the suspicion that there may be some other implicit factors such as interactions of science dimensions between those testlets that affects testlet calibration.

Analyses on the supposedly independent items found that a considerable number of items (60) have a better fit when they are analyzed as testlets, even though there is no specific context developed for the testlets. An additional 54

items (in 24 random testlets) would obtain misfit no matter which model is used. The results demand further study on these developer-designed independent items.

Across the forms, there are only 8 context-dependent items in 2 original testlets (items Testlet 1 in Form 21 and Testlet 2 in Form 23) where the fit statistics are within the normal range regardless of which scoring model is used.

A one-way ANOVA was conducted to verify the existence of local dependence effects within an original testlet and a CI was built for each testlet infit MNSQ. The results provide evidence to support that the testlet-based analysis produces appropriate fit statistics for 75% of the original testlets (30) in this study.

Mean person measures for all five data configurations were compared. For the context-dependent items, there was no significant difference in person fit when the items were analyzed individually or as testlets. For the independent items, the person fit statistics stayed the same regardless of which model was used. For the reformed testlets, even though the testlets were not context-specific, they nevertheless still produced as proper person fit statistics as did those of the original testlets.

Person separation index and the reliability of person separation were described and calculated for all the original testlets, the reformed testlets, and the context-dependent items to see how well a particular data configuration can differentiate the persons in a particular sample. It was

found that when items are context-dependent, the person separation ratios are not statistically different as to whether items are analyzed as testlets or as individual items. When the items are independent, not much difference is presented as to which model is better than the other either. Overall, the results indicate that employing the testlet-based analysis could obtain a test reliability that more truly reflects its nature than the item-based analysis does because the former takes local item dependence effects into account when they are present in the data.

Average category measures provided estimates of the average abilities of the examinees reaching a certain score level of a testlet. It was intended to check for any improbable category use or over prediction. The average category measures for each original and random testlet were computed and compared. It was found that the two kinds of testlets performed similarly for all tryout forms, and there was no pattern as to which type of testlets would more likely have improbable observations or over predictions. However, the ranges of the categories within an original testlet were not as wide as those of the random testlets.

CHAPTER 5

SUMMARIES AND CONCLUSIONS

There are six sections in this last chapter of the study. First, a very brief summary of the study is presented. Then a summary of the results by hypothesis follows. Third, conclusions are made based on the results of the study. Fourth, limitations of the study are discussed. Fifth, generalizability of the study is pursued. In the final section, a few recommendations for further research are proposed.

Summary of the Study

The issue of local item dependence has received increasing attention in the past decade due to progress in the area of IRT item analysis, and more importantly, the increasingly high-stake assessments administered at the different levels of education.

Literature indicates that the testlet concepts have been widely applied in regular classroom testing, computerized adaptive testing, and non-traditional, non-IRT scoring. It has been found that although item-based parameter estimation for the context-dependent items appear to provide more information over most levels of the latent trait continuum,

this extra gain in information may be "fooled" by the excess within-context correlation among the items. This situation is especially true when the assumption of local independence is violated. It has been suggested that one should use testlets to manage the local item dependence problem.

The purpose of this study was to explore the local item dependence effect when context-dependent items in the *Michigan High School Proficiency Test in Science* were analyzed as independent items and as testlets. The family of the Rasch models (partial credit and dichotomous models) were applied to testlets in a large-scale assessment program, and to estimate the person ability measures and the test reliabilities, testlet/item calibrations, and testlet/item fit statistics when the potential violation of the assumption of local independence is controlled by the testlet.

The first tryout data from the newly-developed *Michigan High School Proficiency Test in Science* (1995) were used. The test was designed to examine students' abilities in *using, reflecting,* and *constructing scientific knowledge. Using science* was further divided into *using life, using physical* and *using earth. Reflecting* and *constructing* were embedded across all three content areas. There were ten forms in total for the tryout. Every form had four testlets, each testlet consisted of four multiple-choice items and one or two constructed-response questions. Only multiple-choice questions were used in the study to avoid the inter-rater reliability problem and other related issues in the hand-

scoring of constructed-response questions. In addition, only context-dependent items and an additional 16 independent multiple-choice items in the same form were used in the analysis.

Cluster sampling in combination with stratified sampling was used in the tryout to ensure that the sample was representative of the population. The sampling frame included all Michigan 11th grade students, including alternative education and special education students. There were 10,074 students from 72 schools who took the science tryout test. All ten forms in the tryout were used in this study.

Data were analyzed in five different configurations: as the individual context-dependent items, the original testlets, the reformed testlets, the individual independent items, and the random testlets. Statistical methods of phi coefficient, testlet measure, one-way ANOVA, person ability measure, person separation indices, and average category measure, were used in the analysis.

Summary of the Results by Hypothesis

Mixed results have been generated from the data analyses in this study. They are presented in the order of the research hypotheses.

For context-dependent items:

1a. If the context-dependent items were generated from the same context, the average within-context item correlations were larger than the average across-context item correlation

for a majority (29) of the original testlets.

1b. More than 40% (17) of the original testlets demonstrated a better fit when they were analyzed as testlets. Half (20) of the original testlets had misfit by both models. Only 5% (2) of them obtained good fit as individual items and as testlets.

1c. No matter how the data were organized, whether they were analyzed as individual items or as testlets, the person fit statistics generated from the Rasch dichotomous model were as good as those from the Rasch partial credit model.

1d. The person separation ratios were not statistically different whether items were analyzed as testlets or as individual items. However, the nonsignificantly different person separation ratios between the testlet-based analysis and the item-based analysis indicate that the former had smaller measurement errors than the latter because the former has a larger unit of analysis and it took the local item dependence into account.

For independent MC items:

2a. When the items were analyzed as a testlet by the Rasch partial credit model, the testlet fit statistics were not the same as the items fit statistics when the items were analyzed individually by the Rasch dichotomous model. Sixty items in 15 random testlets obtained a better fit when they were analyzed as (hypothetical) testlets. Another 34% of the items (54) showed misfit both when these items being analyzed as testlets and as items. The results are contradictary to the

intention of the test development in that these items should be context independent. It is suspected that there may be an implicit factor affecting item calibration.

2b. The person fit statistics for the independent items configuration and the random testlets configuration were not significantly different.

2c. The reliability of person separation ratio was the same for both the testlet-based analysis and the item-based analysis.

For the reformed testlets:

3a. When context-dependent items in the original testlets were reconfigured into the same number of new testlets, each with an item from each original testlet (i.e., reformed testlets), their mean correlations were not all smaller than those of the original testlets. Eleven of them had mean within-context phi correlations larger than those of their corresponding original testlets. The remaining reformed testlets obtained smaller average within-context correlations than their corresponding original testlets.

3b. The person fit statistics estimated by the reformed testlets were not significantly different from the person fit statistics estimated by the original testlets.

Conclusions

Based on the results of this study, the following eight conclusions are made.

1. Context-dependent items correlated more closely within-context than across-context for most original testlets in this study, which provides some evidence that local item dependence does exist within a context.

2. Where there is a local item dependence effect in the context-dependent items, the IRT assumption of local independence may be violated for some context-dependent items. Under this circumstance, it would be thereotically preferrable to use the Rasch partial credit model. Evidence in this study showed that such a local dependence effect can be controlled and a better fit for item calibration can be obtained by employing the model for some, but not all original testlets.

3. Caution must be exercised in any revision of the misfit testlets. Often only one or two misfit items causes misfit of the whole testlet. When the problematic item(s) are not highly correlated to other items in the context, the test developers only have to eliminate or revise the bad item(s) instead of discarding the whole testlet.

This conclusion may be more meaningful to test developers than to curriculum specialists or teachers. Very often during the testlet development an item is found to be problematic in measurement or for other concerns such as ethnic or gender bias. As a result, the whole testlet is discarded because of the underlying assumption that a testlet is considered as a complete piece and all of its parts are clustered together closely and should not be separated. If

one part goes wrong, the whole work is terminated. The results from this study imply that when context-dependent items are not highly correlated with each other, deleting the problematic item may not affect the remaining part of the testlet significantly. Therefore, one can still keep the technically sound items, and revise or eliminate the bad item, or, replace it with a new item. It is not necessary to discard the whole testlet or make any changes in other testlets either.

4. It seems that an implicit factor other than the local item dependence affects the misfit original testlets. Even when the Rasch partial credit model was applied unacceptable fit statistics were obtained.

5. Local item dependence effects may even exist in some developer-designed independent items in this study. However, they may be caused by random errors.

6. Truly statistically independent items should be analyzed independently, whether they belong to a context or not.

7. There is no significant different between the Rasch partial credit model and the Rasch dichotomous model in average person ability measures. Competitive estimates were obtained by both models.

8. The Rasch partial credit model, which was usually used to analyze partial credit items, performed efficiently in analyzing the testlet data of this large-scale assessment. The computer program BIGSTEPS provided most of the necessary information for this research in a user-friendly manner.

Limitations

Every study has its limitations. The major limitation of this study may be the quality of the data. Since the data were from a tryout administration, there were no previous item statistics available. Therefore, there was no reference of item quality, testlet formation or other related information.

Another limitation is the nature of the testlet formation. Because the original testlets here were designed to assess students' multiple traits, their items were not linked to a common factor. Therefore, it is unlikely that student abilities would be affected by a single context. If these testlets had been developed as unidimensional instead of multi-dimensional, the results may have been quite different.

In addition, because it was a tryout and not an operational administration, the results did not have any impact on student records, and therefore, it did not matter if they performed seriously or not. Consequently, student attitudes may confound the results of the study.

Furthermore, for the simplicity of the study, neither the response patterns of the testlets nor the constructed-response questions were considered in the research design. Whether this would affect the results is not known.

Generalizability

One of the outstanding features of this study is that the data were collected from a very large and representative sample of approximately 100,000 students per testing instrument. Because every 11th grade student in Michigan public schools is required by the Legislature to take the Michigan High School Proficiency Tests, it was possible to sample from the entire public school student population of the 11th grade, which can help generalize the results to similar situations. However, such a large-scale and high-stake assessment may not be available in every field. So the methods described in the study may not be applicable to every testing situation. Other researchers who want to do similar studies or generalize the results from this study need to be very cautious on this matter.

Another important and practical factor is the cost of data analysis. even though some evidence of local dependence has been shown here, it is almost impossible to score those items as testlets with the Rasch partial credit model and other items with the dichotomous model for such a large-scale statewide assessment because the cost will be increased dramatically. What is more, this approach may also cause a lot of confusion and tension in the education community and to the public, especially parents and school boards.

Recommendations for Further Research

This study demonstrated a technique for analyzing potential local item dependence with context-dependent testlets. Although the models function consistently, the lack-of-quality data leave some uncertainties on the inconsistent final results. To this author's knowledge, all the original testlets in the science tryout have been revised and one context-dependent multiple-choice item has been eliminated from each testlet in the operational forms. There is a need to use full operational data to conduct the study again to verify the outcomes.

