PREDICTING THE DISTRIBUTION OF A GOODNESS-OF-FIT STATISTIC APPROPRIATE FOR USE WITH PERFORMANCE-BASED ASSESSMENTS

by

Mary A. Hansen

B.S., Mathematics and Computer Science, California University of PA, 1994

M.A., Statistics, University of Pittsburgh, 1996

M.S., Research Methodology, University of Pittsburgh, 1999

Submitted to the Graduate Faculty of

the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2004

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Mary A. Hansen

It was defended on July 14, 2004

and approved by

Clement A. Stone, Ph.D., School of Education

Suzanne Lane, Ph.D., School of Education

Carol E. Baker, Ph.D., School of Education

James J. Irrgang, Ph.D., School of Health and Rehabilitation Sciences

Dissertation Director: Clement A. Stone, Ph.D.

PREDICTING THE DISTRIBUTION OF A GOODNESS-OF-FIT STATISTIC APPROPRIATE FOR USE WITH PERFORMANCE-BASED ASSESSMENTS

Mary A. Hansen, PhD

University of Pittsburgh, 2004

One aspect of evaluating model-data fit in the context of Item Response Theory involves assessing item fit using chi-square goodness-of-fit tests. In the current study, a goodness-of-fit statistic appropriate for assessing item fit on performance-based assessments was investigated.

The statistic utilized a pseudo-observed score distribution, that used examinees' entire posterior distributions of ability to form item fit tables. Due to dependencies in the pseudo-observed score distribution, or pseudocounts, the statistic could not be tested for significance using a theoretical chi-square distribution. However, past research suggested that the Pearson and likelihood ratio forms of the pseudocounts-based statistic (χ^{2*} and G^{2*}) may follow scaled chi-square distributions.

The purpose of this study was to determine whether item and sample characteristics could be used to predict the scaling corrections needed to rescale χ^{2*} and G^{2*} statistics, so that significance tests against theoretical chi-square distributions were possible. Test length (12, 24, and 36 items) and number of item score category levels (2 to 5-category items) were manipulated. Sampling distributions of χ^{2*} and G^{2*} statistics were generated, and scaling corrections obtained using the method of moments were applied to the simulated distributions.

Two multilevel equations for predicting the scaling corrections (a scaling factor and degrees of freedom value for each item) were then estimated from the simulated data.

Overall, when scaling corrections were obtained with the method of moments, sampling distributions of rescaled χ^{2*} and G^{2*} statistics closely approximated theoretical chi-square distributions across test configurations.

Scaling corrections obtained using multilevel prediction equations did not adequately rescale simulated χ^{2*} distributions for 2- to 5-category tests, or simulated G^{2*} distributions for 2- and 3- category tests. Applications to real items showed that the prediction equations were inadequate across score category levels when χ^{2*} was used, and for 2- and 3-category items when G^{2*} was used.

However, for 4- and 5-category tests, the predicted scaling corrections did adequately rescale empirical sampling distributions of G^{2*} statistics. In addition, applications to real items indicated that use of the multilevel prediction equations with G^{2*} would result in correct identification of item misfit for 5-category, and potentially 4-category items.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to extend my thanks and appreciation to my dissertation advisor, Dr. Clement A. Stone, for his support and guidance on this project, and throughout my graduate studies. I am also grateful to my dissertation committee, Dr. Suzanne Lane, Dr. Carol E. Baker, and Dr. James J. Irrgang, for their valuable input and suggestions, which enhanced the quality of this study.

Without the help of several family members, the completion of this dissertation would not have been possible. Special thanks to my parents, Marilyn and Charles Doerzbacher, for their unconditional and unending support and love. To my beautiful daughters, Jacqueline and Sarah, thank you for the laughter and smiles you bring to my life each day. And to my husband Michael, thank you for your patience, confidence, and endless encouragement throughout this undertaking.

CHAPTER I

I. INTRODUCTION

I. A. Assessing Item Fit for Large Scale Performance-Based Assessments

I. A. 1. Large Scale Performance-Based Assessments

In recent years, there has been a shift at national, state, district, and school levels from selected-response only tests to tests that contain performance-based items. Performance-based assessments are believed to support the teaching and learning of problem solving and critical thinking skills, increase teacher and student motivation, and improve the alignment between curriculum and instruction (Khattri, Reeve, & Kane, 1998). Due to these attributes, these types of assessments continue to gain popularity.

At the national level, advanced placement exams that consist of performance-based items provide students with a chance to earn college credits. In addition, the National Assessment of Educational Progress (NAEP) provides a continuing assessment of the knowledge of students in grades 4, 8, and 12. NAEP utilizes a matrix-sampling design to administer blocks of items that include performance-based items.

At the state level, performance-based assessment systems are used to monitor student performance as well as hold schools and teachers accountable for student achievement. A

number of states implement testing programs that include both selected-response and performance-based item types (e.g., Florida, Pennsylvania, and Massachusetts). Others have consisted primarily or entirely of performance-based items (e.g., Maryland). The items on these high-stakes assessments can be evaluated using item response theory.

I. A. 2. Item Response Theory

Item response theory (IRT) is a powerful system of mathematical models that allows for the analysis of item response data. IRT has many applications in large scale testing, including test development, equating, and scoring. IRT models postulate a relationship between an individual's response to an item and the underlying (latent) ability measured by a test. The models use item and person characteristics in predicting the probability that an examinee provides a specific response to an item. IRT models and techniques are appropriate for use with tests consisting of selected-response and/or performance-based items.

IRT offers three main advantages over other testing theories when the model assumptions are met: (a) ability estimates are independent of the sample of items on which they are based; (b) item parameters are independent of the sample of examinees from which they were obtained; and (c) examinees at different ability levels have different standard errors of measurement that can be estimated (Hambleton, 1993; Hambleton & Swaminathan, 1985).

The first two advantages of invariant ability and item parameters, respectively, are central to IRT applications. Invariant ability parameter estimates allow examinees taking different sets of test items to be compared. This is because under IRT, examinees will obtain the same ability estimate (apart from measurement error) across different sets of items. In addition, invariant item parameter estimates ensure that the item descriptors do not depend on or vary across

different sets of examinees, apart from measurement error. Further, with IRT the precision of ability estimates at all points along the ability continuum is known.

In order for these advantages of IRT to be attained, certain assumptions relating to the specific model being implemented must be met. A number of assumptions underlie the basic IRT models. The assumptions are: (a) one or more ability dimensions underlie performance; (b) examinees' responses to the items on a test are independent; (c) the test is non-speeded; and (d) the form and shape of the model as defined by the item parameters is appropriate (Hambleton, 1993; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). The model assumptions must be met in order for the benefits of IRT to be attained

Generally, performance-based assessments consist of polytomous items. These items result in examinees receiving one of several score points (e.g., 0 – 5) for each item. Because performance-based assessments are the focus of this paper, concentration will be placed on IRT models appropriate for polytomous data. One particular IRT model, Samejima's (1969) Graded Response Model (GRM), will be highlighted. This model requires the additional assumption that the score categories of the items are ordered, and is appropriate for the applications to be described herein.

The extent to which the advantages of IRT are attained in a particular application depends on the extent to which the IRT model being implemented is appropriate. Model-data fit must be evaluated in order to determine whether the benefits of IRT, as well as the benefits of the particular assessment, can be attained.

I. A. 3. Model-Data Fit in the Context of Item Response Theory

The assessment of model-data fit is necessary whenever data are analyzed by mathematical models. In order to assess model-data fit in the context of IRT, one must

investigate several potential sources of misfit. Model-data misfit can be the result of factors such as (a) failure to meet model assumptions, (b) failure to attain invariant item and ability parameter estimates, (c) failure to select an appropriate item response model, or (d) small sample size (Hambleton, 1993; McKinley & Mills, 1985). Validating the use of an IRT model requires examination of each of these potential sources of misfit.

The assessment of item fit, or the match between examinee responses to an item and the IRT model predictions, must therefore be investigated as part of the model validation process. The assessment of item fit is important for a number of reasons. Item misfit indicates that the item parameters do not accurately represent examinee responses to test items (Reise, 1990). These inaccuracies render the validity of test scores questionable. When decisions concerning students or teachers are based on test results, such inaccuracies can have serious consequences. Item fit must be evaluated so that the validity of IRT applications is not compromised.

There are a number of well-known methods of assessing item fit, including graphical representations and statistical tests. Pictorial displays of the relationship between the observed and expected score distributions, and analysis of residuals (Hambleton, 1993) are two graphical methods. Most statistical methods that have been introduced to detect item misfit are based on the calculation of a chi-square test statistic. In general, these goodness-of-fit methods involve the comparison of an observed to expected, or model-predicted, score distribution, and test for statistical significance using the chi-square distribution.

Chi-square goodness-of-fit statistics in the context of IRT are formed by first rank-ordering examinees according to the point estimate of their ability, and then dividing them into some number of subgroups based on this ranking. Then, within each of the subgroups, the observed and expected score distributions are computed.

The observed score distribution is formed by tabulating the number of examinees in each ability subgroup that responded to each score category of the item. The expected score distribution is formed by using an IRT model to predict for each subgroup the number of examinees who should fall into each of the score categories, if the IRT model fits the item response data. This prediction is made using estimated item parameters, and a single summary measure of ability (i.e., mean or median) for each of the subgroups. The observed and expected score distributions from the item fit table are then compared through the computation of a chi-square test statistic, and tested for significance using the chi-square distribution with the appropriate degrees of freedom.

Since examinees are grouped according to their estimated abilities, it follows that the ability estimate plays a significant role in the computation of the observed and expected score distributions, and in the resulting value of the chi-square test statistic. A precise estimate of ability is therefore desirable.

It is well known that the precision of ability estimates is influenced by test length. As test length increases, the accuracy of ability estimates increases, and their variability decreases. For performance-based assessments, the number of items on the assessments is often less than the number of items found on comparable selected-response item tests. As a consequence of having shorter test lengths, individual examinee ability estimates for performance-based assessments and assessments that utilize matrix-sampling techniques may lack precision (cf., Ankenmann & Stone, 1992).

This imprecision in the estimation of ability can result in classification errors in the item fit tables. Classification errors occur when examinees are assigned to incorrect ability subgroups in the item fit table. Mote and Anderson (1965) caution that classification errors are common in

practical applications of chi-square testing theory, and can result in increased Type I error rates and decreased power for significance tests of fit.

The chi-square based goodness-of-fit statistics of Bock (1972), Yen (1981), and McKinley and Mills (1985) are commonly used to assess item fit. These 'traditional' fit statistics do not take into account the precision of the ability estimates. Research has shown that the extent to which classification errors impact chi-square fit statistics is related to the precision of ability estimates (Ankenmann, 1994; Stone, 2000; Stone, Ankenmann, Lane, & Liu, 1993; Stone & Hansen, 2000; Stone, Mislevy, & Mazzeo, 1994). Classification errors pose a greater risk for shorter tests such as performance-based assessments that provide less precise ability parameter estimates.

As an example, Stone and Hansen (2000) conducted a simulation study focused on performance-based assessments. Their study investigated the effect of precision in ability estimation (represented by test length) on the distribution of goodness-of-fit statistics for IRT models. The researchers found that using traditional chi-square fit statistics in the presence of imprecise ability estimates affected the means and variances of the sampling distributions of the statistics, and resulted in higher than nominal Type I error rates in the assessment of fit. These researchers concluded that imprecision in ability estimation should be considered in the assessment of item fit for shorter performance-based tests.

I. A. 4. A Fit Statistic that Accounts for Uncertainty in Ability Estimates

Stone (2000) described a method for assessing item fit that accounts for imprecision in ability estimation by utilizing examinees' posterior distributions of ability. Using Bayesian estimation techniques, a distribution of posterior expectations across the ability scale is obtained for each examinee. Stone uses these posterior expectations to create "pseudocounts", or a

pseudo-observed score distribution, that classifies each examinee into several cells of the item fit table based on their entire distribution of posterior expectations of ability (Stone, 2000; Stone et al., 1994). These pseudocounts serve as a discrete representation of the posterior ability distributions of examinees (Ankenmann, 1994). A chi-square fit statistic is then formed using the pseudocounts, rather than actual counts, as the observed frequencies.

In contrast to the traditional statistics, this new test statistic based on a pseudo-observed score distribution cannot be compared to a theoretical chi-square distribution with known degrees of freedom for purposes of significance testing. This is because the assumption of independence associated with chi-square tests is not met due to dependencies in the posterior expectations utilized by the pseudocounts-based procedure. However, studies investigating the utility of this procedure suggested that the fit statistic may follow a scaled chi-square distribution (Ankenmann, 1994; Stone, 2000; Stone et al., 1993; Stone et al., 1994).

Stone et al. (1993) computed this pseudocounts-based fit statistic in order to statistically assess the fit of items on a performance-based assessment. The researchers used Monte Carlo methods to generate empirical sampling distributions of the fit statistics for each item, and obtained the critical values needed for significance testing from the empirical distributions.

The authors examined the empirical sampling distributions of the pseudocounts-based fit statistic, and reported the following. First, Stone et al. (1993) found evidence that the fit statistics were distributed as scaled chi-square random variables. This suggested that one of the family of theoretical chi-square distributions could be used for significance testing of the fit statistics, provided that the appropriate scaling corrections (a scaling factor and degrees of freedom value) for each item could be found.

Further, Stone et al. (1993) found evidence suggesting that the scaling corrections could potentially be estimated (or predicted) from characteristics of the sample data. If prediction of the scaling corrections from sample data were possible, then significance testing of the observed fit statistics could be carried out without the empirical generation of sampling distributions.

Stone et al. (1994) and Ankenmann (1994) further investigated the sampling distributions of the pseudocounts-based fit statistics, and again found the statistics to be distributed as scaled chi-square random variables. They attempted to predict the scaling corrections for the fit statistics from characteristics of the sample data. For these studies, if the scaling corrections could be predicted from item and/or sample characteristics, then empirically generated sampling distributions would not be needed for significance testing of the fit statistics. Instead, practitioners could use the researchers' prediction equations to obtain the appropriate scaling corrections for their observed fit statistics, and use a known chi-square distribution for significance testing. Specifically, the observed fit statistic for an item could be rescaled (multiplied or divided) by the predicted scaling factor, and compared to the chi-square distribution with predicted degrees of freedom.

The scaling corrections can be predicted reasonably well from characteristics of the sample data for the Rasch model (Stone et al., 1994) and GRM with 5 item response categories (Ankenmann, 1994). With respect to the Rasch model, Stone et al. (1994) found that test level scaling corrections could be predicted fairly well using the average posterior variance of examinee ability as the predictor variable.

Ankenmann (1994) found that for applications to the 5-category case of the GRM, item characteristics such as item discrimination and item information, along with the mean posterior variance, were needed to determine the scaling corrections. Ankenmann found that item level

prediction equations could be derived from simulated data and applied to real items, so that observed fit statistics could be tested for fit without the generation of sampling distributions. He found that matching the items on discrimination parameter and/or item information was necessary in order to find the most appropriate prediction equations.

While these results are promising, Ankenmann's (1994) study was limited with respect to the number and variety of items represented. In his study, only 4 performance-based items, each with 5 score response categories, were investigated. In practice, assessments may consist of items with varying numbers of score categories (e.g., NAEP). The effect of varying the number of score categories on the distribution of the pseudocounts-based fit statistic and on the prediction of the scaling corrections from characteristics in the sample data is unknown. This must be investigated in evaluating the utility of the fit statistic and prediction method. The current study was designed to provide more general equations for predicting scaling corrections than those found by Stone et al. (1994) and Ankenmann (1994). This may be accomplished by including items that vary in the numbers of score categories and have a variety of item parameters.

I. B. Statement of the Problem

The purpose of this study was to determine whether item and sample characteristics could be used to predict the scaling corrections needed for significance testing of a fit statistic that considers the imprecision in ability estimation found on performance-based assessments. The main goal was to provide a set of two prediction equations that would predict the scaling corrections for real items having different item parameters from the items included in this study.

The following research questions were under study:

1. For items modeled by the GRM having 2, 3, 4, and 5 response categories, can scaling corrections obtained from Monte Carlo based empirical sampling distributions be used to rescale simulated fit statistic distributions so that significance tests against a known chi-square distribution can be performed?
2. For items modeled by the GRM having 2, 3, 4, and 5 response categories, can appropriate scaling corrections be predicted from item and sample characteristics and used to rescale simulated fit statistic distributions so that significance tests against a known chi-square distribution can be performed?
3. Can the prediction equations derived from empirical data be generalized to real items having different parameters than those in the simulation study, for purposes of significance testing of the observed fit statistics?

I. C. Summary of the Study

In this study, data sets were simulated using realistic item parameters, a fit statistic based on examinees' posterior distributions of ability was computed for each item, and sampling distributions of each fit statistic were generated. The empirically generated sampling distributions were examined to verify that they followed scaled chi-square distributions. Prediction equations based on sample characteristics (e.g., mean posterior variance and root mean posterior variance) and item characteristics (e.g., item information, item discrimination, and number of score categories) were then derived. The goal of the prediction equations was to correctly predict the appropriate scaling corrections for each item from these sample/item characteristics. The scaling corrections obtained from the prediction equations were evaluated for each item. Finally, the prediction equations were applied to a new set of items to assess their utility.

Data for this study was simulated under the multidimensional GRM (MGRM) for conditions varying test length (12a, 12b, 12c, 24, and 36 items) and number of score categories (2, 3, 4, and 5). Multidimensional item response data having one main dimension and five minor or nuisance dimensions were generated to better reflect real item responses (Davey, Nering, & Thompson, 1997). It was assumed that the assumption of unidimensionality was met in an applied sense, meaning that in practice, the data would be considered essentially unidimensional.

Item discrimination and difficulty parameters were chosen to be representative of real performance-based items. Three 12-item tests at each of the score category levels were considered, so that 36 distinct items having each of 2, 3, 4 and 5 categories were included. The test lengths of 24 items repeated the items in tests 12a and 12b at each score category level. The 36-item tests contained the three 12-item tests at each score category level. Items with 2-5 score response categories were considered, because performance-based assessments and assessments such as NAEP consist of such items. In addition, data for examinee sample sizes of 2000 were generated for all experimental conditions.

I. D. Significance of the Study

Due to the increased use of performance-based items and continued use of the GRM for modeling these items, a need has arisen for a method to assess item fit in this setting. Methods of assessing item fit appropriate for use in many large-scale testing systems consisting of selected-response items are less appropriate for tests that consist only of performance-based items. This is because performance-based assessments often consist of small numbers of items, and yield only imprecise individual ability estimates. The particular method being investigated in the current study has yielded promising results (Ankenmann, 1994; Stone, 2000; Stone, 2003; Stone et al., 1993; Stone et al., 1994), but in a narrow range of conditions.

Item response data in this study were simulated under the MGRM with one main and five nuisance dimensions. This marks a difference from the study carried out by Ankenmann (1994) in which item response data were simulated under the unidimensional GRM. The prediction technique was found to be promising in that study, but under limited conditions. It is possible that differences between data simulated in that study and more realistic data may have caused the procedure to appear more effective than it actually would be when applied to additional real data sets (Davey et al., 1997). Also, the current study used items with a variety of score categories, because the prediction technique in question had only been investigated for items modeled by the GRM having 5 response categories.

Donoghue & Hombo (2001a, 2001b) conclude that the literature relating to the use of scaled chi-square distributions in assessing item fit using this fit statistic shows promising results, but caution against its use in operational NAEP analyses due to the time-intensive simulations that are required. The need for measures that accurately assess item fit for performance-based items without these simulations exists. This study plays a significant role in assessing whether the current method fills this need.

I. E. Limitations of the Study

As with any simulation study, the results of this study are somewhat limited with respect to the conditions defined within the study. The study is limited in terms of the item parameters, number of items, and IRT model that were used in the simulation procedures. However, the study did investigate the generalizability of the prediction equations to items with parameters different from those used in the simulation.

Another possible limitation in the study lies in the IRT model that was utilized for data generation. Multidimensional response data sets having one main and five nuisance dimensions

were generated. The purpose of generating response data under a multidimensional IRT model was to allow the data to reflect real item responses. However, the data were estimated using a unidimensional IRT model. This could be a limitation if another (multidimensional) IRT model would better describe the data.

An additional limitation to the study is that all results are based on empirical, rather than analytical, evidence. A recent attempt to solve the analytical solution to this problem has yielded theoretical results, but these results have not carried over into applied settings (Donoghue & Hombo, 2001b). Thus, there is a need for further empirical research. Furthermore, Harwell, Stone, Hsu, & Kirisci (1996) state that Monte Carlo studies should be considered if analytic solutions to problems cannot be found.

CHAPTER II

II. REVIEW OF LITERATURE

Item response theory (IRT) is commonly used in the development and scoring of large-scale assessments, including performance-based assessments. IRT offers several advantages that are attained when the item response model fits the data. In order to ensure model-data fit, one must examine several potential sources of model-data misfit, including a) the extent to which the IRT model assumptions are met, b) the extent to which the properties of item and ability invariance are obtained, and c) the accuracy of model predictions. To determine if the benefits of item response theory can be attained, each of these potential sources of misfit must be evaluated. As part of this evaluation, the fit of the model to the test data must be assessed. Most statistical methods that have been introduced to detect item misfit are based on the calculation of a chi-square test statistic. In general, these goodness-of-fit methods involve the comparison of an observed to expected, or model-predicted, score distribution using a chi-square goodness-of-fit statistic.

Several specific statistics for assessing item fit have been introduced and are summarized. Results from empirical studies that investigated their utility are presented, with specific attention being paid to the limitations of these goodness-of-fit statistics. This introduction to chi-square

based goodness-of-fit methods concentrates on their application to testing situations where ability is precisely estimated.

When dealing with performance-based assessments consisting of small numbers of items, imprecision in the estimation of ability becomes a factor in the applicability of the traditional goodness-of-fit statistics. Literature investigating the use of the traditional statistics with shorter tests is discussed. Limitations of the traditional statistics in this context are presented, and literature introducing a specific method of detecting item misfit that accounts for imprecision in ability estimation is discussed. The statistic discussed accounts for some of this imprecision by utilizing examinees' posterior distributions of ability. Studies that have investigated the use of this particular statistic are reviewed, providing a background for the current paper. Alternative methods that have been proposed for assessing item fit for polytomous IRT models are also discussed.

Several IRT models that are appropriate for modeling items on performance-based assessments exist. One particular model, the homogeneous case of Samejima's (1969) Graded Response Model (GRM) is discussed in detail, as it is the focus of the applications of this paper.

II. A. Assessing Model-Data Fit Using Chi-Square Goodness-of-Fit Statistics

In any model-based theory, the fit of the model to the sample data must be assessed. Decisions concerning the extent to which data obtained from a sample compare to what one would expect from a hypothesized distribution are made using goodness-of-fit tests. In the context of IRT, a number of goodness-of-fit tests for assessing item fit are based on the Pearson χ^2 statistic. In general, the Pearson χ^2 statistic is defined as

$$\chi^2 = \sum_{j-1}^{J} \frac{(O_j - E_j)^2}{E_j}, \tag{1}$$

where j is the category or cell,

J is the number of categories,

O_j is the observed count of responses in category j, and

E_j is the expected count of responses in category j.

When Equation 1 is expressed in terms of proportions rather than frequencies, the Pearson χ^2 statistic becomes

$$\chi^2 = \sum_{j=1}^{J} \frac{N_j (O_j - E_j)^2}{E_j (1 - E_j)}, \tag{2}$$

where j is the category or cell,

J is the number of categories,

N_j is the number of respondents in category j,

O_j is the observed proportion of responses in category j, and

E_j is the expected proportion of responses in category j.

In order to validly use Pearson χ^2 statistics in the assessment of model-data fit, a number of assumptions must be met by the data. These assumptions are that: (a) the sample is a simple random sample (SRS); (b) the observations are independent, meaning that every subject contributes to one cross-classification of the variables; and (c) the sample size is large (Howell, 1999; Yen, 1981).

Assumption (c), dealing with overall sample size, is necessary to ensure that the distribution of the Pearson statistic is approximated by the theoretical chi-square distribution. In addition, when the value of each cell expectation is greater than 5, the chi-square distribution will be a good approximation to the actual distribution of the test statistic; and the approximation should be satisfactory when each expected cell frequency is at least 1.5 (DeGroot, 1986). With respect to the size of the expected cell frequencies, Camilli and Hopkins (1978) found that Type

I error rates are not affected by expected cell frequencies as small as 1 or 2 in one or two cells of the contingency table, with a sample size of 20 or more. Thus, if the sample size is large and the expected value criteria are met for each cell, the chi-square distribution will be a good approximation to the actual distribution of the Pearson statistic.

Goodness-of-fit tests can also be based on the likelihood-ratio G^2 statistic, the general form of which is

$$G^2 = 2\sum_{j=1}^{J}\left\{\left(O_j\right)ln\left(\frac{O_j}{E_j}\right)\right\}, \tag{3}$$

where j is the category or cell,

J is the number of categories,

ln is the natural logarithm function,

O_j is the observed count of responses in category j, and

E_j is the expected count of responses in category j.

The likelihood-ratio statistic, presented in terms of proportions rather than frequencies, becomes

$$G^2 = 2\sum_{j=1}^{J}\left\{O_j\, ln\left(\frac{O_j}{E_j}\right)+\left(1-O_j\right)ln\left(\frac{1-O_j}{1-E_j}\right)\right\}, \tag{4}$$

where j is the category or cell,

J is the number of categories,

ln is the natural logarithm function,

O_j is the observed proportion of responses in category j, and

E_j is the expected proportion of responses in category j.

The assumptions associated with likelihood-ratio G^2 statistics are the same as those of the Pearson statistic, and the two statistics χ^2 and G^2 are asymptotically equivalent.

II. A. 1. Traditional Chi-Square Item Fit Statistics for IRT Models

In general, statistical techniques for assessing item fit involve comparing the observed and expected score distributions for some number of ability subgroups using either a Pearson or likelihood ratio chi-square statistic. This comparison involves the following steps:

1. Rank the examinees according to their estimated ability and separate them into some number of subgroups.
2. Construct the observed score distribution for a specific item by tabulating, within each of the subgroups, the number (or proportion) of examinees responding to each score category of that item.
3. Construct the expected score distribution for an item, again within each of the ability subgroups, by utilizing an IRT model to predict the number (or proportion) of examinees who should fall into each of the score categories. This prediction is made using estimated item parameters, and a single summary measure of ability (i.e., mean or median) for each of the subgroups.
4. Compare the observed and expected score distributions through the computation of a chi-square test statistic, and test the statistic for significance using the chi-square distribution with the appropriate degrees of freedom.

The observed and expected score distributions are often summarized in an item fit table. An example of an item fit table for an item having 5 response categories, labeled 0 – 4, and J ability subgroups is given in Table 1.

Table 1. Item Fit Table For an Item With Five Response Categories

	Score Response Category					
θ Group	0	1	2	3	4	
1	f_{10} (e_{10})	f_{11} (e_{11})	f_{12} (e_{12})	f_{13} (e_{13})	f_{14} (e_{14})	$f_{1.}$ ($e_{1.}$)
2	f_{20} (e_{20})	f_{21} (e_{21})	f_{22} (e_{22})	f_{23} (e_{23})	f_{24} (e_{24})	$f_{2.}$ ($e_{2.}$)
3	f_{30} (e_{30})	f_{31} (e_{31})	f_{32} (e_{32})	f_{33} (e_{33})	f_{34} (e_{34})	$f_{3.}$ ($e_{3.}$)
.						
.						
.						
J	f_{J0} (e_{J0})	f_{J1} (e_{J1})	f_{J2} (e_{J2})	f_{J3} (e_{J3})	f_{J4} (e_{J4})	$f_{J.}$ ($e_{J.}$)
	$f_{.0}$ ($e_{.0}$)	$f_{.1}$ ($e_{.1}$)	$f_{.2}$ ($e_{.2}$)	$f_{.3}$ ($e_{.3}$)	$f_{.4}$ ($e_{.4}$)	

In Table 1, f_{jk} and e_{jk} are the observed and expected frequencies, respectively, for individuals with ability subgroup level j and score response category k, $k = 0, \ldots, 4$.

The null hypothesis in the assessment of item fit is that, apart from random error, there is no difference between the observed and expected frequencies (or proportions) for each cell in the item fit table. For a single item, this hypothesis in terms of proportions can be written as H_o: $\pi_{jk}=P_k(\theta_j)$, $\forall$ j and k. Here, π_{jk} references the proportion of examinees from the population having ability level j and score response k. $P_k(\theta_j)$ is the model-based expected proportion of examinees from the population having ability level j and score response k.

As was stated, the tests used to assess item fit can be based on either a Pearson or likelihood-ratio chi-square statistic. Two well-known procedures for assessing item fit developed by Bock (1972) and Yen (1981) are based on the Pearson χ^2 statistic. These methods

are discussed in detail, as is a third likelihood-ratio based statistic developed by McKinley & Mills (1985).

Bock (1972) introduced a measure for detecting item misfit that utilized a Pearson χ^2 test statistic. Bock's chi square, χ^2_B, is defined for dichotomous item *i* as

$$\chi^2_{B_i} = \sum_{j=1}^{J} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}, \tag{5}$$

where *i* is the item number,

J is the number of ability subgroups,

j is the counter for the ability subgroups,

N_j is the number of examinees with an ability estimate falling within ability subgroup *j,*

O_{ij} is the observed proportion of correct responses on item *i* within ability subgroup *j*, and

E_{ij} is the expected proportion of correct responses on item *i* within ability subgroup *j*, equal to

$E_{ij} = \hat{P}_i\left(\hat{\theta}_{med-j}\right)$, where $\hat{\theta}_{med-j}$ is the median of the $\hat{\theta}$ values for examinees in subgroup *j*.