Testlets in this study were multi-dimensional. It is necessary to use the models in this study to investigate the local item dependence with unidimensional testlets. It is anticipated that dimensionality of a testlet has an impact on the validity of the results.

As mentioned above, only the multiple-choice items within the testlets were used in the analysis. To fully investigate the local item dependence effects, full testlets, that is, multiple-choice items and constructed-response items, should be used in future studies.

Local dependence shown in the independent items and the original testlets when they were analyzed as testlets need to be studied further.

An alternative to examine the local item dependence of a test is to study the item relationship only when two or more items are found highly correlated to each other, temporarily

ignoring whether they are from the same testlet or not (see Yen, 1984a).

In this study, only the fit statistic generated from the BIGSTEPS was used. Other statistics such as Q_2 and Q_3 were mentioned but not considered in the analyses. In addition, R. Smith (April, 1996, personal contact) proposed a "between-fit" statistic contrary to Linacre and Wright's (1995) infit and outfit statistics. It will be helpful to the item/testlet analysis field to compare the efficiency of these and other currently available fit statistics.

APPENDICES

APPENDIX A

EXAMPLES OF PARTIAL CREDIT SCORING

Example 1. Mathematics item

$$\sqrt{9.0/0.3-5}=?$$

No steps taken 0
9.0/0.3 = 30 1
30 - 5 = 25 2
$\sqrt{25}$ = 5 3

Example 2. Screening test item

Draw a Circle

0	1	2	3
No response	Scribble, no resemblance to circle	Lack of closure, much overlap, more than 1/3 of figure distorted	Closure, no more than 2/3" overlap, 2/3 figure round

Example 3. Geography item

The capital city of Australia is

a. Wellington 1
b. Canberra 3
c. Montreal 0
d. Sydney 2

APPENDIX B

SAMPLE TESTLET IN THE MHSPT IN SCIENCE

Below is a data table which shows the melting and boiling points of common substances. Study the table. Then do Number 1 through 5.

Substance	Melting Point (°C)	Boiling Point (°C)
Water	0	100
Alcohol	-117	78
Nitrogen	-210	-196
Oxygen	-218	-183

1. Which substance should be a *liquid* at -90 degrees?

A water
B alcohol
C nitrogen
D oxygen

2. As each substance in the table is cooled down, the atoms and molecules undergo a

A physical changes as they move faster
B physical changes as they move slower
C chemical changes as they move faster
D chemical changes as they move slower

3. Because alcohol freezes and boils at lower temperatures than water, mixing alcohol and water could be a useful application for a

A better radiator coolant in cars during the summertime
B better windshield-washer fluid in cars during the wintertime
C clean and inexpensive alternative to gasoline
D clean and inexpensive alternative to engine lubricants

4. In order to change water from a solid to a liquid, energy must be

A removed
B added
C created
D destroyed

5. As water boil, the arrangement and behavior of the water molecules undergo changes. Describe at least two of these changes on the lines provided below.

APPENDIX C

MICHIGAN SCHOOL STRATUM CLASSIFICATION

The Michigan schools are classified into seven strata relative to populations where the schools reside.

1. Large City
   Central city of a Metropolitan Statistical Area (MSA) with a population greater than or equal to 400,000 or a population density greater than or equal to 6,000 people per square mile.

2. Mid-size City
   Central City of an MSA with a population less than 400,000 and a population density less than 6,000 people per square mile.

3. Urban Fringe of Large City
   Place within an MSA of a Large Central City and defined as urban by the Census Bureau.

4. Urban Fringe of Mid-size City
   Place within an MSA of a Mid-size Central City and defined as urban by the Census Bureau.

5. Large Town
   Town not within an MSA and with a population greater than or equal to 25,000 people.

6. Small Town
   Town not within an MSA and with a population less than 25,000 and greater than or equal to 2,500 people.

7. Rural
   A place with fewer than 2,500 people and coded rural by the Census Bureau.

APPENDIX D

ITEM CODE SHEET FOR TRYOUT FORM 22

Item Num.	Item Code	Dimension Content	Item Type	Item Num.	Item Code	Dimension Content	Item Type
1	L04	CELLS-COMP/RESP	MC	28	P13	SPEED/DIR CHANGE	MC
2	L06	CLASSFY ORGANISM	MC	29	R2	REFLECTING	MC
3	L14	ECO RELATIONSHIPS	MC	30	P10	ATOMIC CHANGES	MC
4	L08	FOOD STORAGE/USE	MC	31	C1	CONSTRUCTING	MC
5	R4	REFLECTING	MC	32	P13	SPEED/DIR CHANGE	OE
6	L12	NATURAL SELECTION	MC	33	R1	TEXT CRITICISM	OE
7	C1	CONSTRUCTING	MC	34	R1	TEXT CRITICISM	OE
8	L02	EXPLAIN GROWTH	MC	35	E02	USE MAPS	MC
9	L16	POPULATION SIZE	MC	36	E06	SOIL/SURFACE	MC
10	C1	CONSTRUCTING	MC	37	E09	WATER BELOW SURF	MC
11	L05	CELLS-FOOD/RESP	MC	38	E13	AIR/WEATHER	MC
12	L05	CELLS-FOOD/RESP	MC	39	E16	HUMANS/POPULATION	MC
13	C1	CONSTRUCTING	MC	40	E19	OBSERVE NITE SKY	MC
14	R2	REFLECTING	MC	41	E25	SPACE SCI/TECH	MC
15	L05	CELLS-FOOD/RESP	OE	42	R3	REFLECTING	MC
16	C1	INVESTIGATION	OE	43	C1	CONSTRUCTING	MC
17	C1	INVESTIGATION	OE	44	C1	CONSTRUCTING	MC
18	P01	CLASSFY SUBSTNCS	MC	45	C1	CONSTRUCTING	MC
19	P02	MASS/VOLUME/ DENS	MC	46	E23	EVOLUTION OF UNIVERSE	MC
20	P04	ANALYZE RISK/BEN	MC	47	E23	EVOLUTION OF UNIVERSE	MC
21	P18	SOUNDS/WAVES	MC	48	R1	REFLECTING	MC
22	P21	TYPES OF WAVES	MC	49	E23	EVOLUTION OF UNIVERSE	OE
23	R3	REFLECTING	MC	50	C1	CONSTRUCTING	MC
24	P11	ENERGY CHANGES	MC	51	P12	MEANS SPEED/DIRECTION	MC
25	P15	OBJECTS/FORCE	MC	52	E24	SOLAR SYST.FORM	MC
26	C1	CONSTRUCTING	MC	53	R1	REFLECTING	MC
27	C1	CONSTRUCTING	MC	54	P12	MEANS SPEED/ DIRECTION	OE

MC - Multiple-choice
OE - Open-ended

APPENDIX E

TABLES AND FIGURES

Table 7. Comparison of Original Testlets and Context-Dependent Items on Error and Fit by Form

Form 20

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	1.03			
			11	.07	.93*
			12	.07	1.02
			13	.07	1.05*
			14	.07	1.05
Testlet 2	.04	.95			
			28	.07	.97
			29	.07	1.14*
			30	.07	.98
			31	.07	.98
Testlet 3	.04	.97			
			45	.07	.91*
			46	.07	.94*
			47	.08	.91*
			48	.07	1.10*
Testlet 4	.04	.97			
			50	.07	.91*
			51	.07	1.0
			52	.07	.96
			53	.08	1.09*

* indicates where the standardized infit statistics are greater than ±2.0 SDs.

Table 7. (cont'd)

Form 21

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.94			
			11	.08	.93
			12	.07	1.01
			13	.07	1.10
			14	.07	1.03
Testlet 2	.04	1.17*			
			28	.07	1.03
			29	.09	1.24*
			30	.07	1.07*
			31	.08	1.18*
Testlet 3	.03	.90*			
			45	.07	.82*
			46	.07	.80*
			47	.07	1.00
			48	.07	.94*
Testlet 4	.04	.94			
			50	.07	1.02
			51	.07	.93*
			52	.07	.91*
			53	.07	1.08*

Table 7. (cont'd)

Form 22

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	1.03			
			11	.07	1.03
			12	.07	.89*
			13	.08	.91*
			14	.07	1.05
Testlet 2	.04	.90*			
			28	.07	.97
			29	.07	1.02
			30	.07	1.00
			31	.07	.91*
Testlet 3	.04	1.03			
			45	.08	.83*
			46	.08	1.25*
			47	.07	1.16*
			48	.07	.90*
Testlet 4	.04	.91*			
			50	.07	.89*
			51	.07	.96
			52	.07	1.22*
			53	.07	.96

Table 7. (cont'd)

Form 23

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.96			
			11	.07	.97
			12	.07	1.11*
			13	.09	.92
			14	.07	.99
Testlet 2	.04	.98			
			28	.07	1.05
			29	.07	.96
			30	.07	.96
			31	.08	.97
Testlet 3	.04	1.22*			
			45	.07	1.13*
			46	.08	1.16*
			47	.07	1.09*
			48	.07	1.10*
Testlet 4	.04	.76*			
			50	.09	.88*
			51	.07	.88*
			52	.07	.86*
			53	.09	.86*

Table 7. (cont'd)

Form 24

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.88*			
			11	.08	1.18*
			12	.07	.93*
			13	.07	.91*
			14	.08	.87*
Testlet 2	.04	1.18*			
			28	.08	1.12*
			29	.07	1.07*
			30	.07	1.12*
			31	.09	.92
Testlet 3	.04	.89*			
			45	.07	1.08*
			46	.07	1.12*
			47	.07	.90*
			48	.08	.83*
Testlet 4	.04	.87*			
			50	.07	.98
			51	.07	.99
			52	.07	.98
			53	.07	.85*

Table 7. (cont'd)

Form 25

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.94			
			11	.07	.97
			12	.07	.92*
			13	.07	1.09*
			14	.11	.86
Testlet 2	.04	.93			
			28	.07	1.07
			29	.07	.98
			30	.07	.91*
			31	.07	.95
Testlet 3	.04	.92			
			45	.07	.95*
			46	.07	.99
			47	.07	.95*
			48	.08	1.02
Testlet 4	.04	1.12*			
			50	.07	1.08*
			51	.07	1.03
			52	.07	1.03
			53	.08	1.10*

Table 7. (cont'd)

Form 26

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.94			
			11	.08	1.02
			12	.08	.94
			13	.08	.91*
			14	.08	1.03
Testlet 2	.04	1.08			
			28	.08	1.16*
			29	.08	1.09*
			30	.10	.89
			31	.09	.89*
Testlet 3	.04	.88*			
			45	.08	.95
			46	.08	1.01
			47	.08	.99
			48	.08	.97
Testlet 4	.04	1.01			
			50	.07	.91*
			51	.09	.93
			52	.11	1.27*
			53	.08	1.00