After rank-ordering examinees according to their ability estimate, Bock's (1972) procedure divides the examinees into ability subgroups of approximately equal size. In forming the observed score distribution for the dichotomous case, the examinees are classified, within their ability subgroup, as to whether they answered the item of interest correctly or incorrectly. In forming the expected score distribution, Bock's procedure utilizes the median ability level of all examinees falling within an ability subgroup, along with the item parameter estimates, in generating the IRT model-based predictions E_{ij}. Once observed and expected score distributions are computed, they are compared using Equation 5. This statistic is tested for significance by

comparing it to the χ^2 distribution with degrees of freedom defined as the number of ability subgroups *J* minus the number of estimated item parameters (*m*).

Extending this method to the polytomous case involves classifying examinees within their ability subgroup according to their response category. In the case where an item contains five possible score categories, for example, examinees would be classified into one of the *k* response categories, labeled 0 - 4, based on their response to the item. The expected score distribution would again be computed using the item response function and the median ability level of the examinees in each ability subgroup. The statistic is tested for significance by comparing it to the χ^2 distribution with degrees of freedom $J * (k-1) - m$. In the case of a 5-category item with 11 ability subgroups, the degrees of freedom would be 11 * (5-1) – 5 = 39.

Yen (1981) specified the Q_1 fit statistic, an item fit measure that is similar to χ^2_B, whose computation follows the general steps listed above. Differences between Yen's and Bock's (1972) procedures include that Yen's procedure specifies that 10 equally sized subgroups of examinees be formed, whereas Bock's procedure does not require this specification of exactly 10 groupings. In computing the Q_1 statistic, examinees are again rank ordered on the basis of their estimated ability. In forming the expected score distribution, Yen's procedure utilizes the average of the predicted probabilities of a correct response across examinees falling within the specified ability subgroup. Bock's procedure utilizes the median of the estimated abilities of examinees in the subgroup. Yen's Q_1 statistic for dichotomous item *i* is given by

$$Q_{1_i} = \sum_{j=1}^{10} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}, \tag{6}$$

where *i* is the item number,

j is the counter for ability subgroups,

N_j is the number of examinees with an ability estimate falling within ability subgroup *j,*

O_{ij} is the observed proportion of correct responses on item *i* within ability subgroup *j*, and

E_{ij} is the expected proportion of correct responses on item *i* within ability subgroup *j*, equal to

$E_{ij} = \frac{\sum_{k \in j}^{N_j} \hat{P}_i\left(\hat{\theta}_k\right)}{N_j}$, where $\hat{P}_i\left(\hat{\theta}_k\right)$ is the item characteristic function for item *i,* evaluated at θ_k.

In a similar way to Bock's χ_B^2 statistic, the Q_1 statistic is tested for significance by comparing it to the χ^2 distribution with degrees of freedom equal to the number of ability subgroups minus the number of estimated item parameters, or (*10* - *m*). Yen's statistic can also be extended for polytomous items. For the polytomous case, the degrees of freedom of the Q_1 statistic are *10 * (k-1) - m*.

In determining the degrees of freedom for the Q_1 statistic, Yen (1981) reasoned that the degrees of freedom should be adjusted for the estimation of item parameters, but not for the estimation of ability parameters. For chi-square test statistics, "Degrees of freedom are subtracted to reflect the extent to which the E_{ij} values are calculated from or are dependent on the values of O_{ij} (Lancaster, 1969, pp. 136, 142-150)", Yen (1981, p. 247). In the case of chi-square test statistics that assess item fit, the expected values, or model-based predictions, are functions of both estimated item and ability parameters. Yen stated that the item parameters are highly dependent on the observed values O_{ij}, and so degrees of freedom should be adjusted for the estimation of the item parameter estimates**.**

Yen (1981) further argued that the degrees of freedom for the Q_1 statistic should not be adjusted for the estimation of ability parameters. She argued that for tests consisting of large numbers of items, the contribution of an individual item to the estimate of ability is negligible,

and so no adjustment to the degrees of freedom for the estimation of ability is necessary (Stone, 2000). The effect is that ability is treated as if it is known (Stone & Hansen, 2000). Yen's argument relating to the degrees of freedom of her Q_1 statistic carries over to other item fit statistics such as Bock's χ_B^2.

For shorter tests, the impact of a single item response on an examinee's ability estimate increases. This impacts Yen's (1981) argument concerning the degrees of freedom of the traditional fit statistics. Literature relating to the use of the Q_1 and other traditional fit statistics for shorter tests is discussed in Section II.A.2.

An item fit statistic similar to Yen's (1081) Q_1 statistic, but based on a likelihood-ratio chi-square statistic, was developed by McKinley and Mills (1985). *LCHI* for item *i* is defined as

$$LCHI_i = 2\sum_{j=1}^{10}\left\{O_{ij}\,ln\left(\frac{O_{ij}}{E_{ij}}\right)+\left(1-O_{ij}\right)ln\left(\frac{1-O_{ij}}{1-E_{ij}}\right)\right\}, \tag{7}$$

where *i* is the item number,

j is the counter for ability subgroups,

ln is the natural logarithm function,

O_{ij} is the observed proportion of correct responses on item *i* within ability subgroup *j*, and

E_{ij} is the expected proportion of correct responses on item *i* within ability subgroup *j*.

The difference between *LCHI* and Q_1 is that *LCHI* utilizes a likelihood-ratio based statistic, whereas Q_1 is formed based on a Pearson statistic.

II. A. 1. a) General Limitations

A number of limitations are known to exist with the application of chi-square statistics as measures of fit. Perhaps the most well known limitation is the sensitivity of chi-square statistics to sample size. With small samples, chi-square statistics lack the power to detect item misfit

when it exists. When larger samples are available, chi-square tests have such high power that they will often detect item misfit when it is not present in a practical sense (Hambleton, 1993; Hambleton & Rodgers, 1986).

Another limitation in the use of chi-square test statistics relates to the sample size assumption of the tests. When chi-square statistics are used to assess item fit, problems may arise when the calculated expected frequencies for one or more score categories are small. Small expected cell frequencies could constitute a problem for a number of reasons. In cases where the expected frequencies are small, the question of whether the distribution of the test statistic approximates the chi-square distribution arises. Collins, Fidler, Wugalter, & Long (1993) caution that sparseness in item fit tables can cause the true distribution of the goodness-of-fit indices to differ substantially from known chi-square distributions. In such cases, chi-square significance tests may not be valid.

Additionally, in cases where these expectations are small, the values of the chi-square test statistics will be large. This can cause differences between the expected and observed score distributions that have little or no practical significance to result in statistical significance (Yen & Rosenberger, 1999). Further, expected counts of zero can cause computational problems in the form of divisions by zero and logarithms of zero.

II. A. 1. b) Limitations for IRT Models

Within the context of IRT, additional issues relating to the use of fit statistics arise. One issue relates to the number of ability subgroups that are formed in creating the observed and expected score distributions. Yen (1981) and Bock (1972) both utilized 10 groups of examinees in the applications of their fit statistics, and this use of 10 subgroups has carried over into a number of other goodness-of-fit studies (McKinley & Mills, 1985; Reise, 1990; Stone & Hansen,

2000). It is not through an analytic solution, however, that the specification of 10 groupings has been shown to be ideal. In fact, the number of subgroups could be chosen somewhat arbitrarily (Reise, 1990). Variability in determining the number of ability subgroups that are created can impact the value and the statistical significance of the fit statistic. In practice, however, many researchers follow Bock and Yen and utilize 10 ability groups, limiting the arbitrary nature of this aspect of the goodness-of-fit procedures.

The manner in which the ability groups are formed also poses a potential limitation of the fit statistics. The traditional item fit measures presented above form subgroups of examinees based on the examinees' estimated abilities. This implies that the cell boundaries for the test statistics are dependent on the ability estimates. As stated by Yen (1981), this may affect the null sampling distribution of the fit statistics. Orlando & Thissen (2000) also assert that model dependent observed proportions (dependent on the ability parameter) make it difficult to ascertain the true distributions of the traditional item fit statistics.

II. A. 1. c) Sampling Distributions of Item Fit Statistics for IRT Models

A number of studies have investigated the sampling distributions of the traditional fit statistics such as Bock's and Yen's. Yen (1981) argued that the dichotomous form of the Q_1 statistic follows the chi-square distribution with $J - m$ degrees of freedom, where J is the number of ability subgroups formed and m is the number of estimated item parameters. Yen carried out a simulation study where she simulated item response data sets consisting of 36 items for 1000 examinees under three models (one-parameter, two-parameter, and three-parameter). Yen found her Q_1 statistic to be approximately distributed as a chi-square random variable with $10 - m$ degrees of freedom for tests of this length. Although she found that the mean value of the Q_1

statistic was consistently higher than expected, she contended that Q_1 did approximate a chi-square distribution with 10 – m degrees of freedom for tests consisting of 36 items.

Ansley & Bae (1989), as discussed by Ankenmann (1994), carried out a simulation study investigating the sampling distribution of the Q_1 statistic for the three-parameter dichotomous IRT model. They varied test length (30 and 60 items), and examinee sample size (1000 and 2000). They found evidence that the Q_1 statistic was not distributed as chi-square with J - m degrees of freedom, but instead was distributed as a non-central chi-square random variable. In addition, they found that the non-centrality parameter varied with sample size and test length, such that for a given test length, the non-centrality parameter increased with sample size, and for a given sample size, the non-centrality parameter decreased with test length.

Stone and Hansen (2000) investigated the null sampling distributions of Pearson and likelihood ratio based chi-square item fit statistics for item response data generated under the 5-category case of the GRM. For a condition involving true ability and a test length of 32 items, they found that the sampling distributions approximated the null chi-square distribution fairly well, with some small departures. When ability was estimated and with a test length of 32 items, the sampling distributions of the statistics again approximated the null chi-square distribution fairly well. More serious departures of the sampling distributions from the null chi-square distribution were found for tests consisting of 8 and 16 items.

These studies indicate that there is some uncertainty regarding the null sampling distribution of chi-square item fit statistics. Arguably, the distributions of the traditional fit statistics can be approximated by the theoretical chi-square distribution with nominal degrees of freedom for longer tests. The approximation becomes less appropriate for shorter tests.

II. A. 1. d) Empirical Power of Item Fit Statistics for IRT Models

Several studies have investigated the empirical power of the different fit statistics to detect item misfit. Yen's (1981) simulation study compared the Q_1 statistic, a statistic proposed by Write and Panchapakesan (1969), a statistic similar to Q_1 provided by Write & Mead (1977), and an additional statistic used by Elliot, Murray, and Saunders (1977). Item response data sets consisting of 36 items for 1000 examinees were generated under the one-parameter, two-parameter, and three-parameter models. In general, the Q_1 statistic performed adequately. Yen found that the Write & Mead statistic and Q_1 yielded very similar results. Although she did not include Bock's χ_B^2 statistic in her study, she contended that the results of that statistic would also have been similar to the Q_1 statistic. Yen also noted that the statistic was not able to detect item misfit when the 2-parameter model was fitted to data generated under the 3-parameter model.

McKinley and Mills (1985) compared Bock's χ_B^2, Yen's (1981) Q_1, the Write & Mead (1977) chi-square statistic, and the LCHI statistic in a simulation study. They compared the procedures in terms of the errors made involving erroneous conclusions of misfit and erroneous conclusions of fit. McKinley & Mills simulated item response data for three sample sizes (500, 1000, and 2000 examinees), three ability distributions (low, medium, and high) and four different IRT models (one-parameter logistic, two-parameter logistic, three-parameter logistic, and two-factor linear). Each simulated test consisted of 75 items. For items modeled by the one-parameter, two-parameter, and three-parameter IRT models, the LCHI statistic resulted in the fewest false rejections as compared to the other statistics. Bock's chi-square statistic resulted in the fewest false acceptances of model-data fit. As a result, the authors could not conclude that one measure was the best. They suggested that the choice of the fit statistic be made based on the type of error that is viewed as more serious in a particular setting.

II. A. 2. Traditional Chi-Square Item Fit Statistics for Performance-Based Assessments

The traditional item fit statistics of Bock (1972) and Yen (1981) can be extended to the polytomous case. However, additional issues arise when these statistics are applied to performance-based assessments consisting of small numbers of items.

II. A. 2. a) Limitations for Performance-Based Assessments

Yen's (1981) argument that the degrees of freedom for her Q_1 statistic should not be adjusted for the estimation of ability becomes less convincing for shorter tests, because the impact of a single item response on the ability estimate increases. Yen & Rosenberger (1999) state that when dealing with shorter tests, there is "substantial part-whole contamination" when the observed and expected score distributions are compared. This is because few items are used to obtain the ability estimate, and so a single item can greatly influence the ability estimate. Consequently, the appropriate degrees of freedom of Q_1 are even more in question for shorter tests. Yen and Rosenberger recognize this as a limitation of using the Q_1 statistic for shorter tests.

Imprecision in ability estimation further affects the appropriateness of the Q_1 and other traditional chi-square goodness-of-fit statistics for shorter tests. For shorter, performance-based assessments, imprecision in the ability estimates can cause classification errors in the item fit tables. These classification errors impact the sampling distributions of the statistics (Ankenmann, 1994; Stone et al., 1993; Stone et al., 1994; Stone & Hansen, 2000), and power and Type I error rates of the significance tests (Mote & Anderson, 1965; Stone, 2003).

II. A. 2. b) Sampling Distributions of Item Fit Statistics for Performance-Based Assessments

Stone and Hansen (2000) conducted a simulation study whose purpose was to investigate the effect of precision in ability estimation on the distribution of chi-square goodness-of-fit statistics under the GRM. The researchers generated empirical sampling distributions of the fit statistics under several conditions, varying test length and sample size. Tests lengths of 8, 16, and 32 items, combined with sample sizes of 1000 and 2000 examinees were investigated. The test lengths represented the range of cases where examinee-level ability estimates would be estimated from fairly imprecisely (8 items) to precisely (32 items). In addition, the 8-item test represented a lower bound for the number of items on a typical performance-based assessment.

The item parameters in the study were obtained from one form of a mathematics performance assessment that utilized the GRM (Lane, 1993). Eight 5-category items from this form were used in the study, with the item parameters being replicated for test lengths of 16 and 32 items. Examinee sample sizes of 1000 and 2000 were chosen to avoid sparseness in the item fit tables, and to investigate the sensitivity of the goodness-of-fit statistics to sample size. Both Pearson and likelihood-ratio chi-square item fit statistics were computed, and the empirical sampling distributions of these statistics were formed over 1000 replications. Bayes expected a posteriori (EAP) ability estimates were found, and the expected score distributions were based on the median ability level in each ability subgroup, as is done in Bock's procedure.

To evaluate the distributions of the goodness-of-fit statistics, the researchers used Q-Q plots comparing the empirical distributions to the null chi-square distribution, Type I error rates, and descriptive measures of the sampling distributions. In addition, a baseline measure was created by generating empirical distributions of the fit statistics using the true ability parameter.

The effect of imprecision in the estimation of ability on the fit statistic distributions could be seen by comparing results based on estimates of ability to results based on this baseline distribution.

The researchers found only small departures from the null chi-square distributions in the empirical sampling distributions generated with a known ability level. The differences between the baseline and null distributions resulted in slightly increased Type I error rates for the condition when ability was known.

Stone and Hansen (2000) compared the conditions involving true and estimated ability, by looking at the means and variances of the sampling distributions, Q-Q plots, and Type I error rates. Under the condition involving a 32-item test, sample size of 1000, and ability estimate, the distribution of the Pearson χ^2 statistic yielded results similar to those found with the true ability parameter. Again, the results were fairly consistent with the null chi-square distribution.

For tests consisting of fewer items (8 and 16), however, results showed marked differences in the empirical distributions of the Pearson statistics from the null distributions. These differences were found especially in the form of increased Type I error rates, or concluding misfit when the model actually fits the data. The differences were more pronounced for smaller numbers of item. In other words, the less precise the ability estimate, the greater the differences between empirical and null distributions of the Pearson fit statistics. The more precise the ability estimates (as represented by the 32 item case), the closer the empirical distribution of fit statistics to the null sampling distribution. Results based on the likelihood-ratio chi-square statistic and for sample sizes of 2000 were less consistent, but showed the same general trend.

As a result, Stone and Hansen (2000) concluded that the distributions of Pearson and likelihood ratio item fit statistics were affected by imprecision in ability estimation. They suggested that this imprecision should be considered if chi-square fit statistics are to be used to assess item fit for shorter assessments.

II. A. 3. Accounting for Uncertainty in Ability Estimation with Goodness-of-Fit Statistics

Stone and colleagues (1994) described a chi-square goodness-of-fit statistic that does take into account the imprecision with which ability is estimated on shorter tests such as performance-based assessments. The applicability of this method has been investigated by Ankenmann (1994), Stone (2000), Stone (2003), Stone et al. (1993), Stone et al. (1994), and Stone and Zhang (2003). The method accounts for imprecision in ability estimation by utilizing each examinee's posterior distribution of ability, rather than only the (imprecise) point estimate of ability used in traditional chi-square item fit measures.

In traditional item fit tables, examinees are classified into one and only one cell of the item fit table, corresponding to their item response and point estimate of ability. The method described by Stone (2000) uses posterior expectations to classify examinees into several cells of an item fit table, based on their item response and estimated posterior expectation of ability at different ability levels. Instead of using only the mean or mode of this posterior distribution of ability as a single point estimate, the entire distribution is utilized. The posterior expectations are used to create pseudocounts, or a pseudo-observed score distribution, that is distributed across the ability scale. These pseudocounts (rather than actual counts) are then used in the item fit table, and differentiate the statistic being described from the traditional goodness-of-fit statistics. Table 2 provides an example of an item fit table using the pseudocounts procedure for an item

having 5 response categories, with item parameters a = -1.77, b_1 = -2.84, b_2 = 0.576, b_3 = 1.100, and b_4 = 1.540.

Table 2. Example Distribution of Pseudocounts for Three Students With Score Responses of 0, 3, and 4

	Score Response				
θ Group	0	1	2	3	4
-3.16					
-2.74					
-2.32					
-1.90	0.00				
-1.05	0.02				
-0.63	0.10			0.00	
-0.21	0.27			0.03	
+0.21	0.35			0.19	
+0.63	0.21			0.41	0.00
+1.05	0.05			0.29	0.05
+1.47	0.01			0.07	0.23
+1.90	0.00			0.00	0.37
+2.32					0.25
+2.74					0.08
+3.16					0.01

The variability of the pseudo-observed score distribution is dependent on the precision with which ability is estimated. As ability is measured more precisely, as is the case with longer tests, the posterior distribution for an examinee will be concentrated over a smaller range of abilities. This means that the pseudo-observed score distribution will take on values over a

smaller range of the ability scale. For less precise ability estimates, the posterior distribution will be spread out over a wider range of ability.

The posterior expectations from the posterior distribution are conditional probabilities that are formed by considering the examinee's response pattern and an assumed prior ability distribution. They are based on Bayes' Theorem, which relates conditional and marginal probabilities, and is given by

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}. \tag{8}$$

In the context of ability estimation, the response vector $\boldsymbol{x}$, a set of observed responses on n items, corresponds to A, and θ corresponds to B, so that the posterior expectations of ability are given by Bayes' Theorem as:

$$P(\theta \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid \theta)P(\theta)}{P(\boldsymbol{x})}, \tag{9}$$

where $P(\theta \mid \boldsymbol{x})$ is the posterior distribution of θ,

$P(\boldsymbol{x} \mid \theta)$ is the conditional probability of response pattern $\boldsymbol{x}$,

$P(\theta)$ is the prior distribution of θ, and

$P(\boldsymbol{x})$ is the unconditional or marginal probability of a response pattern $\boldsymbol{x}$ for an examinee of unknown θ randomly sampled from a population with a given distribution .

The relationship defined by Equation 9 can be stated as posterior $\propto$ likelihood $\times$ prior, meaning that the posterior distribution is proportional to the product of the likelihood function and the prior distribution. When item scores for an examinee are substituted into $P(\boldsymbol{x} \mid \theta)$, the resulting expression becomes the likelihood function $L(\boldsymbol{x} \mid \theta)$ defined as

$$L(\boldsymbol{x} \mid \theta) = \prod_{j=1}^{n} \prod_{k=0}^{m_j} \left[P_{jk}(\theta)\right]^{x_{jk}} , \tag{10}$$

where n is the number of items,

$m_j + 1$ is the number of response categories for an item, and

$\boldsymbol{x} = (x_{1k}, x_{2k}, \ldots, x_{nk})$ is the observed response pattern for an examinee, with $x_{jk} = 1$ if the examinee's response to item j falls in the k^{th} response category and $x_{jk} = 0$ otherwise.

In practice, the continuous ability distribution is approximated by quadrature points, which are a set of discrete ability levels. The estimated posterior expectation of an item for score response j and ability level k, given by r_{jk}, is defined as:

$$r_{jk} = \sum_{n=1}^{N} \frac{x_{jn} P(\boldsymbol{x_n} \mid X_k) A(X_k)}{P(\boldsymbol{x_n})}, \tag{11}$$

where n is the counter for examinees,

N is the total number of examinees in the sample,

x_{jn} is a dichotomized score equal to 1 if the item response of the n^{th} examinee for the item is j, and equal to 0 otherwise,

$\boldsymbol{x_n}$ is the response pattern of the n^{th} examinee to the set of items,

X_k is the ability level corresponding to the k^{th} quadrature point in the discrete representation of the ability distribution,

$P(\boldsymbol{x_n} \mid X_k)$ is the conditional probability of the n^{th} examinee's response pattern at quadrature point X_k,

$A(X_k)$ is a weight corresponding to the normal density function at quadrature point X_k,

$P(\boldsymbol{x_n})$ is the conditional probability of response pattern $\boldsymbol{x}_n$ for an examinee of unknown θ from a population in which θ is normally distributed, and

$$P(\boldsymbol{x}_n) \cong \sum_{k=1}^{K} P(\boldsymbol{x}_n \mid X_k) A(X_k).$$

Summing the posterior expectations over the ability distribution for a single examinee results in the numerical value of one, and corresponds to that examinee's observed frequency in a traditional item fit table. Summing the posterior expectations across examinees yields pseudocounts, or a pseudo-observed score distribution, containing the number of examinees at each ability subgroup and each item response (Stone, 2000). Summing the posterior expectations for an item across examinees and score responses yields $r_{.k}$ the posterior expectation of the number of attempts at quadrature point X_k.

The expected proportions E_{ijk} are evaluated using the item response function, the item parameter estimates for item i, and the ability level represented by quadrature point X_k. A Pearson χ^{2*} or likelihood-ratio G^{2*} chi-square test statistic is then computed using the pseudocounts, rather than actual counts, as the observed frequencies. The * denotes that the chi-square statistic utilizes pseudocounts, rather than actual counts. The formula for the Pearson chi-square, χ^{2*}, for item i is given by

$$\chi_i^{2*} = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{r_{i.k}\left(\left(r_{ijk} / r_{i.k}\right) - E_{ijk}\right)^2}{E_{ijk}} \quad , \tag{12}$$

where i is the item number,

k is the counter for quadrature points,

K is the number of quadrature points,

j is the counter for the response categories for item i,

J is the number of response categories for item i,

$r_{i.k}$ is the posterior expectation of the number of attempts at quadrature point X_k, which is the pseudocount of the number of respondents at quadrature point X_k, for item i,

r_{ijk} is the pseudocount of responses to response category j of item i at quadrature point X_k, and

E_{ijk} is the expected proportion of respondents in response category j for item i at quadrature point X_k.

The formula for the likelihood ratio form of the statistic, G^{2*}, for item i is given by

$$G_i^{2*} = 2\sum_{k=1}^{K}\sum_{j=1}^{J}\frac{r_{ijk}}{r_{i.k}} ln\left(\frac{\left(r_{ijk}/r_{i.k}\right)}{E_{ijk}}\right), \tag{13}$$

where all terms are as defined in Equation 12.

As an alternative to having the weights $A\left(X_j\right)$ used in Equation 11 correspond to the normal density function, posterior weights $A^*\left(X_j\right)$, can be estimated from the data using the formula

$$A^*\left(X_j\right) = \frac{r_{.k}}{\sum_{k=1}^{K} r_{.k}}, \tag{14}$$

where $r_{.k}$ is the posterior expectation of the number of attempts at $\theta = k$, and is found by summing the posterior expectations r_{jk} over examinees and across the j score responses,

k is the index for quadrature points, and

K is the number of quadrature points.

The mean of the posterior distribution of ability for an examinee given by Equation 9 is called the *expected a posteriori* (EAP) estimator, and is given by

$$EAP(\theta) = E(\theta \mid \boldsymbol{x}) = \int P(\theta \mid \boldsymbol{x})\theta d\theta = \frac{\int L(\boldsymbol{x} \mid \theta)P(\theta)\theta d\theta}{\int L(\boldsymbol{x} \mid \theta)P(\theta)d\theta}. \tag{15}$$

where the terms are as defined above. This value is approximated by

$$EAP\left(\hat{\theta}\right) \cong \frac{\sum_{k=1}^{K} P\left(\boldsymbol{x}_n \mid X_k\right) A\left(X_k\right)}{\sum_{k=1}^{K} P\left(\boldsymbol{x}_n \mid X_k\right) A\left(X_k\right)}. \tag{16}$$

The variance of the posterior distribution, $Var(\theta)$ is defined as

$$Var(\theta) = Var(\theta \mid \boldsymbol{x}) = \int \theta^2 P(\theta \mid \boldsymbol{x}) d\theta - \left[E(\theta \mid \boldsymbol{x})\right]^2, \tag{17}$$

where the terms are again as defined above. ($Var(\hat{\theta})$), is approximated by

$$Var\left(\hat{\theta}\right) \cong \frac{\sum_{k=1}^{K} \left(X_k - \hat{\theta}\right)^2 P\left(\boldsymbol{x} \mid X_k\right) A\left(X_k\right)}{\sum_{j=1}^{J} P\left(\boldsymbol{x} \mid X_k\right) A\left(X_k\right)}, \tag{18}$$

where all terms are as previously defined.

II. A. 3. a) Studies Utilizing the G^{2*} and χ^{2*} Test Statistics

These test statistics based on pseudocounts, G^{2*} and χ^{2*}, cannot be compared to theoretical chi-square distributions with known degrees of freedom as the traditional fit statistics arguably can. This is because a key assumption of chi-square tests, that the observations are independent, is not met when pseudocounts are utilized (See Table 2). However, studies investigating the utility of the pseudocounts procedure have shown that the G^{2*} statistic closely approximates a scaled chi-square random variable (Ankenmann, 1994; Stone, 2000; Stone et al., 1993; Stone et al., 1994). These studies implemented Monte Carlo techniques to generate empirical sampling distributions of the G^{2*} fit statistics, and found that these distributions could be approximated by scaled chi-square random variables.

Stone et al. (1993) fit the GRM to data from the QUASAR Cognitive Assessment Instrument (QCAI), a performance-based assessment in which each examinee received a form containing 9 performance-based tasks (Lane, 1993; Lane, Liu, Ankenmann & Stone, 1996; Lane,

Stone, Ankenmann & Liu, 1995). Due to of the small number of tasks given to each examinee, the researchers utilized the G^{2*} fit statistic in assessing the fit of the items for their test.

In order to statistically test the fit of the items using G^{2*}, Stone et al. (1993) utilized Monte Carlo resampling techniques to empirically generate sampling distributions of the fit statistics. The method implemented by the researchers in generating the sampling distributions involved a number of steps. The observed data were calibrated using MULTILOG (Thissen, 1991), and a G^{2*} statistic was calculated for each item. To form the sampling distributions, the researchers simulated item responses using the item parameter estimates obtained from MULTILOG, and an ability value sampled from the posterior distribution of ability. From the simulated data sets, a G^{2*} statistic was calculated for each item.

The process of simulating item response data sets and computing a G^{2*} statistic for each item was repeated 1000 times, forming the sampling distributions of the G^{2*} statistics for each item. Statistical significance was determined by comparing the observed fit statistic with the 95th percentile of these empirically generated G^{2*} fit statistic distribution for that item. If the observed fit statistic was larger than this value, the item was flagged as misfitting.

Stone et al. (1993) studied the empirically generated sampling distributions of the G^{2*} fit statistics, and found evidence that the G^{2*} statistics were distributed as scaled chi-square random variables. This conclusion was drawn after examination of Quantile-Quantile (Q-Q) plots of the empirical data versus theoretical chi-square distribution.

Q-Q plots can be used to assess whether an empirical distribution follows a theoretical distribution. If a theoretical distribution is a close approximation to an empirically generated distribution, the quantiles of the empirical distribution will match those of the theoretical distribution. The points of the Q-Q plot would fall close to the line $y = x$ (Chambers, Cleveland,

Kleiner, & Tukey, 1983). Large deviations from this reference line indicate differences between the empirical and theoretical distributions.

Chambers et al. (1983) state that the straightness of Q-Q plots can be used to determine whether empirical and theoretical distributions come from the same family of theoretical distributions. Q-Q plots that do not follow linear patterns indicate that the empirical and theoretical distributions have different shapes.

Chambers et al. (1983) further show that certain types of deviations from the line $y = x$ indicate specific distributional differences between the theoretical and empirical distributions. For instance, if the data points fall on a line parallel to the line $y = x$, the two distributions are similar except for a difference in location. Adding or subtracting the appropriate constant to each observed data point would result in a shift of the Q-Q plot onto the line $y = x$.

If the data points do not fall on the line $y = x$, but instead fall on a line passing through the origin but not parallel to $y = x$, the two distributions are similar except for a difference in spread. In this case, multiplying all data points by the appropriate single positive constant (scaling factor) would result in a shift of the data onto the line $y = x$. If the data is compressed, an appropriate scaling factor less than 1 could be applied to the empirical data. If the data is stretched out, the appropriate scaling factor would be greater than 1. Rescaling the empirical distribution in this manner would shift the data in the Q-Q plot onto the line $y = x$.

If the data points do not fall on the line $y = x$, but instead follow a different pattern, it is possible that transformed empirical data (e.g., the natural logarithm of the data) would follow the theoretical distribution.