Table 7. (cont'd)

Form 27

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	1.00			
			11	.08	.96
			12	.08	1.16*
			13	.07	1.04
			14	.07	.93*
Testlet 2	.04	1.12*			
			28	.07	1.02
			29	.08	1.31*
			30	.09	1.04
			31	.07	.92*
Testlet 3	.04	.79*			
			45	.09	.83*
			46	.07	.90*
			47	.08	.92*
			48	.08	.86*
Testlet 4	.04	.95			
			50	.09	1.31*
			51	.07	.90*
			52	.08	.88*
			53	.08	.85*

Table 7. (cont'd)

Form 28

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.84*			
			11	.07	.90*
			12	.07	.91*
			13	.07	.94*
			14	.07	.95
Testlet 2	.04	1.20*			
			28	.08	1.31*
			29	.08	1.10*
			30	.08	1.11*
			31	.07	.95
Testlet 3	.04	.88*			
			45	.07	.98
			46	.07	.90*
			47	.07	.85*
			48	.07	.96
Testlet 4	.04	1.05			
			50	.08	1.09*
			51	.07	.97
			52	.08	1.05
			53	.07	1.07*

Table 7. (cont'd)

Form 29

1 Orig. Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Context Depend. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.90*			
			11	.08	1.08*
			12	.07	.95
			13	.08	.92
			14	.09	.87*
Testlet 2	.04	.86*			
			28	.08	1.06
			29	.08	.87*
			30	.10	.94
			31	.07	.92*
Testlet 3	.04	1.24*			
			45	.08	.98
			46	.07	1.09*
			47	.07	1.12*
			48	.08	1.26*
Testlet 4	.04	.90*			
			50	.09	.81*
			51	.07	1.05
			52	.07	.95
			53	.07	1.04

Table 10. Comparison of Random Testlets and Independent Items on Error and Fit by Form

Form 20

1 Random Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Indep. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.05	.78*			
			1	.14	.95
			8	.09	.91
			24	.07	1.10*
			38	.08	.84*
Testlet 2	.04	1.11*			
			2	.07	1.03
			9	.07	1.19*
			25	.07	1.01
			40	.08	.92*
Testlet 3	.04	.98			
			3	.07	1.04
			18	.07	.94*
			27	.07	.88*
			41	.07	1.00
Testlet 4	.04	.88*			
			4	.07	1.00
			20	.08	1.18*
			37	.08	.90*
			43	.07	.97

* indicates where the standardized infit statistics are greater than ±2.0 SDs.

Table 10. (cont'd)

Form 21

1	2	3	4	5	6
Random Testlet	SE of Testlet	Testlet Infit MNSQ	Indep. Item	SE of Item	Item Infit MNSQ
Testlet 1	.04	.85*			
			1	.08	1.11*
			8	.07	.90*
			24	.07	.93*
			38	.07	1.01
Testlet 2	.04	1.05			
			2	.07	.93*
			9	.08	.92*
			25	.07	1.00
			40	.09	1.14*
Testlet 3	.04	1.11*			
			3	.07	1.02
			18	.08	.97
			27	.07	1.06*
			41	.07	1.04
Testlet 4	.04	.87*			
			4	.07	1.00
			20	.07	1.00
			37	.09	1.00
			43	.07	.95

Table 10. (cont'd)

Form 22

1	2	3	4	5	6
		Testlet			Item
Random	SE of	Infit	Indep.	SE of	Infit
Testlet	Testlet	MNSQ	Item	Item	MNSQ
Testlet 1	.04	.76*			
			1	.07	.88*
			8	.07	.94*
			24	.07	.96
			38	.07	1.07*
Testlet 2	.04	1.10*			
			2	.07	1.09*
			9	.07	.93*
			25	.08	1.01
			40	.07	1.19*
Testlet 3	.04	1.20*			
			3	.07	.93*
			18	.11	1.26*
			27	.07	.97
			41	.07	1.03
Testlet 4	.04	.77*			
			4	.07	.87*
			20	.07	1.05*
			37	.07	.99
			43	.07	.88*

Table 10. (cont'd)

Form 23

1 Random Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Indep. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.84*			
			1	.07	1.10*
			8	.09	.83*
			24	.07	1.06*
			38	.07	.95*
Testlet 2	.04	1.11*			
			2	.07	.93*
			9	.07	.97
			25	.07	.98
			40	.08	1.17*
Testlet 3	.04	.97			
			3	.07	1.03
			18	.07	1.03
			27	.07	.93*
			41	.08	.89*
Testlet 4	.04	.90*			
			4	.07	.97
			20	.07	.96
			37	.09	.91
			43	.08	1.26*

Table 10. (cont'd)

Form 24

1 Random Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Indep. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.95			
			1	.08	.97
			8	.07	1.11*
			24	.08	.95
			38	.07	1.03
Testlet 2	.04	.98			
			2	.08	.94
			9	.07	1.15*
			25	.08	.82*
			40	.07	1.10*
Testlet 3	.04	.98			
			3	.08	.89*
			18	.09	1.10
			27	.09	.93
			41	.07	1.03
Testlet 4	.04	.81*			
			4	.09	1.01
			20	.07	1.04
			37	.10	.89
			43	.07	.92*

Table 10. (cont'd)

Form 25

1 Random Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Indep. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.83*			
			1	.07	.90*
			8	.10	.97
			24	.07	1.00
			38	.07	.94*
Testlet 2	.04	.98			
			2	.08	.93
			9	.07	.96
			25	.07	.93*
			40	.08	1.14*
Testlet 3	.04	1.00			
			3	.07	.91*
			18	.07	1.00
			27	.07	.94*
			41	.07	1.09*
Testlet 4	.04	1.03			
			4	.07	1.10*
			20	.07	1.08*
			37	.07	1.10*
			43	.07	.95

Table 10. (cont'd)

Form 26

1 Random Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Indep. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.05	.85*			
			1	.08	.93
			8	.13	.92
			24	.08	1.04
			38	.08	.90*
Testlet 2	.04	.94			
			2	.09	.90*
			9	.08	1.01
			25	.08	1.08*
			40	.08	.90*
Testlet 3	.04	1.20*			
			3	.07	1.09*
			18	.08	.99
			27	.07	1.04
			41	.09	1.10*
Testlet 4	.04	.85*			
			4	.07	1.15*
			20	.09	.98
			37	.08	.93*
			43	.08	.95

Table 10. (cont'd)

Form 27

1 Random Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Indep. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.85*			
			1	.07	1.07*
			8	.08	.93*
			24	.09	1.17*
			38	.08	.89*
Testlet 2	.05	1.08			
			2	.09	.93
			9	.08	.92*
			25	.07	1.17*
			40	.08	.99
Testlet 3	.04	.99			
			3	.07	.93*
			18	.07	1.04
			27	.07	.95
			41	.07	.98
Testlet 4	.04	.87*			
			4	.07	1.08*
			20	.07	1.01
			37	.07	.95*
			43	.09	.99

Table 10. (cont'd)

Form 28

1 Random Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Indep. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.77*			
			1	.08	.83*
			8	.07	.99
			24	.07	.99
			38	.07	.91*
Testlet 2	.04	1.14*			
			2	.07	1.06*
			9	.08	1.10*
			25	.08	1.15*
			40	.07	.88*
Testlet 3	.04	1.04			
			3	.07	.91*
			18	.07	1.08*
			27	.08	.93
			41	.07	1.02
Testlet 4	.04	.92			
			4	.07	1.08*
			20	.08	1.09*
			37	.07	.95
			43	.07	.99

Table 10. (cont'd)

Form 29

1 Random Testlet	2 SE of Testlet	3 Testlet Infit MNSQ	4 Indep. Item	5 SE of Item	6 Item Infit MNSQ
Testlet 1	.04	.80*			
			1	.07	1.05
			8	.08	.87*
			24	.07	.89*
			38	.07	1.09*
Testlet 2	.04	1.04			
			2	.08	.99
			9	.08	.92*
			25	.08	1.27*
			40	.07	.92*
Testlet 3	.05	1.10*			
			3	.08	1.01
			18	.08	1.02
			27	.11	.94
			41	.09	1.10
Testlet 4	.04	.82*			
			4	.08	.94
			20	.08	1.10*
			37	.07	.91*
			43	.07	.91*

Table 11. Degrees of Freedom for the Context-dependent Items

FORM 20

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	.42	1029.00	.93	-.07	.88	-.13
1	TEL12	.41	1028.00	1.02	.02	1.03	.03
1	TEL13	-.04	1025.00	1.05	.05	1.07	.07
1	TEL14	-1.02	1025.00	1.05	.05	1.08	.08
2	TEL28	-.13	1024.00	.97	-.03	.97	-.03
2	TEL29	.17	1024.00	1.14	.13	1.21	.19
2	TEL30	-.79	1022.00	.98	-.02	.98	-.02
2	TEL31	-.46	1025.00	.98	-.02	.99	-.01
3	TEL45	-.08	1021.00	.91	-.09	.89	-.12
3	TEL46	.24	1017.00	.94	-.06	.94	-.06
3	TEL47	.77	1018.00	.91	-.09	.91	-.09
3	TEL48	.50	1018.00	1.10	.10	1.16	.15
4	TEL50	-.26	1014.00	.91	-.09	.89	-.12
4	TEL51	.31	1018.00	1.00	.00	1.00	.00
4	TEL52	-1.11	1020.00	.96	-.04	.96	-.04
4	TEL53	1.05	1019.00	1.09	.09	1.23	.21

FORM 21

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	-1.12	1044.00	.93	-.07	.87	-.14
1	TEL12	.34	1045.00	1.01	.01	1.04	.04
1	TEL13	-.73	1045.00	1.03	.03	1.07	.07
1	TEL14	-.60	1044.00	1.03	.03	1.04	.04
2	TEL28	-.55	1038.00	1.03	.03	1.01	.01
2	TEL29	1.88	1036.00	1.24	.22	1.63	.49
2	TEL30	-.39	1038.00	1.07	.07	1.15	.14
2	TEL31	1.47	1038.00	1.18	.17	1.39	.33
3	TEL45	-.49	1036.00	.82	-.20	.73	-.31
3	TEL46	-.06	1033.00	.80	-.22	.75	-.29
3	TEL47	-.03	1036.00	1.00	.00	1.00	.00
3	TEL48	-.23	1035.00	.94	-.06	.93	-.07
4	TEL50	-.37	1024.00	1.02	.02	1.02	.02
4	TEL51	-.29	1033.00	.93	-.07	.87	-.14
4	TEL52	.43	1032.00	.91	-.09	.86	-.15
4	TEL53	.75	1019.00	1.08	.08	1.12	.11