The finding that the G^{2*} statistics were distributed as scaled chi-square random variables (Stone et al., 1993), was largely due to comparisons of the empirically generated sampling

distributions to theoretical chi-square distributions using Q-Q plots. The theoretical chi-square distributions had degrees of freedom equal to the means of the empirically generated sampling distributions. In general, the shapes of these Q-Q plots were linear, and the data were compressed. This suggested that the empirical distributions followed scaled chi-square distributions. For some items, the authors noted degeneration in the linearity at the tails of the distributions, possibly indicating the presence of outliers in the distributions.

These researchers then attempted to estimate the scaling factor and degrees of freedom values, or scaling corrections, for each item. To estimate the scaling corrections, they used the means and variances from the empirical sampling distributions with the method of moments. The assumption was that $G^{2*} \sim \gamma G^2$, where G^{2*} is the fit statistic obtained from the empirically derived distribution, and G^2 is distributed as a chi-square random variable with ν degrees of freedom. For chi-square random variables, the mean and variance are given by $E[G^2] = \nu$ and $Var[G^2] = 2\nu$, respectively. Thus, $E[G^{2*}] = E[\gamma G^2] = \gamma E[G^2] = \gamma\nu$, and $Var[G^{2*}] = Var[\gamma G^2] = \gamma^2 Var[G^2] = 2\nu\gamma^2$.

When numerical values for the mean $E[G^{2*}]$ and the variance $Var[G^{2*}]$ of the empirical distributions are substituted into these equations, it is possible to solve the system of equations for the scaling factor γ and the degrees of freedom ν. Using the mean $\bar{x}_{G^{2*}}$ and the variance $s^2_{G^{2*}}$ of the empirical sampling distributions in the above equations and solving the system of equations for γ and ν yields:

$$\gamma = \frac{s^2_{G^{2*}}}{2\bar{x}_{G^{2*}}} \text{ and} \tag{19}$$

$$\nu = \frac{2\bar{x}^2_{G^{2*}}}{s^2_{G^{2*}}}. \tag{20}$$

Based on the assumption that G^{2*} was a scaled chi-square random variable, the observed fit statistics for each item could be divided by the derived scaling factor γ, and compared to the theoretical chi-square distribution with ν degrees of freedom for purposes of significance testing. Stone et al. (1993) did not carry out the significance testing in this manner. However, they did find that in general, the transformed empirical data corresponded fairly well to the theoretical chi-square distribution with ν degrees of freedom for the test items. They felt that significance testing could be accurately carried out using the appropriate theoretical chi-square distribution for each item.

Chambers et al. (1983) caution that conclusions drawn from Q-Q plots with estimated shape (scaling) parameters are only as robust as the estimation procedures used for determining the shape parameters. Here, this implies that conclusions regarding the distribution of the fit statistics are only as robust as the procedure for finding the scaling factor from the empirically generated sampling distributions. The sampling distributions of the scaling corrections were never directly studied, and so the robustness of the estimation procedure is somewhat unknown. This illustrates a possible limitation in the procedures for determining the scaling factors in the current study. Stone et al. (1993) found some variation in the scaling factors and effective degrees of freedom across items, and attributed the variation to sample sizes, inadequacies in the estimation of item parameters, or task difficulty.

Stone (2000) investigated the use of resampling methods to estimate the scaling factor (γ) and degrees of freedom value (ν) for G^{2*} item fit statistics. The method involved simulating item response data according to the GRM using item parameter estimates and a randomly generated ability (N(0,1)), calculating the G^{2*} fit statistic for each item using the original item parameter estimates, and repeating these steps to generate sampling distributions. This marks a difference

in the generation of sampling distributions from the method used by Stone et al. (1993). Stone (2000) used the original item parameter estimates, and randomly selected θ for each replication, while Stone et al. (1993) re-estimated the item parameter estimates for each replication.

Stone (2000) then used Equations 19 and 20 to estimate the scaling corrections for each item. Each fit statistic in the empirically generated sampling distribution for an item was rescaled by the scaling factor, and the rescaled G^{2*} distributions were compared to theoretical chi-square distributions. Rather than compare the rescaled fit statistic to a chi-square distribution with ν degrees of freedom, the degrees of freedom were adjusted for the estimation of item parameters, as is done with the traditional methods. Q-Q plots comparing the rescaled G^{2*} distributions to theoretical chi-square distributions showed that the family of chi-square distributions provided good approximations to the rescaled fit statistic distributions.

Stone (2000) compared two methods of assessing item fit using G^{2*} fit statistics. The first procedure based decisions of item fit on percentile rank of the empirical sampling distributions as was done by Stone et al. (1993). For the second procedure, decisions of item fit were made by comparing rescaled G^{2*} statistics to theoretical chi-square distributions with the ν degrees of freedom adjusted for the estimation of item parameters. The scaling factors and degrees of freedom values were found using Equations 19 and 20. Stone found the percent agreement with respect to decisions of item fit to range between .90 and .95, indicating that the procedures are fairly interchangeable. He also found that the resampling method used in Stone (2000) yielded slightly less power than the method used by Stone et al. (1993).

II. A. 3. b) Predicting the Scaling Corrections of G^{2*} for the Rasch Model

Stone et al. (1993) examined the sampling distributions of G^{2*} fit statistics, and found that the statistics approximated scaled chi-square distributions. Further, they provided evidence that

related the scaling corrections (scaling factor and degrees of freedom values) to sample characteristics, namely the variance of the posterior distribution of ability. Specifically, they found that as the posterior variance of ability decreased, the scaling factor increased to 1, and the degrees of freedom increased. The researchers suggested that estimating these scaling corrections from this or other sample characteristics may be possible. If the scaling factors γ and degrees of freedom ν could be estimated using the sample characteristics, then the empirical generation of G^{2*} or χ^{2*} sampling distributions would not be necessary for significance testing of this statistic.

Stone et al. (1994) carried out a simulation study that investigated whether the scaling corrections could in fact be estimated from sample characteristics, as the results by Stone et al. (1993) had suggested. The focus of this paper was the Rasch model, and the factors in the study were test length (5, 10, 20, and 40 items) and the interval width used to create ability subgroups (0.8, 0.4, and 0.2). The sample characteristics investigated were the mean posterior variance of the examinees (posterior variance of ability averaged over examinees and replications of a condition of the study), and the ability subgroup width.

In this Stone et al. (1994) study, empirical sampling distributions of the G^{2*} statistics were generated. Then, the scaling factor γ and degrees of freedom ν were derived for each item using Equations 19 and 20, respectively.

Regression equations for predicting the scaling factor $\hat{\gamma}$ and degrees of freedom $\hat{\nu}$ from the mean posterior variance and the ability subgroup width were then formed. It was found that ability interval width did little to improve the predictions of these scaling corrections, and this variable was eliminated as an independent variable. Again, the goal was to rescale the empirical distributions for each item by the predicted scaling factor $\hat{\gamma}$, and use the chi-square distribution

with predicted degrees of freedom $\hat{\nu}$ for significance testing. Using scaling corrections predicted from sample data in this way would eliminate the need for the generation of empirical sampling distributions for significance testing.

Stone et al. (1994) evaluated their prediction equations by applying the predicted scaling corrections $\hat{\gamma}$ and $\hat{\nu}$ to the empirically generated sampling distributions. The correspondence between the distribution of fit statistics transformed by $\hat{\gamma}$ and the theoretical chi-square distribution with $\hat{\nu}$ degrees of freedom was then evaluated. Correspondence between these two distributions for an item implied that the transformed sampling distribution of G^{2*} fit statistics followed the chi-square distribution with $\hat{\nu}$ degrees of freedom for that item.

In general, the researchers found that the scaling corrections were predicted fairly well from the mean posterior variance for the Rasch model. For test lengths of 10, 20, and 40-items, they found that test-level prediction equations could be used to assess the fit of each item. This finding is relevant because it implies that with the Rasch model, different prediction equations are not needed for items with different difficulty values. Instead, only a single pair of test-level prediction equations is needed to predict the scaling corrections for each item on a test. That is, one equation predicting the scaling factor $\hat{\gamma}$ and one predicting the degrees of freedom $\hat{\nu}$, is needed.

This result did not carry over to the 5-item test, with the finding in that case being that the difficulty value may play a part in obtaining the prediction equations. In addition, close correspondence between the transformed empirical distributions and the chi-square with $\hat{\nu}$ degrees of freedom was observed for ability interval widths of 0.2 and 0.4, but not 0.8.

II. A. 3. c) Predicting the Scaling Corrections of G^{2*} for the GRM

Ankenmann (1994) furthered the research of Stone et al. (1994) by investigating whether the scaling corrections needed for significance testing of G^{2*} fit statistics could be estimated from sample data under the GRM with 5 score categories. His simulation study varied test length (4, 8, 16, and 32 items) and ability interval width (0.2, 0.4, and 0.8), with the item parameters for the 4-item case being repeated to construct longer test conditions.

Using the same methodology as Stone et al. (1994), Ankenmann (1994) generated empirical sampling distributions of G^{2*} fit statistics, and obtained the scaling factor γ and degrees of freedom ν for each item using Equations 19 and 20, respectively. Then, regression equations for predicting the scaling factors $\hat{\gamma}$ and degrees of freedom $\hat{\nu}$ were formed. Average posterior variance served as the independent variable (ability subgroup width was again not a useful addition to the prediction equations). These prediction equations were evaluated through comparisons of the sampling distributions rescaled by $\hat{\gamma}$ to the theoretical chi-square distributions with $\hat{\nu}$ degrees of freedom for an item.

Ankenmann found that under the GRM, item-level prediction equations were needed. That is, one equation for predicting $\hat{\gamma}$ and one equation for predicting $\hat{\nu}$ for each item were necessary. The predicted item-level scaling factors ($\hat{\gamma}$) were fairly effective in transforming the observed G^{2*} fit statistic distributions so that they approximated the theoretical chi-square distributions with predicted degrees of freedom $\hat{\nu}$.

In evaluating the utility of the item-level prediction equations, Ankenmann (1994) applied the prediction equations to data sets with different item parameters than were used in his simulation study. The purpose of doing this was to determine how well the prediction equations generalized to real applications utilizing the GRM with 5 score categories. He found that it was

necessary to consider item characteristics as matching tools in deciding which set of item-level prediction equations to use.

To find the best prediction equations, Ankenmann (1994) matched items based on their item discrimination and/or item information, to an item in his simulation study. Once matched in this way, the prediction equations were useful for detecting misfit for items with parameters different from those used in the simulation study. Ankenmann found that for most items, the prediction equations could be validly applied to real test data for purposes of significance testing of the observed G^{2*} fit statistics. This conclusion was based on evidence from Q-Q plots and Type I error rates. In addition, it was found that item information served as a better matching criterion than item discrimination. The current study builds upon the results of Stone et al. (1994) and Ankenmann (1994) by investigating the utility of the prediction technique under broader simulation conditions.

II. A. 3. d) Null Sampling Distribution of the G^{2*} and χ^{2*} Item Fit Statistics

Stone et al. (1994) investigated the null sampling distribution of the G^{2*} fit statistic while manipulating test length (5, 10, 20, and 40) and interval width for ability subgroups (.8, .4, and. 2). Item response data was generated under a one-parameter IRT model. The researchers found a close approximation to the null chi-square distribution for the condition involving the 40 item test (where ability is more precisely estimated) and ability interval with of .8. This indicated that with longer test lengths and wider interval widths, the G^{2*} fit statistic becomes equivalent to the traditional fit statistics that use point estimates of ability. For tests of shorter lengths, the researchers found that the sampling distributions of the G^{2*} fit statistics were from the family of chi-square distributions, but differed in location and spread from the null distribution.

Ankenmann (1994) investigated the null sampling distribution of the G^{2*} fit statistic for tests of different lengths when data was generated under the GRM. Ankenmann considered three test lengths (8 item, 32 item using calibrated point estimates of ability to compute G^{2*}, and 32 item using true ability values to calculate G^{2*}). Under conditions using true ability, a close approximation to the null chi-square distribution was observed. The sampling distributions for the 32 item tests and estimated ability also showed a close approximation to the null chi-square distribution.

For the 8 item test, the sampling distributions of the G^{2*} fit statistics did not approximate the chi-square distributions with nominal degrees of freedom. Instead, evidence was found that the sampling distributions were from the family of chi-square distributions, but different in shape or both shape and spread. These results indicate that it is not appropriate to use the chi-square distribution with nominal degrees of freedom for testing the significance of G^{2*} in the presence of imprecise ability estimates. Instead, the appropriate scaling factor and degrees of freedom values must be found.

II. A. 3. e) Performance of the G^{2*} and χ^{2*} Test Statistics

Stone (2003) investigated the empirical power and Type I error rates of both the G^{2*} and χ^{2*} statistics by conducting a simulation study in which number of items (6 and 12 items), shape of the ability distribution (normal and positively skewed), and sample size (500, 1000, and 2000 examinees) were manipulated factors. Item parameters for the study were obtained from items on the 1994 NAEP Reading Assessment fit to the 2- and 3-category GRM.

In order to investigate Type I error rates and power, model-data misfit was introduced into the study. For the misfitting items, the item parameters used to compute the item fit statistics were altered from those used to generate the item response data. Misfit was introduced

to one 2-category or one 3-category item by altering either the slope parameter by .3 or .5, or by altering the threshold parameter(s) by .25 or .5. To investigate Type I error rates, the percent of misfit detected across 100 replications of the study at α = .10, .05, and .01 was calculated for the items whose parameters were not altered. To investigate empirical power, the percent of item misfit detected across the 100 replications for the items with altered item parameters was calculated.

Stone (2003) found that for data simulated using the normal ability distribution, Type I error rates matched their nominal levels for tests consisting of 6 and 12 items, regardless of sample size. In addition, Type I error rates for the G^{2*} and χ^{2*} statistics were similar. For the skewed ability distributions, Type I error rates were well above their nominal levels for all conditions, and increased as sample size increased and test length decreased.

In considering empirical power for the conditions where ability was normally distributed, Stone (2003) found that power was high for sample sizes of 1000 and 2000, and adequate for sample sizes of 500, when misfit was introduced into the 12-item tests. Results for the 6-item tests showed less promise. In general, Stone found that power decreased as alpha decreased, and increased as sample size increased. It was also found that the power of the G^{2*} statistic was slightly less than that for χ^{2*}. Empirical power for the conditions involving skewed ability distributions were not investigated due to the high Type I error rates found under those conditions.

Stone and Zhang (2003) investigated the performance of the χ^{2*} statistic for dichotomously scored items. This study looked at the empirical power and Type I error rates of the χ^{2*} statistic in comparison to two other goodness of fit methods. The authors manipulated number of items (10, 20, and 40 items), and sample size (500, 1000, and 2000 examinees).

To examine the empirical power and Type I error rates, Stone and Zhang (2003) introduced model misfit in two ways. First, a different IRT model was used to calibrate the data and assess fit than was used to simulate the data. For example, a 3-parameter model was used to simulate the data, but the data were assessed for fit under a 1-parameter model. Second, item misfit was introduced for specific items by altering the item parameters used to calibrate the data from those used to generate the data. For instance, the authors altered either the slope parameters (by .5) or threshold parameters (by .25) for two items when calculating the fit statistics for those items.

The authors examined the Type I error rates by examining the percent of misfit detected across 100 replications of each experimental condition α = .10, .05, and .01 for the items generated to fit the model. Empirical power was examined by looking at the percent of misfit detected across 100 replications of each experimental condition for the items for which misfit was introduced.

Type I error rates for the χ^{2*} statistic were found to be at the nominal levels regardless of test length or sample size. The power of the χ^{2*} statistic was evaluated under two types of model misfit. When model misfit was introduced by calibrating the data under a different model than was used to simulate the data, power was found to increase as sample size increased, and remained relatively consistent across test length.

Power levels for the χ^{2*} statistic depended on the combination of models used to generate and calibrate the data. For data generated for 1000 or more examinees under the 2- or 3-parameter models and fitted under the 1-parameter model, the power of χ^{2*} was adequate. However, the χ^{2*} statistic exhibited inadequate power for detecting misfit when data were modeled under the 3-parameter model and evaluated under the 2-parameter model. When model

misfit was introduced by altering item parameters, the χ^{2*} statistic exhibited high power when examinee sample sizes were 1000 or more.

The results from the Stone (2003) study indicate that the G^{2*} and χ^{2*} statistics can be used to detect item misfit for tests which yield imprecise individual ability estimates, such as performance-based assessments or assessments utilizing a matrix-sampling design. Further, Stone and Zhang (2003) supported the use of the χ^{2*} statistic for detecting item misfit for tests consisting of as few as 10 dichotomous items, when examinee sample sizes were 1000 or more.

II. B. Additional Methods for Assessing Item Fit for IRT Models

The limitations of chi-square goodness-of-fit statistics indicate that a single type of evidence by itself may not be sufficient for assessing item fit. Several graphical methods that can aid in the assessment of item fit have been suggested (Hambleton, 1993; Hambleton & Rodgers, 1986; Hambleton & Swaminathan, 1985; Kingston & Dorans 1985). Also, additional methods of assessing fit for shorter, performance-based assessments have been suggested (Donoghue & Hombo,1999; Donoghue & Hombo 2001a, Donoghue & Hombo 2001b; Orlando & Thissen, 2000). Summaries of these methods follow.

II. B. 1. Graphical Methods for Detecting Item Misfit

In addition to using goodness-of-fit tests, it is suggested that other methods of detecting item misfit, such as graphical comparisons of observed and expected score distributions, be utilized (Hambleton, 1993; Hambleton & Rodgers, 1986; Hambleton & Swaminathan, 1985). Plotting the observed versus expected score distributions allows for a visual representation of the fit between the two distributions.

Hambleton (1993) also suggested the analysis of residuals as a means for assessing item fit. This analysis involves obtaining the observed and expected score distributions across 10-15 ability subgroups for an item. The residuals, or the differences between actual and expected item performance for a subgroup, are then obtained. Plots of the residuals or standardized residuals versus ability provide evidence of item fit or misfit. Residual plots that show a random scatter about zero indicate item fit, while residual plots that show patterns indicate item misfit.

Also, Kingston & Dorans (1985) suggested an exploratory technique of assessing fit that involves the analysis of item-ability regressions. This method graphically compares the regression of the observed proportion of correct responses on the estimated ability with the estimated item response function.

First, the ability scale is divided into 15 intervals between $-3 \leq \theta \leq 3$, and the proportion of examinees in each interval that provide a correct response (adjusted for omits) is computed. The estimated item response function is plotted, and these proportions are plotted on the same curve as squares whose areas are proportional to the sample size in each interval. Then, vertical lines serving as rough estimates of a .95 CI are marked on the item response function at each ability subgroup. As a summary statistic, the number of times the midpoints of the proportion correct boxes fell off of the approximated .95 CI are counted.

These two types of graphical analyses build on the information provided by chi-square significance tests and the procedures can be used in conjunction in the assessment of item fit.

II. B. 2. Alternative Methods of Assessing Item Fit for Performance-Based Assessments

Orlando and Thissen (2000) introduced a method for assessing item fit for dichotomous items in which examinees are grouped according to an aspect of the observed data rather than

their ability estimates. The chi-square test statistic for each item is computed using observed and expected proportions of correct and incorrect responses for each possible number correct score.

The method for obtaining the expected proportions requires the calculation of two likelihood functions for each number correct score. One is formed for each number correct score including all items, and another for each number correct score when each item is omitted. The number correct score likelihoods with and without each item are combined with the omitted item to produce the expected score distribution for each number correct score and score response. The observed proportions for each item and number correct score group are computed from the data.

The observed and expected proportions are compared using either a Pearson (S-χ^2) or likelihood-ratio (S-G^2) chi-square goodness-of-fit statistic. Because the expected frequencies are computed for the total test score, and the model-dependent ability estimates are not used to obtain the observed proportions, Orlando and Thissen (2000) are able to compute a chi-square test statistic.

Orlando and Thissen (2000) conducted a simulation study evaluating the performance of their fit statistics. They evaluated the performance of S-χ^2 and S-G^2 for three test lengths (10, 40, and 80 items) and data generated under 3 IRT models (1-, 2-, and 3-parameter models). The authors investigated the Type I error rates and empirical power of their statistics. In the evaluation of power, model misfit was introduced by using a different model to calibrate the data than was used to generate the data. The performance of S-χ^2 was also evaluated for different degrees of model misfit.

Results showed that nominal level Type I error rates held for S-χ^2 across sample size, but Type I error rates for S-G^2 increased as sample size increased. Results further showed that the

power of S-χ^2 was relatively high for detecting pronounced misfit when data were generated under the 2- and 3-parameter models and fitted under the 1-parameter model. The power of S-χ^2 was inadequate for detecting misfit for data generated under the 2-parameter and fitted under the 3-parameter model. While the authors state that S-χ^2 is promising for detecting item misfit for dichotomous items, its application to the polytomous item case was not investigated.

Donoghue & Hombo (1999), as discussed by Donoghue & Hombo (2001a), analytically examined the sampling distribution of the χ^{2*} statistic, which they call Q_{DH}, for dichotomous items. They showed that the fit measure, under the assumption that the item parameters are fixed and known, is asymptotically distributed as a quadratic form of a normal random variable. The authors extended this result to the polytomous item case, and showed with the same assumption, that the distribution of the fit statistic is a quadratic form of normals (Donoghue & Hombo, 2001a).

While obtaining an analytic solution is desirable, the usefulness of the statistic applied to real data must also be investigated. The introduction of estimated, rather than known, item parameters affected the distribution of the fit statistic for the dichotomous case. Donoghue & Hombo (2001b) state that the impact of estimated item parameters on the null distribution of the fit statistic would likely extend to the polytomous item case. The authors are currently investigating methods for determining the distribution of the fit statistic with estimated, rather than known, item parameters.

Stone and Zhang (2003) compared the performance of the χ^{2*} statistic to the methods introduced by Orlando and Thissen (2000) and Donoghue and Hombo (1999, 2001a, 2001b). Test length and examinee sample size were manipulated. Type I error rates and empirical power of the methods were compared for tests consisting of dichotomously scored items. Item misfit

was introduced by generating data under one IRT model and calibrating it under another, and also by altering the item parameters used in generating and calibrating the data.

The Orlando and Thissen (2000) statistic and χ^{2*} exhibited nominal level type I error rates regardless of test length or sample size. The Donohue and Hombo method was only applied to tests consisting of 10 items due to excessive computational demands for tests of longer lengths. For the 10 item tests, the Type I error rates of the Donohue and Hombo method were below the nominal level.

The performance of the procedures with respect to empirical power depended on the combinations of models used in data generation and calibration. In general, the χ^{2*} statistic displayed higher power than the other methods, with the differences lessening as the sample increased. The Orlando and Thissen (2000) statistic displayed adequate power only for sample sizes of 2000, and none of the methods displayed adequate power for detecting misfit when 3-paremeter data were fitted with a 2-parameter model. Again, performance of the Donohue and Hombo method was assessed only for tests consisting of 10 items.

When model misfit was introduced by altering either the slope or threshold parameters of two items, χ^{2*} exhibited greater power than the other two methods, and displayed adequate power for sample sizes of at least 1000 examinees. The power of the Orlando and Thissen (2000) statistic depended on the discrimination parameter of the item, and the procedure was inadequate for detecting misfit in higher discriminating items. Power was investigated for the Donoghue and Hombo method only for tests having 10 items. For tests of this length, power levels were similar to those found by the Orlando and Thissen method, and were lower than those found for the χ^{2*} statistic.

Stone and Zhang (2003) compared the performance of the χ^{2*} statistic to two alternative procedures for assessing item fit in tests consisting of dichotomously scored items. Due to computational demands, results for the Donoghue and Hombo method were incomplete across conditions of the study. While the study did not consider polytomously scored items, the authors provided support for using χ^{2*} and the Orlando and Thissen (2000) statistic for assessing item misfit for dichotomous tests.

II. C. Graded Response Model

The applications in this study utilize Samejima's (1969) graded response model (GRM). The GRM is an IRT model appropriate for items having ordered response categories, where higher categories indicate greater ability. Under this model, examinee responses fall in only one of the ordered categories for each item. The GRM is a direct extension of the two parameter logistic (2PL) IRT model to the polytomous item case where item i has $m_i + 1$ score categories. Under the homogeneous case of the GRM, it is assumed that the thinking process involved in solving the item is homogenous; therefore the item is described by a single discrimination parameter. References to the GRM made throughout this paper refer to the homogenous case of the model.

Under the GRM, boundary characteristic curves (Baker, 1992; Samejima, 1969) are produced for each response category. These curves represent the probability that an examinee with ability θ will respond in response category k or higher for item i. The curves are defined mathematically as $P_{ik}^{*}(\theta)$:

$$P_{ik}^{*}(\theta) = \frac{1}{1 + exp\left[-Da_i(\theta - b_{ki})\right]}, \tag{21}$$

where i is the item number,

$k = 0, 1, \ldots m_i$ is the response category,

$m_i + 1$ is the number of ordered response categories for item i,

D is 1 or a scaling constant of 1.702 that may be introduced,

θ is the latent trait parameter,

a_i is the slope parameter for item i, and

b_{ki} are the threshold parameters for category k of item i.

By definition, $P_{i0}^{*}(\theta)$, the probability of an examinee responding in the lowest item response category or higher, is one, and $P_{i,m_i+1}^{*}(\theta)$, the probability of responding with the highest response, is zero.

Under the GRM, each item is described by one discrimination, or slope, parameter. This implies that the graphs of the $P_{ik}^{*}(\theta)$ values for each of the response categories of a single item will be parallel. The steepness of these curves is determined by the numerical value of the discrimination (a_i) parameter, with larger values yielding steeper curves. Larger discrimination values indicate that the item response categories discriminate or differentiate well between examinees at different levels of ability.

Each item modeled by the GRM is also described by m_i threshold parameters (b_{ki}). These threshold values increase monotonically for each item. They provide an indication of spread of the boundary characteristic curves. Threshold values covering a wide range of the ability scale result in boundary characteristic curves that are spread out. Threshold values covering a more narrow range on the ability scale result in curves that fall closer together. An example of the four boundary characteristic curves for a five category item having item parameters (a = 1.7, b_1 = -2, b_2= -0.6667, b_3 = 0.6667, b_4 = 2) is given in Figure 1.

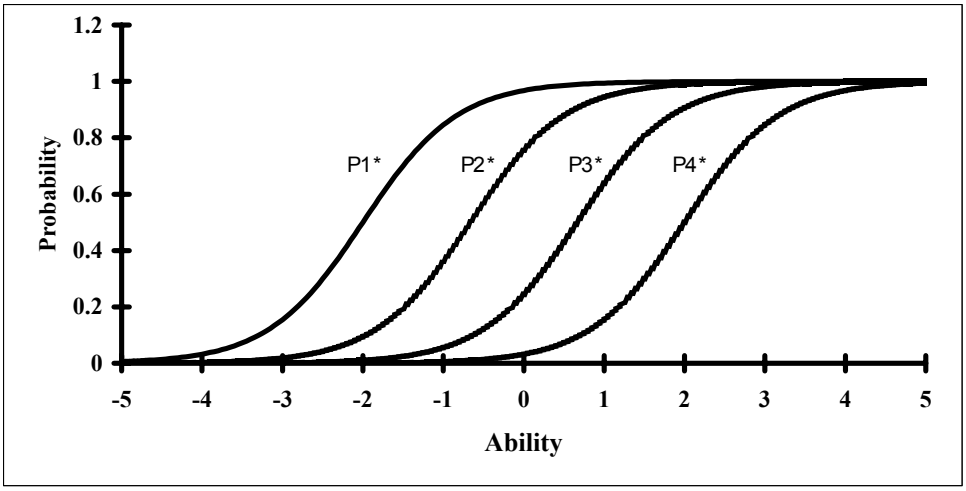

Figure 1. Boundary Characteristic Curves for a 5 Category Item

For the GRM, $P_{ik}^{*}(\theta)$ represents the cumulative probability of a randomly selected examinee with ability θ responding in category *k* or higher on item *i*. In order to find the probability that an examinee responds to a specific category *k* of item *i*, one must compute the difference between the cumulative probabilities for two adjacent categories. In general, the probability of an examinee responding to category *k* of item *i*, $P_{ik}(\theta)$, is given by the difference $P_{ik}^{*}(\theta)$ - $P_{i,k+1}^{*}(\theta)$. As an example, the probability of an examinee responding to the third response category of item *i*, having response categories labeled 0, 1, … m_i +1, is given by $P_{i2}(\theta) = P_{i2}^{*}(\theta)$ - $P_{i3}^{*}(\theta)$.

A plot of the $P_{ik}(\theta)$ values provides the operating characteristic of a particular item response category (Samejima, 1969). These functions are also called item response category characteristic curves (Baker, 1992), and trace lines (Thissen, Steinberg, & Mooney, 1989). The plots of these response category probabilities do not share a common form. The curve representing the highest response category probabilities for an item is always monotonically

increasing, while the curve representing the lowest response category is monotonically decreasing. The curves for the intermediate categories are not monotonic; instead they increase up to a point and then decrease

The threshold parameters can be interpreted from the response category characteristic curves. For the highest and lowest response categories, the threshold values are the points along the ability scale at which the probability that the response will be allocated to that category is .5. For the other response categories, the threshold values provide the modal point of the item response category characteristic curves.

An example of the item response category characteristic curves for a five category item having item parameters (a = 1.7, b_1 = -2, b_2= -0.6667, b_3 = 0.6667, b_4 = 2) is given in Figure 2.