FORM 22

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	-.05	1041.00	1.03	.03	1.04	.04
1	TEL12	.45	1039.00	.89	-.12	.85	-.16
1	TEL13	-1.27	1038.00	.91	-.09	.81	-.21
1	TEL14	.66	1038.00	1.05	.05	1.10	.10
2	TEL28	-.82	1038.00	.97	-.03	.96	-.04
2	TEL29	-.37	1040.00	1.02	.02	1.03	.03
2	TEL30	.84	1040.00	1.00	.00	1.06	.06
2	TEL31	-.42	1038.00	.91	-.09	.88	-.13
3	TEL45	-1.53	1035.00	.83	-.19	.67	-.40
3	TEL46	1.10	1035.00	1.25	.22	1.51	.41
3	TEL47	.48	1035.00	1.16	.15	1.22	.20
3	TEL48	-.72	1033.00	.90	-.11	.88	-.13
4	TEL50	-.33	1028.00	.89	-.12	.85	-.16
4	TEL51	.79	1034.00	.96	-.04	1.00	.00
4	TEL52	.83	1033.00	1.22	.20	1.33	.29
4	TEL53	.35	1033.00	.96	-.04	.98	-.02

Table 11. (cont'd)

FORM 23

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	-.27	1050.00	.97	-.03	.97	-.03
1	TEL12	-.22	1040.00	1.11	.10	1.19	.17
1	TEL13	-1.64	1049.00	.92	-.08	.81	-.21
1	TEL14	.36	1045.00	.99	-.01	1.07	.07
2	TEL28	.99	1040.00	1.05	.05	1.08	.08
2	TEL29	-.17	1040.00	.96	-.04	.94	-.06
2	TEL30	-.46	1038.00	.96	-.04	.88	-.13
2	TEL31	1.61	1039.00	.97	-.03	1.10	.10
3	TEL45	-.08	1032.00	1.13	.12	1.20	.18
3	TEL46	1.39	1030.00	1.16	.15	1.27	.24
3	TEL47	.90	1032.00	1.09	.09	1.13	.12
3	TEL48	1.19	1031.00	1.10	.10	1.17	.16
4	TEL50	-1.63	1024.00	.88	-.13	.81	-.21
4	TEL51	.31	1028.00	.88	-.13	.85	-.16
4	TEL52	-.66	1028.00	.86	-.15	.75	-.29
4	TEL53	-1.62	1026.00	.86	-.15	.70	-.36

FORM 24

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	1.58	1016.00	1.18	.17	1.58	.46
1	TEL12	-.60	1023.00	.93	-.07	.94	-.06
1	TEL13	-.65	1020.00	.91	-.09	.87	-.14
1	TEL14	-1.04	1023.00	.87	-.14	.78	-.25
2	TEL28	1.76	1018.00	1.12	.11	1.47	.39
2	TEL29	-.30	1020.00	1.07	.07	1.05	.05
2	TEL30	.81	1018.00	1.12	.11	1.27	.24
2	TEL31	-2.00	1021.00	.92	-.08	.82	-.20
3	TEL45	.60	1015.00	1.08	.08	1.15	.14
3	TEL46	.98	1014.00	1.12	.11	1.30	.26
3	TEL47	-.67	1012.00	.90	-.11	.87	-.14
3	TEL48	-1.34	1015.00	.83	-.19	.72	-.33
4	TEL50	-.20	1007.00	.98	-.02	1.01	.01
4	TEL51	.44	1010.00	.99	-.01	1.02	.02
4	TEL52	1.09	1008.00	.98	-.02	.97	-.03
4	TEL53	-.46	1009.00	.85	-.16	.79	-.24

FORM 25

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	-.67	1013.00	.97	-.03	.95	-.05
1	TEL12	-.96	1013.00	.92	-.08	.86	-.15
1	TEL13	.26	1010.00	1.09	.09	1.13	.12
1	TEL14	-2.49	1014.00	.86	-.15	.65	-.43
2	TEL28	1.12	1005.00	1.07	.07	1.17	.16
2	TEL29	-.90	1008.00	.98	-.02	1.06	.06
2	TEL30	-.66	1005.00	.91	-.09	.86	-.15
2	TEL31	.85	1005.00	.95	-.05	1.10	.10
3	TEL45	-.26	999.00	.95	-.05	.92	-.08
3	TEL46	.22	996.00	.99	-.01	1.02	.02
3	TEL47	.36	993.00	.95	-.05	.92	-.08
3	TEL48	1.39	995.00	1.02	.02	1.14	.13
4	TEL50	.35	989.00	1.08	.08	1.13	.12
4	TEL51	.18	993.00	1.03	.03	1.08	.08
4	TEL52	-.03	992.00	1.03	.03	1.03	.03
4	TEL53	1.23	987.00	1.10	.10	1.23	.21

Table 11. (cont'd)

FORM 26

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	-.67	894.00	1.02	.02	1.02	.02
1	TEL12	.99	892.00	.94	-.06	.99	-.01
1	TEL13	-.93	895.00	.91	-.09	.80	-.22
1	TEL14	-.27	894.00	1.03	.03	1.05	.05
2	TEL28	.87	894.00	1.16	.15	1.26	.23
2	TEL29	.93	893.00	1.09	.09	1.13	.12
2	TEL30	-2.01	893.00	.89	-.12	.74	-.30
2	TEL31	-1.61	891.00	.89	-.12	.76	-.27
3	TEL45	.26	888.00	.95	-.05	.95	-.05
3	TEL46	.41	889.00	1.01	.01	1.05	.05
3	TEL47	.71	885.00	.99	-.01	1.00	.00
3	TEL48	-.33	888.00	.97	-.03	.99	-.01
4	TEL50	.04	889.00	.91	-.09	.87	-.14
4	TEL51	-1.44	883.00	.93	-.07	.90	-.11
4	TEL52	2.62	889.00	1.27	.24	2.24	.81
4	TEL53	.43	890.00	1.00	.00	.98	-.02

FORM 27

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	-1.10	944.00	.96	-.04	.88	-.13
1	TEL12	.84	936.00	1.16	.15	1.31	.27
1	TEL13	.42	943.00	1.04	.04	1.05	.05
1	TEL14	-.03	943.00	.93	-.07	.90	-.11
2	TEL28	.17	931.00	1.02	.02	1.03	.03
2	TEL29	1.05	931.00	1.31	.27	1.61	.48
2	TEL30	1.68	932.00	1.04	.04	1.31	.27
2	TEL31	.27	931.00	.92	-.08	.92	-.08
3	TEL45	-1.52	923.00	.83	-.19	.72	-.33
3	TEL46	-.32	923.00	.90	-.11	.87	-.14
3	TEL47	-.96	922.00	.92	-.08	.88	-.13
3	TEL48	-1.24	923.00	.86	-.15	.73	-.31
4	TEL50	2.06	913.00	1.31	.27	2.95	1.08
4	TEL51	.20	920.00	.90	-.11	.87	-.14
4	TEL52	-.80	920.00	.88	-.13	.83	-.19
4	TEL53	-.72	919.00	.85	-.16	.77	-.26

FORM 28

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	-.15	942.00	.90	-.11	.87	-.14
1	TEL12	-.64	941.00	.91	-.09	.87	-.14
1	TEL13	-.53	942.00	.94	-.06	.93	-.07
1	TEL14	-.57	940.00	.95	-.05	.91	-.09
2	TEL28	.67	933.00	1.31	.27	1.40	.34
2	TEL29	1.14	934.00	1.10	.10	1.21	.19
2	TEL30	.60	935.00	1.11	.10	1.12	.11
2	TEL31	-.39	934.00	.95	-.05	.94	-.06
3	TEL45	.20	926.00	.98	-.02	.97	-.03
3	TEL46	-.16	924.00	.90	-.11	.87	-.14
3	TEL47	-.58	925.00	.85	-.16	.79	-.24
3	TEL48	-.42	926.00	.96	-.04	.93	-.07
4	TEL50	.50	907.00	1.09	.09	1.13	.12
4	TEL51	-.27	921.00	.97	-.03	.96	-.04
4	TEL52	.53	919.00	1.05	.05	1.07	.07
4	TEL53	.08	920.00	1.07	.07	1.11	.10

Table 11. (cont'd)

FORM 29

ORIGIN TESTLET	ITEM NAME	ITEM CALIBR	DF	INFIT MNSQ	LN(INFIT)	OUTFIT MNSQ	LN(OUTFIT)
1	TEL11	-.20	943.00	1.08	.08	1.07	.07
1	TEL12	.27	944.00	.95	-.05	.94	-.06
1	TEL13	-.99	943.00	.92	-.08	.78	-.25
1	TEL14	-1.45	944.00	.87	-.14	.67	-.40
2	TEL28	1.88	939.00	1.06	.06	1.46	.38
2	TEL29	-.28	939.00	.87	-.14	.78	-.25
2	TEL30	-1.82	939.00	.94	-.06	.79	-.24
2	TEL31	.36	942.00	.92	-.08	.88	-.13
3	TEL45	-.81	941.00	.98	-.02	.97	-.03
3	TEL46	.34	940.00	1.09	.09	1.14	.13
3	TEL47	.54	940.00	1.12	.11	1.17	.16
3	TEL48	1.81	941.00	1.26	.23	1.50	.41
4	TEL50	-1.29	938.00	.81	-.21	.65	-.43
4	TEL51	.41	941.00	1.05	.05	1.09	.09
4	TEL52	.73	937.00	.95	-.05	.93	-.07
4	TEL53	.49	940.00	1.04	.04	1.07	.07

Table 12. Discrepancies For Testlets in the Tryout Forms

Form	Testlet 1	Testlet 2	Testlet 3	Testlet 4
20	4	3	4	5
21	1	2	3	9
22	3	2	2	6
23	5	2	2	4
24	7	3	3	3
25	4	3	6	6
26	3	3	4	7
27	8	1	1	7
28	2	2	2	14
29	1	3	1	4

Table 13. CIs for One-Way ANOVA for Context-dependent Items

Form 20

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	.0112	.0575	.0287	-.0803 TO .1027
Testlet 2	4	.0150	.0775	.0387	-.1082 TO .1383
Testlet 3	4	-.0388	.0907	.0454	-.1831 TO .1055
Testlet 4	4	-.0122	.0761	.0381	-.1334 TO .1089

F ratio = .4236 probability = .7396

Form 21

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	-.0009	.0487	.0243	-.0783 TO .0766
Testlet 2	4	.1195	.0857	.0429	-.0169 TO .2558
Testlet 3	4	-.1209	.1073	.0537	-.2917 TO .0499
Testlet 4	4	-.0175	.0801	.0400	-.1450 TO .1099

F ratio = 5.6103 probability = .0122

Form 22

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	-.0331	.0843	.0422	-.1673 TO .1011
Testlet 2	4	-.0262	.0499	.0249	-.1056 TO .0531
Testlet 3	4	.0200	.1967	.0983	-.2930 TO .3329
Testlet 4	4	.0002	.1372	.0686	-.2181 TO .2184

F ratio = .1431 probability = .9322

Form 23

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	-.0049	.0791	.0396	-.1308 TO .1210
Testlet 2	4	-.0158	.0434	.0217	-.0848 TO .0532
Testlet 3	4	.1130	.0281	.0141	.0683 TO .1578
Testlet 4	4	-.1393	.0133	.0066	-.1604 TO -.1182

F ratio = 18.6909 probability = .0001

Form 24

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	-.0352	.1366	.0683	-.2526 TO .1823
Testlet 2	4	.0527	.0933	.0466	-.0957 TO .2011
Testlet 3	4	-.0254	.1438	.0719	-.2541 TO .2034
Testlet 4	4	-.0532	.0730	.0365	-.1694 TO .0629

F ratio = .6559 probability = .5946

Form 25

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	-.0446	.1002	.0501	-.2040 TO .1147
Testlet 2	4	-.0245	.0686	.0343	-.1336 TO .0846
Testlet 3	4	-.0232	.0346	.0173	-.0783 TO .0319
Testlet 4	4	.0578	.0335	.0168	.0045 TO .1112

F ratio = 1.9327 probability = .1782

Table 13. (Cont'd)

Form 26

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	-.0267	.0609	.0305	-.1237 TO .0702
Testlet 2	4	.0004	.1374	.0687	-.2182 TO .2190
Testlet 3	4	-.0205	.0264	.0132	-.0624 TO .0215
Testlet 4	4	.0180	.1527	.0764	-.2250 TO .2611

F ratio = .1431 probability = .9321

Form 27

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	.0186	.0985	.0493	-.1382 TO .1753
Testlet 2	4	.0614	.1491	.0746	-.1759 TO .2987
Testlet 3	4	-.1315	.0461	.0231	-.2048 TO -.0581
Testlet 4	4	-.0314	.2023	.1012	-.3534 TO .2905

F ratio = 1.4697 probability = .2722

Form 28

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	-.0782	.0257	.0129	-.1192 TO -.0373
Testlet 2	4	.1046	.1313	.0657	-.1044 TO .3136
Testlet 3	4	-.0822	.0647	.0323	-.1851 TO .0207
Testlet 4	4	.0430	.0513	.0257	-.0386 TO .1247

F ratio = 5.5278 probability = .0128

Form 29

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean
Testlet 1	4	-.0492	.0917	.0458	-.1951 TO .0966
Testlet 2	4	-.0566	.0832	.0416	-.1890 TO .0758
Testlet 3	4	.1026	.1032	.0516	-.0617 TO .2669
Testlet 4	4	-.0435	.1203	.0601	-.2349 TO .1478

F ratio = 2.3075 probability = .1284

Table 14. Summary of Measured (Non-Extreme) Persons Fit by Form

1	2	3	4
Item/testlet Composition	Mean Measure	Infit MNSQ	Outfit MNSQ
Form 20 (n=1030)			
16 context-dependent items	-.27	1.00	1.01
4 original testlets	-.28	.97	.97
4 reformed testlets	-.37	.96	.96
16 MC independent items	.18	1.00	1.02
4 random testlets	.12	.94	.93
Form 21 (n=1046)			
16 context-dependent items	.06	.99	1.03
4 original testlets	.04	.95	.97
4 reformed testlets	-.02	.93	.94
16 MC independent items	-.29	1.00	1.02
4 random testlets	-.38	.97	.97
Form 22 (n=1044)			
16 context-dependent items	-.03	1.00	1.01
4 original testlets	-.04	.96	.97
4 reformed testlets	-.11	.95	.96
16 MC independent items	-.26	.99	1.05
4 random testlets	-.30	.92	.96
Form 23 (n=1051)			
16 context-dependent items	.25	1.00	.99
4 original testlets	.20	.95	.96
4 reformed testlets	.22	.94	.95
16 MC independent items	.36	1.00	1.00
4 random testlets	.37	.95	.94
Form 24 (n=1024)			
16 context-dependent items	.10	.99	1.04
4 original testlets	.08	.94	.95
4 reformed testlets	.00	.93	.96
16 MC independent items	.57	.99	1.02
4 random testlets	.71	.92	.92

Table 14 (cont'd)

1	2	3	4
Item/testlet Composition	Mean Measure	Infit MNSQ	Outfit MNSQ
Form 25 (n=1016)			
16 context-dependent items	.07	1.00	1.02
4 original testlets	-.03	.96	.98
4 reformed testlets	.06	.96	.98
16 MC ındependent items	.21	1.00	1.04
4 random testlets	.10	.96	.96
Form 26 (n=896)			
16 context-dependent items	.15	.99	1.05
4 original testlets	.16	.96	.97
4 reformed testlets	.05	.93	.96
16 MC independent items	.63	1.00	1.00
4 random testlets	.78	.95	.96
Form 27 (n=945)			
16 context-dependent items	.14	.98	1.10
4 original testlets	.03	.93	.95
4 reformed testlets	.05	.92	.93
16 MC independent items	.47	.99	.99
4 random testlets	.61	.94	.96
Form 28 (n=944)			
16 context-dependent items	-.22	1.00	1.01
4 original testlets	-.36	.95	.97
4 reformed testlets	-.33	.96	.96
16 MC independent items	.01	1.00	1.00
4 random testlets	-.09	.96	.97
Form 29 (n=947)			
16 context-dependent items	.47	1.00	.99
4 original testlets	.47	.97	.97
4 reformed testlets	.53	.94	.94
16 MC independent items	.09	.99	1.03
4 random testlets	.14	.92	.96

Table 15. Person Separation Ratios for Different Configurations by Form

1	2	3	4	5
Item/testlet Composition	Mean Measure	Real RMSE	Adj. SD	Separa. Ratio
Form 20 (n=1030)				
16 context-dependent items	-.27	.61	.78	1.28
4 original testlets	-.28	.70	.85	1.21
4 reformed testlets	-.37	.71	.97	1.36
16 MC independent items	.18	.66	.85	1.82
4 random testlets	.12	.79	1.20	1.52
Form 21 (n=1046)				
16 context-dependent items	.06	.63	.89	1.42
4 original testlets	.04	.70	.89	1.27
4 reformed testlets	-.02	.76	1.20	1.59
16 MC independent items	-.29	.63	.55	.87
4 random testlets	-.38	.76	.81	1.06
Form 22 (n=1044)				
16 context-dependent items	-.03	.63	.84	1.33
4 original testlets	-.04	.72	.98	1.36
4 reformed testlets	-.11	.74	1.06	1.45
16 MC independent items	-.26	.62	.77	1.25
4 random testlets	-.30	.74	1.05	1.42
Form 23 (n=1051)				
16 context-dependent items	.25	.66	.87	1.32
4 original testlets	.20	.72	.91	1.26
4 reformed testlets	.22	.75	1.14	1.51
16 MC independent items	.36	.63	.75	1.19
4 random testlets	.37	.77	1.11	1.45
Form 24 (n=1024)				
16 context-dependent items	.10	.65	.84	1.30
4 original testlets	.08	.74	1.01	1.36
4 reformed testlets	.00	.75	1.03	1.38
16 MC independent items	.57	.68	.95	1.40
4 random testlets	.71	.79	1.25	1.59

Table 15 (Cont'd)

1	2	3	4	5
Item/testlet Composition	Mean Measure	Real RMSE	Adj. SD	Separation
Form 25 (n=1016)				
16 context-dependent items	.07	.64	.71	1.12
4 original testlets	-.03	.70	.75	1.06
4 reformed testlets	.06	.73	.89	1.22
16 MC independent items	.21	.62	.72	1.16
4 random testlets	.10	.74	.97	1.31
Form 26 (n=896)				
16 context-dependent items	.15	.66	.90	1.36
4 original testlets	.16	.73	.96	1.32
4 reformed testlets	.05	.79	1.19	1.50
16 MC independent items	.63	.66	.86	1.30
4 random testlets	.78	.77	1.08	1.41
Form 27 (n=945)				
16 context-dependent items	.14	.66	.87	1.32
4 original testlets	.03	.73	.93	1.27
4 reformed testlets	.05	.77	1.22	1.58
16 MC independent items	.47	.64	.80	1.25
4 random testlets	.61	.76	1.13	1.49
Form 28 (n=944)				
16 context-dependent items	-.22	.61	.77	1.26
4 original testlets	-.36	.68	.81	1.19
4 reformed testlets	-.33	.72	1.04	1.43
16 MC independent items	.01	.61	.78	1.27
4 random testlets	-.09	.73	1.04	1.43
Form 29 (n=947)				
16 context-dependent items	.47	.67	.94	1.41
4 original testlets	.47	.74	1.01	1.36
4 reformed testlets	.53	.77	1.17	1.53
16 MC independent items	.09	.64	.81	1.26
4 random testlets	.14	.76	1.06	1.39

Table 17. Comparisons of Average Measures for the Original and Random Testlets

Form 20 (n=1030)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.38	1.11
1	-.78	1.08
2	-.21	1.02
3	.43	1.04
4	1.32	.92
Random Testlet 1		
0	-2.57	.91
1	-1.79	.78
2	-.88	.88
3	.00	.75
4	1.57	.74
Original Testlet 2		
0	-1.61	.93
1	-.98	.85
2	-.28	.98
3	.45	.89
4	.93	1.15
Random Testlet 2		
0	-1.31	1.07
1	-.55	1.08
2	.36	1.04
3	1.38	1.15
4	2.39	1.44
Original Testlet 3		
0	-1.14	1.10
1	-.59	.98
2	-.07	1.02
3	.74	.90
4	1.46	.81
Random Testlet 3		
0	-1.40	1.19
1	-.80	.93
2	.01	.98
3	1.01	.90
4	2.30	.97
Original Testlet 4		
0	-1.45	.98
1	-.78	.96
2	-.18	1.01
3	.56	.91
4	1.24	1.03
Random Testlet 4		
0	-1.52	.90
1	-.62	.81
2	.33	.91
3	1.64	.88
4	2.76	1.08

Table 17. (cont'd)

Form 21 (n=1046)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.51	.99
1	-.92	.91
2	-.29	.91
3	.34	1.03
4	1.19	.92
Random Testlet 1		
0	-2.02	.84
1	-1.15	.84
2	-.35	.87
3	.49	.77
4	.99	1.03
Original Testlet 2		
0	-1.15	1.02
1	-.33	1.20
2	.30	1.11
3	.89	1.25
4	1.61	1.55
Random Testlet 2		
0	-1.61	1.12
1	-.86	1.00
2	-.12	1.02
3	.58	1.02
4	.89	1.27
Original Testlet 3		
0	-1.15	1.00
1	-.65	.88
2	-.27	.90
3	.47	.84
4	1.18	.88
Random Testlet 3		
0	-1.77	1.06
1	-.89	1.17
2	-.24	1.05
3	.35	1.12
4	1.08	1.20
Original Testlet 4		
0	-1.11	1.04
1	-.60	.79
2	.01	.98
3	.75	.88
4	1.48	.97
Random Testlet 4		
0	-1.70	.92
1	-.84	.83
2	-.04	.88
3	.65	.91
4	1.54	.84