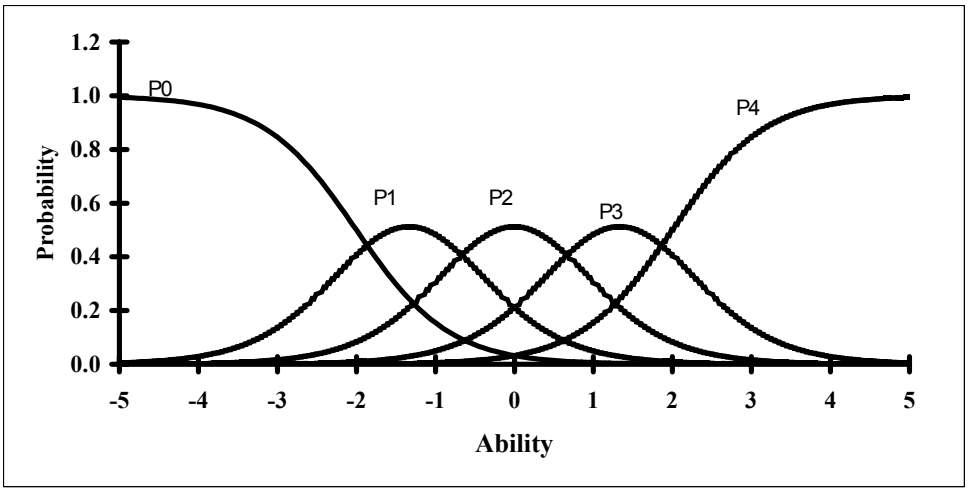

Figure 2. Item Response Category Characteristic Curves for a 5 Category Item

As with all IRT models, the GRM has certain assumptions that must be met in order for the benefits of this testing theory to be attained. The assumptions of the GRM are that: (a) the test is unidimensional; (b) at a specific ability level, examinees' responses to the items on the test are independent; (c) the test is non-speeded; (d) the form of the model is appropriate, and (e) the score categories of the items are ordered. Inherent in the use of the homogenous case of the

GRM is the idea that the reasoning process for a single item is homogenous across the item response categories for that item.

II. C. 1. Item Parameter Estimation for the GRM

In applications of IRT to real testing situations, both item and ability parameters are estimated. To ensure the validity of all test applications utilizing IRT, it is important that accurate parameter estimates are obtained. Several studies have investigated parameter estimation for the GRM. Reise & Yu (1990) conducted a simulation study that investigated the effects of sample size (250, 500, 1000, and 2000 examinees), ability distribution (normal, uniform, and skewed), and true discrimination parameter distribution (3 uniform distributions) on parameter recovery for the GRM. The authors used a test consisting of 25 5-category items, each modeled by the GRM, and found that sample sizes of at least 500 were needed to obtain accurate item parameter estimates under the conditions of their study. These authors further concluded that having 1000 or 2000 examinees is desirable for more accurate item parameter estimation.

Ankenmann & Stone (1992) also carried out a simulation study investigating parameter recovery for the GRM. These authors varied test length (5, 10, and 20 items), sample size (250, 500, and 1000 examinees), number of score categories for the items (3 and 5 categories), and ability distribution (normal and skewed positive). They found that for the GRM, samples sizes of 500 were the minimum that should be used for accurate item parameter estimation when normal ability distributions were assumed. The authors suggested that larger sample sizes are needed for skewed ability distributions.

De Ayala (1994) suggested a 5:1 ratio of examinees to item parameter estimates to allow for accurate parameter estimation based on tests consisting of 15 or 30 items. He further stated

that it is the distribution of responses across item categories, rather than only sample size itself, that impacts the accuracy of item parameter estimates.

II. C. 2. Ability Parameter Estimation for the GRM

In terms of ability parameter estimation, Reise & Yu (1990) found that for tests consisting of 25 items, sample size was not a major factor in producing accurate ability estimates under the GRM. The authors found that for tests of this length, sample sizes of 250 and 500 resulted in adequate ability parameter estimation. Ankenmann & Stone (1992), echoing the results of Reise & Yu, found that sample size and shape of ability distribution are not important factors in the estimation of ability parameters. However, these authors found that test length was a relevant factor in the estimation of ability. Specifically, they found that the accuracy of ability parameter estimates increased and the variability of the estimates decreased as test length increased.

In the present context, the precision of ability estimates is related to the extent to which classification errors impact the goodness-of-fit statistics used in assessing item fit for performance-based assessments (Ankenmann, 1994; Stone, 2000; Stone & Hansen, 2000; Stone et al., 1993; Stone et al., 1994). As a consequence of having shorter test lengths for these assessments, individual ability estimates will be less precise and more variable. The inaccuracy of ability estimates can result in classification errors in the item fit tables when traditional fit statistics are used. The increased variability in individual ability estimates also affects item and test information.

II. C. 3. Item and Test Information for the GRM

Item information has been shown to be useful for constructing tests, describing items and tests, and comparing tests (Hambleton, 1993). Item information functions display the contribution items make to ability estimation at points along the ability continuum. The contribution of a single item to the test can be seen through inspection of the item information function. In addition, the points along the ability continuum at which an item provides the most information about examinees can also be seen with this function.

For the GRM, the item information is the sum of the information that is provided by each of the response categories. The amount of information provided by each of the response categories at ability level θ is given by

$$I_k(\theta) = \frac{\{P_k'(\theta)\}^2 - P_k(\theta)P_k''(\theta)}{\{P_k(\theta)\}^2}, \tag{22}$$

where $P_k'(\theta)$ and $P_k''(\theta)$ are the first and second derivatives, respectively, of $P_k(\theta)$. The information for an item response category is a measure of how well the responses in that category estimate the examinee's ability (Baker, 1992). The amount of information share provided by category *k* of item *i*, which is the amount of information that category *k* contributes to item *i*, is defined to be

$$I_k(\theta)P_k(\theta) = \frac{\{P_k'(\theta)\}^2}{P_k(\theta)} - P_k''(\theta). \tag{23}$$

where the terms are as defined above.

As was stated, item information functions are useful for test development and for describing test items. For instance, assessing the shape of the information function over a range of abilities provides information on the types of examinees for which the test would be most

appropriate. If a test were to be used for high ability examinees, it would be desirable to have test items that provide most of their information in the high ability range.

The point on the ability scale at which an item provides its maximum information can also be useful in test construction. Perhaps more useful is the amount of information that is provided by an item over a range of ability values. By taking the integral of the information function over a specified range of abilities, the area under the information function over those abilities can be quantified into a single value. Obtaining the total area under the item information function over a specific range of abilities can be beneficial as this quantity provides a measure of item quality across that ability range.

Item information for the GRM is given by

$$I_i(\theta) = \sum_{k=1}^{m_i} \frac{\{P_k'(\theta)\}^2}{P_k(\theta)} - P_k''(\theta). \tag{24}$$

The formula for determining the area under the information function for a specified ability range, i.e., the range $-3 \leq \theta \leq 3$, is provided by Equation 25.

$$\int_{-3}^{3} I_i(\theta) d\theta = \int_{-3}^{3} \left[\sum_{k=1}^{m_i} \frac{\{P_k'(\theta)\}^2}{P_k(\theta)} - P_k''(\theta) \right] d\theta . \tag{25}$$

An approximation to this quantity is given by Equation 26,

$$\int_{-3}^{3} I_i(\theta) d\theta \cong \sum_{\theta_j=-3}^{3} \left\{ \left[\sum_{k=1}^{m_i} \frac{\{P_k'(\theta_j)\}^2}{P_k(\theta_j)} - P_k''(\theta) \right] \Delta\theta \right\}, \tag{26}$$

where θ_j takes on ascending values between –3 and 3 in small increments of $\Delta\theta$. All quantities in Equations 24 - 26 are as defined above.

Test information, or the contribution of the test toward ability estimation, is obtained by summing the item information functions for all items on a test. The test information at a point

along the ability continuum is inversely related to the precision with which ability is estimated at that point, as is seen by:

$$SE(\theta_o) = \frac{1}{\sqrt{I(\theta_o)}}, \tag{27}$$

where $SE(\theta_o)$ is the standard deviation of the distribution associated with estimates of ability for examinees having ability θ_o (Hambleton, 1993).

Test information is related to the variance of the posterior distribution of ability such that tests with lower information at a particular level of ability have higher posterior variances at that ability level. Introducing variation into the amount of information provided by items on a test therefore introduces variation in the posterior variances.

In addition, the quality of the test items also influences the standard errors. Smaller standard errors tend to be associated with highly discriminating items, and with items having difficulty parameters that match the ability parameters of the examinees (Hambleton, 1993). Varying the item parameters and the item information will together impact the posterior variances of ability. This is important to the current research because the methods of this study utilize the posterior distributions of ability.

CHAPTER III

III. METHODOLOGY

The methodology of the current study is presented in this chapter. The first section describes the data simulation aspect of the study. This includes discussion of the test configurations, item parameters, generation of item response data, and generation of sampling distributions. Following is a discussion related to the derivation of the scaling factor and degrees of freedom values (scaling corrections) from the empirical distributions. The steps involved in modeling and validating the prediction equations for predicting the scaling corrections from item and sample characteristics are then presented. Finally, the methodology for validating the use of the prediction equations is presented. This includes assessing the applicability of the prediction equations to real data sets containing items different from those in the simulation study.

III. A. Data Simulation

Empirical sampling distributions of χ^{2*} and G^{2*}, the Pearson and likelihood ratio forms of the pseudocounts-based fit statistic, respectively, were generated for each item within each test configuration. The sampling distributions were examined to provide evidence that both forms of the statistics were distributed as scaled chi-square random variables. The steps of this process were as follows:

1. Simulate realistic item response data given a multidimensional graded response model (MGRM) and model parameters for a test configuration.
2. Calibrate the simulated response data Samejima's (1969) unidimensional Graded Response Model (GRM) with MULTILOG.
3. Calculate the pseudocounts-based fit statistics for each item.
4. Repeat steps 1) to 3) 1000 times to generate sampling distributions of the pseudocounts-based fit statistics for each item.
5. Compute scaling corrections for each sampling distribution using the means and variances of the sampling distributions with the method of moments, using Equations 19 and 20,
6. Rescale the empirically generated sampling distributions using the scaling corrections given by Equations 19, and
7. Compare the rescaled empirical distributions to theoretical chi-square distributions with degrees of freedom given by Equation 20 using the slopes and intercepts of regression lines fitted to Quantile-Quantile (Q-Q) plots, and Type I error rates.
8. Repeat steps 1) to 7) for each test configuration.

In addition, the mean of the posterior variances for individual item response vectors was computed across replications for each test configuration. This value then served as a predictor variable in a later stage of this study.

The data simulation aspect of this study provided verification that the scaling corrections obtained from empirical data could be applied to the sampling distributions, and one of the family of theoretical chi-square distributions used for significance testing of the fit statistics.

Details relating to the test configurations, item parameters, and generation of item response data and sampling distributions follow.

III. A. 1. Test Configurations

Two factors were manipulated in this study: test length and number of item score categories. In particular, three levels of test length (12, 24, and 36 items) and four item score category levels (2, 3, 4, and 5) were considered.

Test length was chosen as a factor in this study in order to manipulate the precision of ability estimation and the variance of the estimated posterior ability distributions. Varying test length introduces variability in the accuracy of individual ability estimates. This in turn affects the spread of the posterior distributions of ability, and hence the posterior expectations (pseudocounts) found in Table 2. Imprecise ability estimates obtained from shorter tests lead to the examinees' pseudo-observed score distributions being spread out over a wide range of the ability scale. More precise estimates of ability lead to this distribution being more concentrated.

Previous studies (Ankenmann, 1994; Stone et al., 1994) suggested that the scaling corrections were strongly related to the mean of the variances of estimated posterior ability distributions. Since the mean posterior variance was used as a predictor variable for the scaling corrections in this study as well, it was important to introduce variation into the posterior ability distributions.

Three different tests lengths (12, 24, and 36 items) were investigated. However, in order to include a wide range of item parameters in the study, three 12 item tests (12a, 12b, and 12c) were designed at each item score category level. A total of 5 levels of the test length variable (12a, 12b, 12c, 24, and 36 items) were investigated for each item score category level.

A length of 12 items for a single test was chosen to be representative of a typical performance-based assessment, or an assessment such as NAEP that utilizes matrix-sampling methods. The 24 item tests consisted of the combination of the two 12 item tests labeled 12a and 12b at each score category level. Doubling the test length variable from 12 to 24 items allowed the effects of a longer test length on the posterior distribution of ability to be examined. Further, 24 item tests are consistent with some NAEP assessments as well as with other assessments consisting of mixtures of dichotomously and polytomously scored items.

The 36 item tests consisted of items in the three 12 item tests, labeled 12a, 12b, and 12c, at each item score category level. Including the test length of 36 items in the simulation study allowed the behavior of χ^{2*} and G^{2*} to be evaluated in cases where the individual ability estimates were more precise. While some conditions involving 36 item tests may be unrealistic in practice (e.g., a 36 item test consisting entirely of 5-category items), these conditions served as upper bounds on practical situations. They also allowed the behavior of the procedures in this study to be investigated across a wide range of conditions. For these reasons, a completely crossed design was chosen, rather than a design that excluded some of the less realistic conditions.

The number of item score categories was manipulated for several reasons. The number of score categories factors into the computation of the degrees of freedom for traditional chi-square statistics. It was expected that this variable would serve as an important predictor for the degrees of freedom for the statistic used in the current study as well. Also, performance-based assessments and assessments such as NAEP often consist of items with varying numbers of score categories. Prediction of the scaling corrections needed for significance testing of the G^{2*} fit statistics had only been investigated for 5-category items under the GRM (Ankenmann, 1994).

Thus, it was necessary to investigate the effect that varying this variable has on the methodology of the current study. Item category levels of 2- though 5-categories were chosen because they are representative of performance-based items used in practice.

Sample size was not chosen as a factor in the current study. Selection of a sample size of 2000 in the current study allowed for precise item parameter estimation. In addition, using samples sizes of 2000 eliminated some of the sparseness in the item fit tables. However, by not varying sample size, any effects of this variable on the sensitivity of the pseudocounts-based fit statistics and the quality of the prediction equations were not evaluated.

The number of ability subgroups created was also not chosen as an independent variable in this study. This factor had been manipulated in past studies investigating the distribution of G^{2*} (Ankenmann, 1994; Stone et al., 1994). It had been hypothesized in these two studies that ability interval width would be needed as a predictor for the degrees of freedom values, because changing the ability interval width changes the number of rows in the item fit tables. However, the studies found that the number of ability subgroups did not contribute to the prediction of the appropriate scaling corrections. Therefore, ability interval width was not manipulated in the current study.

III. A. 2. Item Parameters

Tables 3 – 6 list the item parameters, average difficulty, and item information for the 2-category, 3-category, 4-category, and 5-category items, respectively. The parameters for three sets of 12 items are listed in each table, with the first 12 items in each table comprising the test to be labeled 12a, the second set of 12 items comprising the test 12b, and the third set of 12 items comprising the test 12c. The discrimination and threshold parameters for each item are provided.

An overall measure of the item difficulty is provided in Tables 4 - 6. The measure of overall item difficulty is provided for the 3-, 4-, and 5-category items. The numeric value that is provided is the ability level at which the expected score on the item divided by the number of possible points for the item is equal to 0.5. That is, the ability value at which examinees are most likely to receive half of the possible score points on the item. This value provides a single measure of the item difficulty for the multiple category items that is analog to the difficulty parameter for 2-category items. The total information provided by each item, given by Equation 26 with $\Delta\theta = 0.05$, is also provided in Tables 3 – 6.