Table 17. (cont'd)

Form 22 (n=1044)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.56	.97
1	-.64	1.16
2	-.07	1.07
3	.45	1.25
4	1.61	.75
Random Testlet 1		
0	-2.05	.83
1	-1.33	.68
2	-.54	.80
3	.37	.69
4	1.22	.83
Original Testlet 2		
0	-1.64	.99
1	-.93	.86
2	-.17	.95
3	.62	.84
4	1.47	.92
Random Testlet 2		
0	-1.77	1.19
1	-1.06	1.08
2	-.26	1.10
3	.70	.97
4	1.38	1.38
Original Testlet 3		
0	-1.73	.98
1	-.95	.92
2	-.09	1.06
3	.80	.85
4	1.02	1.52
Random Testlet 3		
0	-1.22	1.24
1	-.49	1.04
2	.28	1.19
3	.1.29	1.24
4	1.35	1.97
Original Testlet 4		
0	-1.32	1.01
1	-.56	.89
2	.23	.90
3	1.03	.87
4	1.87	.91
Random Testlet 4		
0	-1.99	.85
1	-1.23	.77
2	-.47	.83
3	.40	.72
4	1.41	.72

Table 17. (cont'd)

Form 23 (n=1051)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.19	1.16
1	-.90	.88
2	-.17	.94
3	.50	.98
4	1.37	.94
Random Testlet 1		
0	-1.91	.82
1	-1.01	.90
2	-.17	.84
3	.71	.86
4	1.87	.81
Original Testlet 2		
0	-.93	1.05
1	-.34	.99
2	.23	1.01
3	.95	.96
4	2.00	.91
Random Testlet 2		
0	-1.23	1.06
1	-.46	1.03
2	.49	1.15
3	1.38	1.10
4	1.93	1.41
Original Testlet 3		
0	-.77	1.19
1	-.12	1.21
2	.46	1.04
3	1.31	1.22
4	1.56	1.76
Random Testlet 3		
0	-1.75	.94
1	-.93	.88
2	.05	.98
3	.84	.95
4	1.76	1.12
Original Testlet 4		
0	-1.52	.96
1	-1.05	.70
2	-.46	.75
3	.27	.67
4	1.16	.76
Random Testlet 4		
0	-1.80	.84
1	-.76	.92
2	.10	.82
3	1.25	.85
4	1.96	1.21

Table 17. (cont'd)

Form 24 (n=1024)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.76	.81
1	-.98	.84
2	-.13	.91
3	.77	.83
4	1.53	1.11
Random Testlet 1		
0	-1.97	.99
1	-.88	.88
2	.21	.93
3	1.29	1.01
4	2.32	.98
Original Testlet 2		
0	-1.70	.99
1	-.59	1.27
2	.06	1.11
3	.80	1.29
4	1.85	1.15
Random Testlet 2		
0	-1.64	.91
1	-.49	.94
2	.50	1.01
3	1.53	.95
4	2.41	1.14
Original Testlet 3		
0	-1.69	.93
1	-.93	.86
2	-.02	.77
3	.79	.99
4	1.79	.93
Random Testlet 3		
0	-1.82	.92
1	-.73	.90
2	.40	1.01
3	1.49	.94
4	2.35	1.22
Original Testlet 4		
0	-1.34	.88
1	-.69	.83
2	.16	.84
3	.98	.83
4	1.71	.99
Random Testlet 4		
0	-2.03	.95
1	-1.22	.64
2	.07	.75
3	1.15	.82
4	2.13	.95

Table 17. (cont'd)

Form 25 (n=1016)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.74	.86
1	-1.06	.91
2	-.49	.95
3	.15	93
4	.75	1.01
Random Testlet 1		
0	-1.75	.87
1	-1.09	.82
2	-.31	.84
3	.55	.80
4	1.61	.87
Original Testlet 2		
0	-1.26	.88
1	-.60	.99
2	-.07	.98
3	.61	.89
4	1.41	.96
Random Testlet 2		
0	-1.37	1.07
1	-.88	1.00
2	-.10	.94
3	.79	.88
4	1.67	1.10
Original Testlet 3		
0	-.98	1.02
1	-.51	.89
2	.12	.83
3	.70	.97
4	1.74	.82
Random Testlet 3		
0	-1.28	1.13
1	-.76	.94
2	-.09	1.05
3	.80	.87
4	1.52	1.09
Original Testlet 4		
0	-.94	1.15
1	-.41	1.05
2	.10	1.05
3	.69	1.19
4	1.55	1.18
Random Testlet 4		
0	-.95	1.11
1	-.39	1.04
2	.33	1.08
3	1.26	.93
4	2.19	1.01

Table 17. (cont'd)

Form 26 (n=896)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.30	1.08
1	-.78	.92
2	-.09	.94
3	.60	.98
4	1.60	.86
Random Testlet 1		
0	-1.60	.77
1	-.76	.88
2	.05	.86
3	.98	.81
4	2.14	.89
Original Testlet 2		
0	-1.74	.89
1	-1.09	.76
2	.04	1.25
3	.45	1.22
4	1.46	1.18
Random Testlet 2		
0	-1.54	.73
1	-.64	1.03
2	.05	.85
3	.99	.98
4	1.91	1.00
Original Testlet 3		
0	-1.13	1.00
1	-.57	.81
2	.13	.93
3	1.01	.73
4	1.68	.94
Random Testlet 3		
0	-.56	1.34
1	-.11	1.05
2	.78	1.13
3	1.68	1.22
4	2.33	1.35
Original Testlet 4		
0	-1.18	1.09
1	-.54	.94
2	.21	1.01
3	1.11	.90
4	1.86	1.50
Random Testlet 4		
0	-1.16	.92
1	-.42	.93
2	.30	.82
3	1.25	.88
4	2.44	.74

Table 17. (cont'd)

Form 27 (n=945)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.40	1.01
1	-.74	.94
2	-.04	.92
3	.68	1.10
4	1.48	1.04
Random Testlet 1		
0	-1.36	.92
1	-.63	.79
2	.36	.86
3	1.26	.96
4	2.38	.80
Original Testlet 2		
0	-1.01	1.06
1	-.29	1.12
2	.31	1.22
3	.99	1.27
4	2.19	.88
Random Testlet 2		
0	-1.50	1.13
1	-1.11	.86
2	-.20	1.06
3	.88	1.01
4	1.58	1.29
Original Testlet 3		
0	-1.77	.81
1	-1.23	.77
2	-.69	.79
3	.06	.80
4	.91	.77
Random Testlet 3		
0	-1.21	1.05
1	-.53	1.00
2	.27	1.04
3	1.19	.94
4	2.13	.96
Original Testlet 4		
0	-1.35	.98
1	-.81	.86
2	.01	.91
3	.84	.84
4	1.35	1.54
Random Testlet 4		
0	-1.30	.88
1	-.39	.95
2	.44	.78
3	1.54	.90
4	2.49	.88

Table 17. (cont'd)

Form 28 (n=944)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.55	.85
1	-1.05	.89
2	-.61	.87
3	.08	.78
4	.86	.83
Random Testlet 1		
0	-2.14	.67
1	-1.21	.71
2	-.28	.80
3	.38	.84
4	1.40	.85
Original Testlet 2		
0	-1.24	1.05
1	-.56	1.28
2	-.02	1.11
3	.47	1.41
4	1.81	1.12
Random Testlet 2		
0	-1.32	1.15
1	-.45	1.12
2	.14	1.22
3	1.10	1.07
4	1.80	1.28
Original Testlet 3		
0	-1.35	1.03
1	-.95	.84
2	-.56	.96
3	.23	.74
4	.94	.83
Random Testlet 3		
0	-1.77	1.10
1	-.98	.99
2	-.24	1.06
3	.45	1.13
4	1.39	.98
Original Testlet 4		
0	-1.22	1.05
1	-.81	1.01
2	-.15	.96
3	.42	1.14
4	1.42	1.08
Random Testlet 4		
0	-1.64	.94
1	-.71	.96
2	.02	.94
3	.89	.88
4	1.90	.94

Table 17. (cont'd)

Form 29 (n=947)

Category Label	Average Measure	Infit MNSQ
Original Testlet 1		
0	-1.48	.97
1	-.97	.75
2	-.10	.96
3	.57	.93
4	1.54	.94
Random Testlet 1		
0	-2.20	.75
1	-1.32	.80
2	-.44	.86
3	.45	.78
4	1.45	.80
Original Testlet 2		
0	-1.31	.90
1	-.64	.86
2	.23	.86
3	1.03	.87
4	2.12	.83
Random Testlet 2		
0	-1.69	1.04
1	-.99	.89
2	.11	.97
3	.91	1.05
4	1.39	1.45
Original Testlet 3		
0	-1.01	1.13
1	-.17	1.31
2	.41	1.23
3	1.19	1.27
4	1.92	1.25
Random Testlet 3		
0	-1.49	.99
1	-.40	1.08
2	.42	1.15
3	1.26	1.24
4	2.27	1.00
Original Testlet 4		
0	-1.22	.88
1	-.45	.94
2	.18	1.00
3	1.03	.83
4	1.94	.91
Random Testlet 4		
0	-1.43	.91
1	-.66	.83
2	.17	.86
3	1.09	.80
4	1.92	.76

Table 18. Ranges for Average Measures for Original and Random Testlets

Tryout Form	Testlet Composition	Avg. Measure Range
Form 20	Original Testlet 1	2.70
	Random Testlet 1	4.14
	Original Testlet 2	2.54
	Random Testlet 2	3.70
	Original Testlet 3	2.60
	Random Testlet 3	3.70
	Original Testlet 4	2.69
	Random Testlet 4	4.19
Form 21	Original Testlet 1	2.70
	Random Testlet 1	3.01
	Original Testlet 2	2.76
	Random Testlet 2	2.50
	Original Testlet 3	2.33
	Random Testlet 3	2.85
	Original Testlet 4	2.59
	Random Testlet 4	3.24
Form 22	Original Testlet 1	3.17
	Random Testlet 1	3.27
	Original Testlet 2	3.11
	Random Testlet 2	3.15
	Original Testlet 3	2.75
	Random Testlet 3	2.57
	Original Testlet 4	3.19
	Random Testlet 4	3.40
Form 23	Original Testlet 1	2.56
	Random Testlet 1	3.78
	Original Testlet 2	2.93
	Random Testlet 2	3.16
	Original Testlet 3	2.33
	Random Testlet 3	3.51
	Original Testlet 4	2.68
	Random Testlet 4	3.76

Table 18. (cont'd)

Tryout Form	Testlet Composition	Avg. Measure Range
Form 24	Original Testlet 1	3.29
	Random Testlet 1	4.29
	Original Testlet 2	3.55
	Random Testlet 2	4.05
	Original Testlet 3	3.48
	Random Testlet 3	4.17
	Original Testlet 4	3.05
	Random Testlet 4	4.16
Form 25	Original Testlet 1	2.49
	Random Testlet 1	3.35
	Original Testlet 2	2.67
	Random Testlet 2	2.93
	Original Testlet 3	2.72
	Random Testlet 3	3.10
	Original Testlet 4	2.49
	Random Testlet 4	3.80
Form 26	Original Testlet 1	2.90
	Random Testlet 1	3.74
	Original Testlet 2	3.20
	Random Testlet 2	3.45
	Original Testlet 3	2.81
	Random Testlet 3	2.89
	Original Testlet 4	3.04
	Random Testlet 4	3.60

Table 18. (cont'd)

Tryout Form	Testlet Composition	Avg. Measure Range
Form 27	Original Testlet 1	2.88
	Random Testlet 1	3.74
	Original Testlet 2	3.20
	Random Testlet 2	3.08
	Original Testlet 3	2.68
	Random Testlet 3	3.34
	Original Testlet 4	2.70
	Random Testlet 4	3.79
Form 28	Original Testlet 1	2.41
	Random Testlet 1	3.54
	Original Testlet 2	3.05
	Random Testlet 2	3.12
	Original Testlet 3	2.29
	Random Testlet 3	3.16
	Original Testlet 4	2.64
	Random Testlet 4	3.54
Form 29	Original Testlet 1	3.02
	Random Testlet 1	3.65
	Original Testlet 2	3.43
	Random Testlet 2	3.08
	Original Testlet 3	2.93
	Random Testlet 3	3.76
	Original Testlet 4	3.16
	Random Testlet 4	3.35

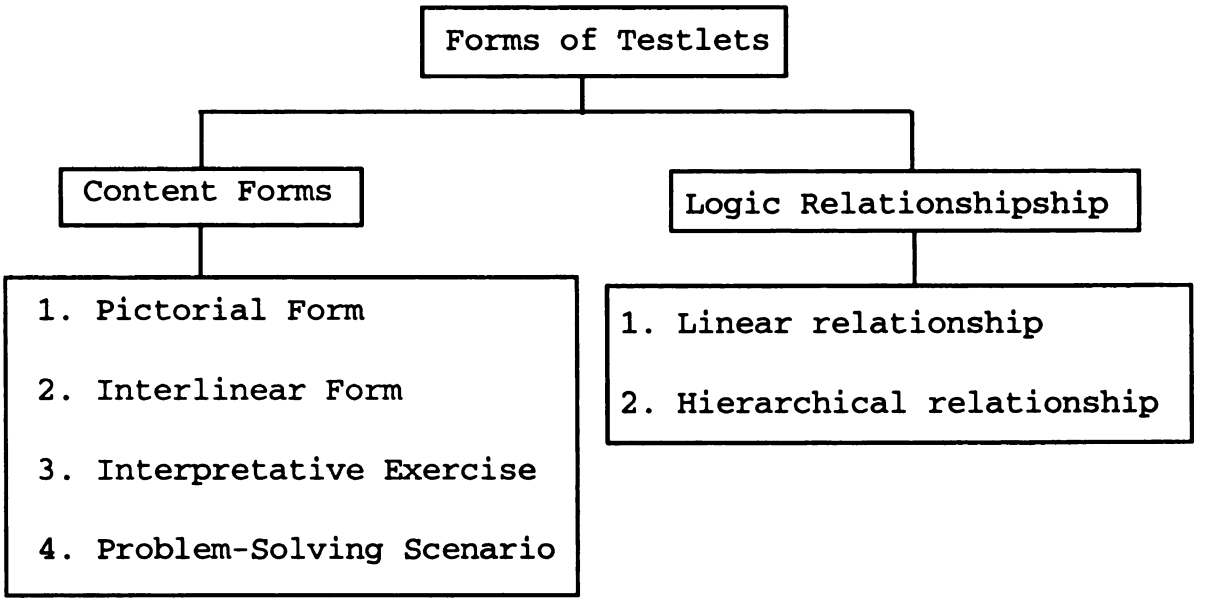


Figure 1. Classification of Testlets

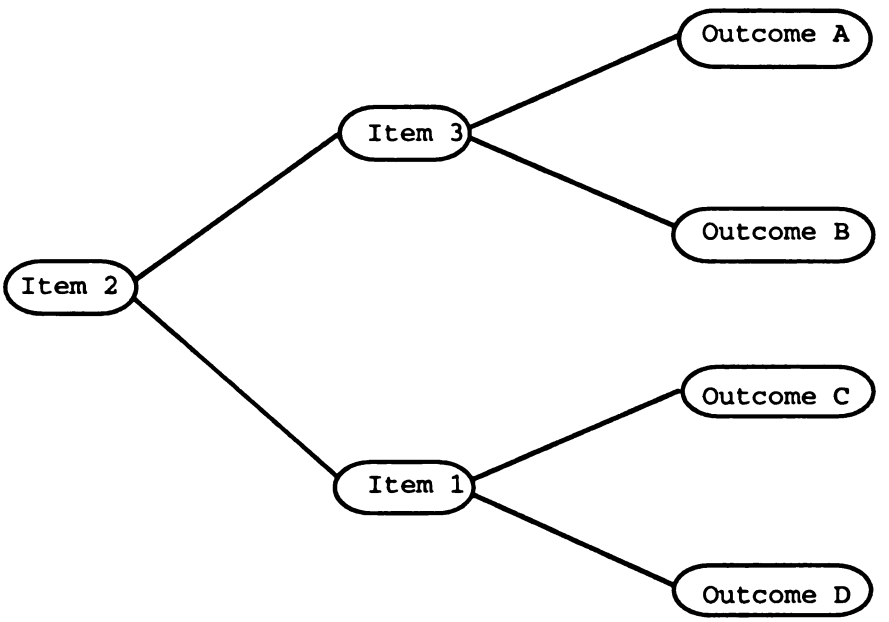


Figure 2. An example of two-level, 3-item, 4-outcome hierarchical testlet

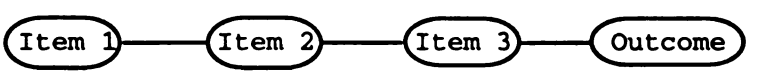


Figure 3. An exampleof three-level, 3-item linear testlet

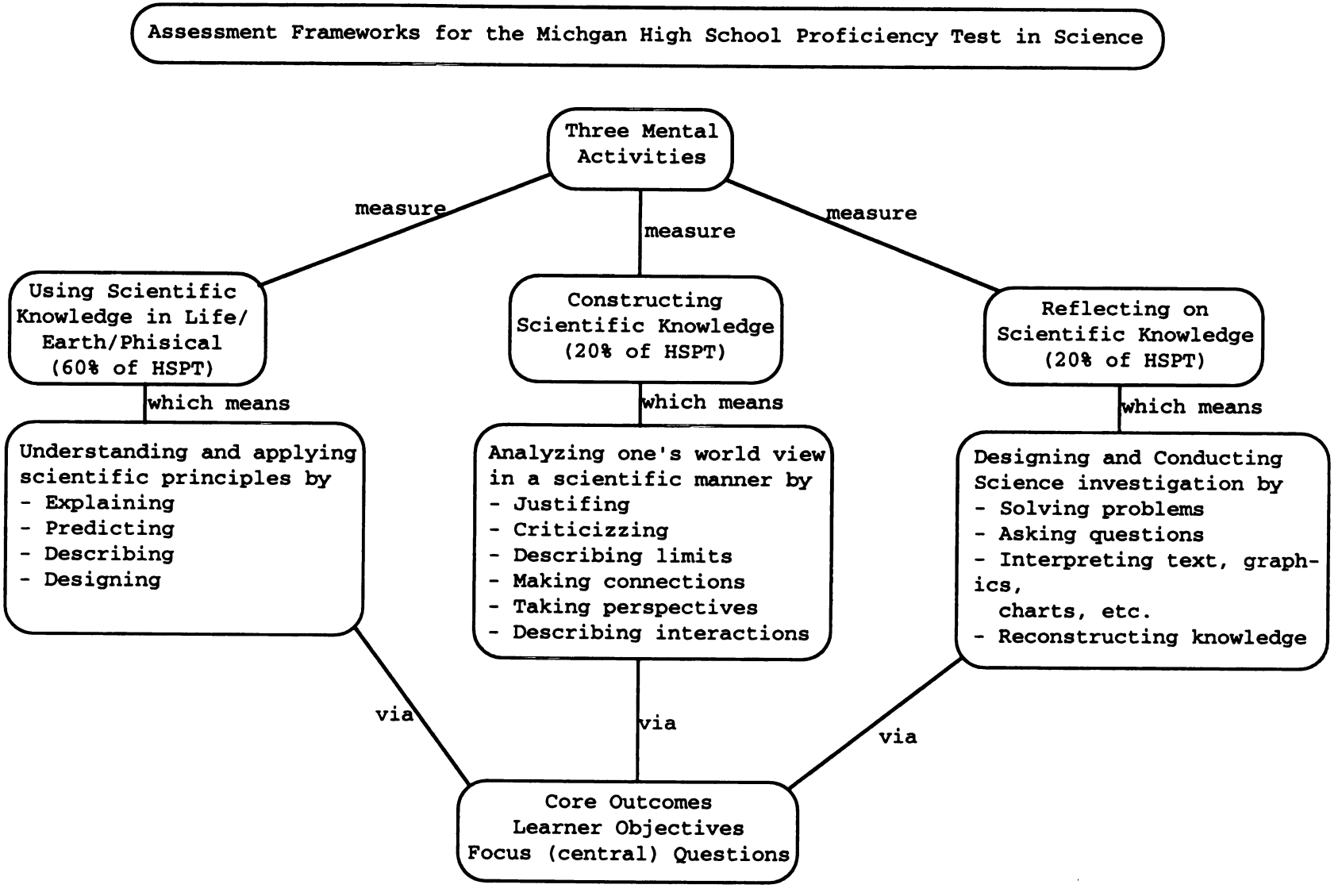


Figure 4. Assessment Frameworks for the Michgan High School Proficiency Test in Science

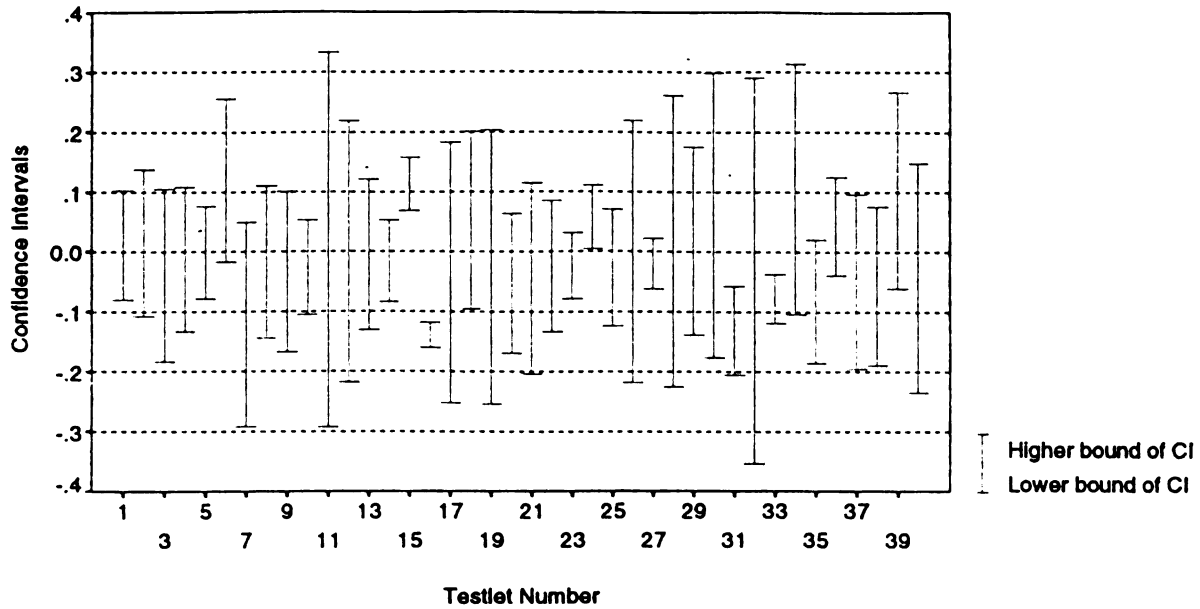


Fig. 5. CIs of ln(infit MNSQ) for Original Testlets

(Note: The first 4 testlets are from Form 20, and so on.)

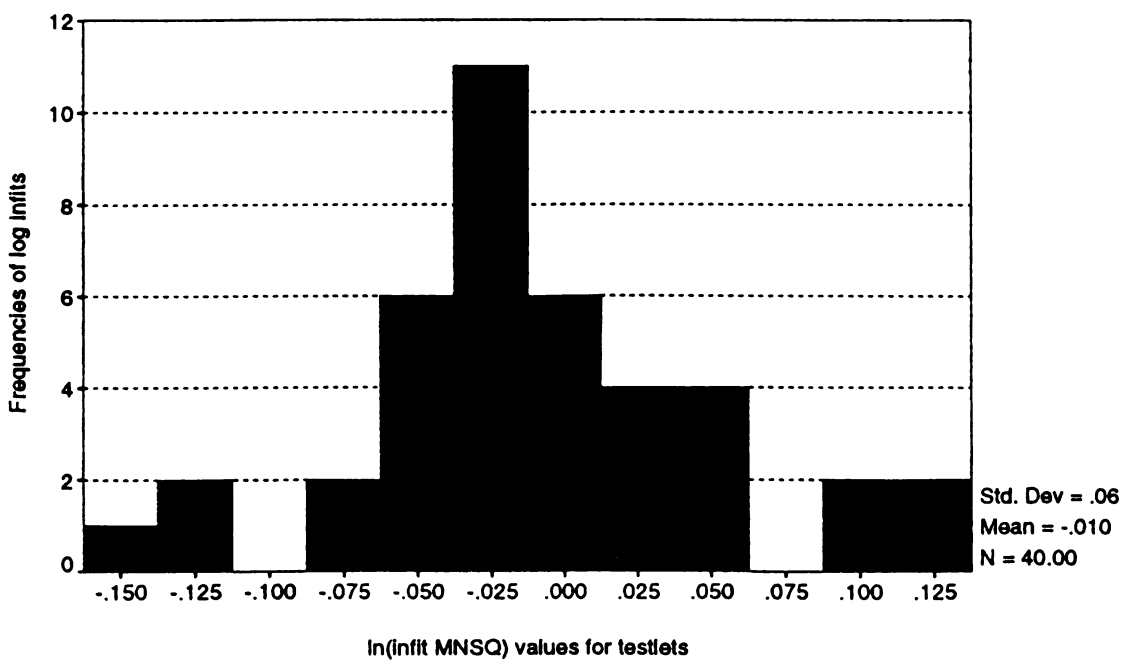


Fig. 6. Frequency Distribution of ln(infit MNSQ) for Original Testlets

LIST OF REFERENCES

LIST OF REFERENCES

Anastasi, A. (1961). Psychological testing (2d ed.). New York: Macmillan.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Biggs, J. B. & Collis, K. F. (1982). Evaluating the quality of learning: The SOLO taxonomy (structure of observed learning outcomes). New York: Academy Press.

Cattell, R., & Burdsal, C., Jr. (1975). The radial parcel double faCtoring design: A solution to the item-vs.-parcel controversy. Multivariate Behavioral Research, 10, 165-179.

Collis, K. F., Romberg, T. A., & Jurdak, M. E. (1986). A technique for assessing mathematical problem solving ability. Journal of Research in Mathematics Education, 17, 206-211.

Cureton, E. E. (1965). Reliability and validity: Basic assumptions and experimental designs. Educational and Psychological Measurement, 25(2), 327-346.

CTB Macmillan/McGraw-Hill. (1989). Comprehensive Tests of Basic Skills (4th ed., Technical Bulletin No. 1). Monterey, CA: Author.

De Ayala, R. (1991, April). The Influence of Dimensionality on Estimation in the Partial Credit Model. Paper presented at the International Objective Measurement Workshop.

De Ayala, R., Dodd, B., & Kock, W. R. (1988). A Comparison of the Nominal and Graded Response Models in Computerized Testing. Paper presented at American Educational Research Association, New Orleans, LA.

Donohue, J. R. (1993). An Empirical Examination of the IRT Information in Polytomously Scored Reading Items. (Research Report-93-12). Priceton, NJ: Educational Testing Services.

Ebel, R. L. (1951). Writing the test item. In E.F. Lindquist (Ed.), Educational Measurement (1st ed., pp. 185-249). Washington, DC: American Council on Education.

Educational Testing Services. (1989). GRE Educational Test Sample Test. Princeton, NJ: Author.

Engelhart, M. D. (1942). Unique types of achievement test exercises. Psychometrika, 7(2), 103-116.

Ercikan, K. (1993). Measurement Accuracy in Testlet Methodology. Paper presented at the National Council on Measurement in Education. Atlanta, GA.

Gerberich, J. R. (1956). Specimen Objective Test Items. New York: Longmans, Green and Co.

Gronlund, N. E. (1965). Measurement and Evaluation in Testing (5th ed.). New York: Macmillan.

Guilford, J. P. (1936). Psychometric Methods (1st ed.). New York: McGraw-Hill.

Haladyna, T. M. (1991). Generic questioning strategies in the teaching of statistics. Educational Technology: Research and Development, 39(1), 73-82.

Haladyna, T. M. (1992). Context-dependent item sets. Educational Measurement: Issues and Problems, 11, 21-25.

Huynh, H. (1994). On equivalence between a partial credit items and a set of independent Rasch binary items. Psychometrika, 59(1) 111-119.

Joreskog, K. G. & Sorbom, D. (1989). LISREL 7: A Guide to the Program and Application (2d ed.). Chicago: SPSS Inc.

Lewis, C. (1989). Validity-based scoring. Unpublished manuscript.

Lewis, C. & Sheehan, K. (1988). Using Bayesian decision theory to design a computerized mastery test. Unpublished manuscript.

Linacre, J. M. & Wright, B. D. (1995) BIGSTEPS, version 2.6 [computer software]. Chicago, IL:MESA Press.

Linacre, J. M. & Wright, B. D. (1995). BIGSTEPS, the User's Guide. Chicago,IL: MESA Press.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

Mehrens, A. W., & Lehmann, I. J. (1984). Measurement and Evaluation in Education and Psychology (4th ed.). New York: Holt, Rinehart and Winston, Inc.

Michigan State Board of Education (1991). Michigan Essential Goals and Objectives for Science Education (K-12). Lansing, MI: Author.

Michigan State Board of Education (1991). Michigan Core Curriculum Outcomes. Lansing, MI: Author.

Michigan State Board of Education (1994). Assessment Frameworks for the Michigan High School Proficiency Test in Science. Lansing, MI: Author.

Rasch, G. (1980). Probabilitic Models for Some Intelligence and Attainment Tests. Chicago, IL: University of Chicago Press. (Original work published by Copenhagan: Danmarks Paedogogiske Institut, 1960).

Rosenbaum, P. R. (1988). A note on item bundles. Psychometrika, 53, 349-359.

Sireci, S., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237-247.

Szeberënyi, J. & Tigyi, A. (1987). The use of application test, a novel type of problem-solving exercise, as a tool of teaching and assessment of competence in medical biology. Medical Teacher, 9(1), 73-82.

Thissen, D., & Steinberg, L. (1988). Data analysis using Item Response Theory. Psychological Bulletin, 104, 385-395.

Thissen, D., Steinberg, L. & Fitzpatrick A. (1989). Multiple-choice models: The distractors are also part of the item. Journal of Educational Measurement, 26, 161-176.

Thissen, D., Steinberg, L. & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical-response models. Journal of Educational Measurement, 26, 247-260.

van de Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. Psychometrika, 47, 123-140.

Wainer, H., & Kiely, G. (1987). Item clusters and computer adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of simulated hierarchical and linear testlets. Journal of Educational Measurement, 29, 243-251.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. Journal of Educational Measurement, 28, 311-324.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. Applied Psychological Measurement, 12(4), 353-364.

Wilson, M. & Iventosch, L. (1988). Using the partial credit model to investigate responses to structured subtests. Applied Measurement in Education, I(4), 319-334.

Wright, B. D. (1992). IRT in the 1990s: Which models work best? Invited debate at the AERA Annual Meeting. 1992.

Wright, B. D. (1996). Reliability and separation. Rasch Measurement Transactions, 9:4, 472.

Wright, B. D. & Masters, G. N. (1982). Rating Scale Analysis. Chicago,IL: MESA Press.

Yen, W. M. (1984a). Effect of local item dependence on the fit and equating performance of the three parameter logistic model. Applied Psychological Measurement, 8(2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for Managing local item dependence. Journal of Educational Measurement, 30(3), 187-213.