All item parameters were chosen to be representative of parameters obtained in real test settings, namely the QCAI (Lane, 1993; Stone et al., 1993) and NAEP (U.S. Department of Education, 1999; U.S. Department of Education, 2001). The item parameters are also representative of items used in previously published simulation studies (Ankenmann, Witt, & Dunbar, 1999; Ankenmann & Stone, 1992; Cohen & Kim, 1998; Cohen, Kim, & Baker, 1993; Kim & Cohen, 1998**;** Stone, 2003; Stone & Hansen, 2000).

Table 3. Item Parameters of the Two Category Items for the Three 12 Item Tests

Item	*a*	b_1	*Info*
1a-2Cat	0.7	-1.5	0.493
2a-2Cat	1.0	-1.5	0.811
3a-2Cat	1.4	-1.5	1.250
4a-2Cat	1.7	-1.5	1.581
5a-2Cat	2.1	-1.5	2.018
6a-2Cat	2.4	-1.5	2.340
7a-2Cat	0.7	1.0	0.524
8a-2Cat	1.0	1.0	0.866
9a-2Cat	1.4	1.0	1.317
10a-2Cat	1.7	1.0	1.646
11a-2Cat	2.1	1.0	2.070
12a-2Cat	2.4	1.0	2.381
1b-2Cat	0.7	-0.5	0.543
2b-2Cat	1.0	-0.5	0.897
3b-2Cat	1.4	-0.5	1.350
4b-2Cat	1.7	-0.5	1.673
5b-2Cat	2.1	-0.5	2.088
6b-2Cat	2.4	-0.5	2.394
7b-2Cat	0.7	0.5	0.543
8b-2Cat	1.0	0.5	0.897
9b-2Cat	1.4	0.5	1.350
10b-2Cat	1.7	0.5	1.673
11b-2Cat	2.1	0.5	2.088
12b-2Cat	2.4	0.5	2.394
1c-2Cat	0.7	-1.0	0.524
2c-2Cat	1.0	-1.0	0.866
3c-2Cat	1.4	-1.0	1.317
4c-2Cat	1.7	-1.0	1.646
5c-2Cat	2.1	-1.0	2.070
6c-2Cat	2.4	-1.0	2.381
7c-2Cat	0.7	1.5	0.493
8c-2Cat	1.0	1.5	0.811
9c-2Cat	1.4	1.5	1.250
10c-2Cat	1.7	1.5	1.581
11c-2Cat	2.1	1.5	2.018
12c-2Cat	2.4	1.5	2.340

Table 4. Item Parameters of the Three Category Items for the Three 12 Item tests

Item	*a*	b_1	b_2	*Average b*	*Info*
1a-3Cat	0.7	-2.5	1.5	-0.5	0.753
2a-3Cat	1.0	-2.5	1.5	-0.5	1.334
3a-3Cat	1.4	-2.5	1.5	-0.5	2.150
4a-3Cat	1.7	-2.5	1.5	-0.5	2.766
5a-3Cat	2.1	-2.5	1.5	-0.5	3.588
6a-3Cat	2.4	-2.5	1.5	-0.5	4.207
7a-3Cat	0.7	-1.5	2.5	0.5	0.753
8a-3Cat	1.0	-1.5	2.5	0.5	1.334
9a-3Cat	1.4	-1.5	2.5	0.5	2.150
10a-3Cat	1.7	-1.5	2.5	0.5	2.766
11a-3Cat	2.1	-1.5	2.5	0.5	3.588
12a-3Cat	2.4	-1.5	2.5	0.5	4.207
1b-3Cat	0.7	-1.5	1.5	0.0	0.753
2b-3Cat	1.0	-1.5	1.5	0.0	1.334
3b-3Cat	1.4	-1.5	1.5	0.0	2.150
4b-3Cat	1.7	-1.5	1.5	0.0	2.766
5b-3Cat	2.1	-1.5	1.5	0.0	3.588
6b-3Cat	2.4	-1.5	1.5	0.0	4.207
7b-3Cat	0.7	-2.0	2.0	0.0	0.760
8b-3Cat	1.0	-2.0	2.0	0.0	1.357
9b-3Cat	1.4	-2.0	2.0	0.0	2.214
10b-3Cat	1.7	-2.0	2.0	0.0	2.871
11b-3Cat	2.1	-2.0	2.0	0.0	3.756
12b-3Cat	2.4	-2.0	2.0	0.0	4.420
1c-3Cat	0.7	-2.0	1.0	-0.5	0.760
2c-3Cat	1.0	-2.0	1.0	-0.5	1.393
3c-3Cat	1.4	-2.0	1.0	-0.5	2.314
4c-3Cat	1.7	-2.0	1.0	-0.5	3.008
5c-3Cat	2.1	-2.0	1.0	-0.5	3.911
6c-3Cat	2.4	-2.0	1.0	-0.5	4.570
7c-3Cat	0.7	-1.0	2.0	0.5	0.760
8c-3Cat	1.0	-1.0	2.0	0.5	1.393
9c-3Cat	1.4	-1.0	2.0	0.5	2.314
10c-3Cat	1.7	-1.0	2.0	0.5	3.008
11c-3Cat	2.1	-1.0	2.0	0.5	3.911
12c-3Cat	2.4	-1.0	2.0	0.5	4.570

Table 5. Item Parameters of the Four Category Items for the Three 12 Item Tests

Item	*a*	b_1	b_2	b_3	*Average b*	*Info*
1a-4Cat	0.7	-2.5	-0.5	1.5	-0.5	0.842
2a-4Cat	1.0	-2.5	-0.5	1.5	-0.5	1.624
3a-4Cat	1.4	-2.5	-0.5	1.5	-0.5	2.878
4a-4Cat	1.7	-2.5	-0.5	1.5	-0.5	3.896
5a-4Cat	2.1	-2.5	-0.5	1.5	-0.5	5.277
6a-4Cat	2.4	-2.5	-0.5	1.5	-0.5	6.304
7a-4Cat	0.7	-1.5	0.5	2.5	0.5	0.842
8a-4Cat	1.0	-1.5	0.5	2.5	0.5	1.624
9a-4Cat	1.4	-1.5	0.5	2.5	0.5	2.878
10a-4Cat	1.7	-1.5	0.5	2.5	0.5	3.896
11a-4Cat	2.1	-1.5	0.5	2.5	0.5	5.277
12a-4Cat	2.4	-1.5	0.5	2.5	0.5	6.304
1b-4Cat	0.7	-1.5	0.0	1.5	0.0	0.812
2b-4Cat	1.0	-1.5	0.0	1.5	0.0	1.572
3b-4Cat	1.4	-1.5	0.0	1.5	0.0	2.815
4b-4Cat	1.7	-1.5	0.0	1.5	0.0	3.846
5b-4Cat	2.1	-1.5	0.0	1.5	0.0	5.274
6b-4Cat	2.4	-1.5	0.0	1.5	0.0	6.351
7b-4Cat	0.7	-2.0	0.0	2.0	0.0	0.850
8b-4Cat	1.0	-2.0	0.0	2.0	0.0	1.648
9b-4Cat	1.4	-2.0	0.0	2.0	0.0	2.943
10b-4Cat	1.7	-2.0	0.0	2.0	0.0	4.002
11b-4Cat	2.1	-2.0	0.0	2.0	0.0	5.446
12b-4Cat	2.4	-2.0	0.0	2.0	0.0	0.812
1c-4Cat	0.7	-2.0	-0.5	1.0	-0.5	6.516
2c-4Cat	1.0	-2.0	-0.5	1.0	-0.5	0.803
3c-4Cat	1.4	-2.0	-0.5	1.0	-0.5	1.548
4c-4Cat	1.7	-2.0	-0.5	1.0	-0.5	2.764
5c-4Cat	2.1	-2.0	-0.5	1.0	-0.5	3.777
6c-4Cat	2.4	-2.0	-0.5	1.0	-0.5	5.190
7c-4Cat	0.7	-1.0	0.5	2.0	0.5	6.264
8c-4Cat	1.0	-1.0	0.5	2.0	0.5	0.803
9c-4Cat	1.4	-1.0	0.5	2.0	0.5	1.548
10c-4Cat	1.7	-1.0	0.5	2.0	0.5	2.764
11c-4Cat	2.1	-1.0	0.5	2.0	0.5	3.777
12c-4Cat	2.4	-1.0	0.5	2.0	0.5	5.190

Table 6. Item Parameters of the Five Category Items for the Three 12 Item tests

Item	*a*	b_1	b_2	b_3	b_4	*Average b*	*Info*
1a-5Cat	0.7	-2.5	-1.17	0.17	1.5	-0.5	0.862
2a-5Cat	1.0	-2.5	-1.17	0.17	1.5	-0.5	1.703
3a-5Cat	1.4	-2.5	-1.17	0.17	1.5	-0.5	3.138
4a-5Cat	1.7	-2.5	-1.17	0.17	1.5	-0.5	4.382
5a-5Cat	2.1	-2.5	-1.17	0.17	1.5	-0.5	6.177
6a-5Cat	2.4	-2.5	-1.17	0.17	1.5	-0.5	7.580
7a-5Cat	0.7	-1.5	-0.17	1.17	2.5	0.5	0.862
8a-5Cat	1.0	-1.5	-0.17	1.17	2.5	0.5	1.703
9a-5Cat	1.4	-1.5	-0.17	1.17	2.5	0.5	3.138
10a-5Cat	1.7	-1.5	-0.17	1.17	2.5	0.5	4.382
11a-5Cat	2.1	-1.5	-0.17	1.17	2.5	0.5	6.177
12a-5Cat	2.4	-1.5	-0.17	1.17	2.5	0.5	7.580
1b-5Cat	0.7	-1.5	-0.5	0.5	1.5	0.0	0.821
2b-5Cat	1.0	-1.5	-0.5	0.5	1.5	0.0	1.609
3b-5Cat	1.4	-1.5	-0.5	0.5	1.5	0.0	2.944
4b-5Cat	1.7	-1.5	-0.5	0.5	1.5	0.0	4.101
5b-5Cat	2.1	-1.5	-0.5	0.5	1.5	0.0	5.787
6b-5Cat	2.4	-1.5	-0.5	0.5	1.5	0.0	7.124
7b-5Cat	0.7	-2.0	-.67	0.67	2.0	0.0	0.870
8b-5Cat	1.0	-2.0	-.67	0.67	2.0	0.0	1.728
9b-5Cat	1.4	-2.0	-.67	0.67	2.0	0.0	3.204
10b-5Cat	1.7	-2.0	-.67	0.67	2.0	0.0	4.490
11b-5Cat	2.1	-2.0	-.67	0.67	2.0	0.0	6.346
12b-5Cat	2.4	-2.0	-.67	0.67	2.0	0.0	7.793
1c-5Cat	0.7	-2.0	-1.0	0.0	1.0	-0.5	0.812
2c-5Cat	1.0	-2.0	-1.0	0.0	1.0	-0.5	1.585
3c-5Cat	1.4	-2.0	-1.0	0.0	1.0	-0.5	2.893
4c-5Cat	1.7	-2.0	-1.0	0.0	1.0	-0.5	4.032
5c-5Cat	2.1	-2.0	-1.0	0.0	1.0	-0.5	5.703
6c-5Cat	2.4	-2.0	-1.0	0.0	1.0	-0.5	7.037
7c-5Cat	0.7	-1.0	0.0	1.0	2.0	0.5	0.812
8c-5Cat	1.0	-1.0	0.0	1.0	2.0	0.5	1.585
9c-5Cat	1.4	-1.0	0.0	1.0	2.0	0.5	2.893
10c-5Cat	1.7	-1.0	0.0	1.0	2.0	0.5	4.032
11c-5Cat	2.1	-1.0	0.0	1.0	2.0	0.5	5.703
12c-5Cat	2.4	-1.0	0.0	1.0	2.0	0.5	7.037

Six different discrimination parameters for items comprising the 12 item tests at each item score category level were selected. The discrimination parameters were paired with 2 different sets of difficulty parameters within each test. Ankenmann (1994) found that the discrimination parameter was useful in predicting the appropriate scaling factor for the sampling distributions, so items with varying discrimination parameters were selected for the current study. Because the threshold parameters could also contribute to the prediction of the scaling corrections, some variation in these values was also introduced.

By design, a wide range of difficulty levels (-2 to 2 in the case of two category items) was crossed with a wide range of discrimination levels. This provided unique combinations of difficulty and discrimination parameters for all test conditions. It also provided a diverse set of items for generating the prediction equations. It was hoped that basing the prediction equations on this diverse set of items would allow them to generalize to a diverse set of tests.

In the GRM, a scaling constant D is absorbed into the estimation of the slope parameters. In MULTILOG, the value of D is set equal to 1.0. The discrimination values used in the current study of a = 0.7, 1.0, 1.4, 1.7, 2.1, and 2.4, with D = 1.0, correspond to discrimination parameters of a = 0.412, 0.588, 0.824, 1.0, 1.235, and 1.412 when D = 1.7.

The amount of item information provided by the items was also a consideration for their inclusion in the study. This is because item information affects the variance of the estimated posterior distributions of ability. As was discussed, test information is inversely related to the standard error of the ability estimates. The higher the test information, the lower the posterior variances, and vice versa. Test information at a specific level of ability is given by the sum of the item information functions at that ability level. So, varying the item information introduced variability into the posterior distributions of ability.

Items providing varying amounts of information across the ability scale were included in this study. In general, items providing information across wide ranges of the ability scale were selected. For example, plots of the item information functions for 4 items can be found in Figures 3 - 6.

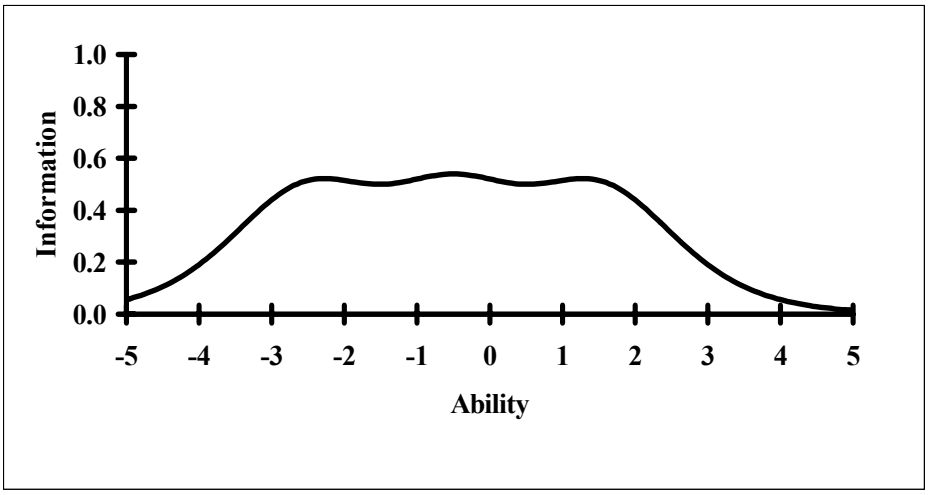

Figure 3. Item Information Function for 2-Category Item 3a-2cat

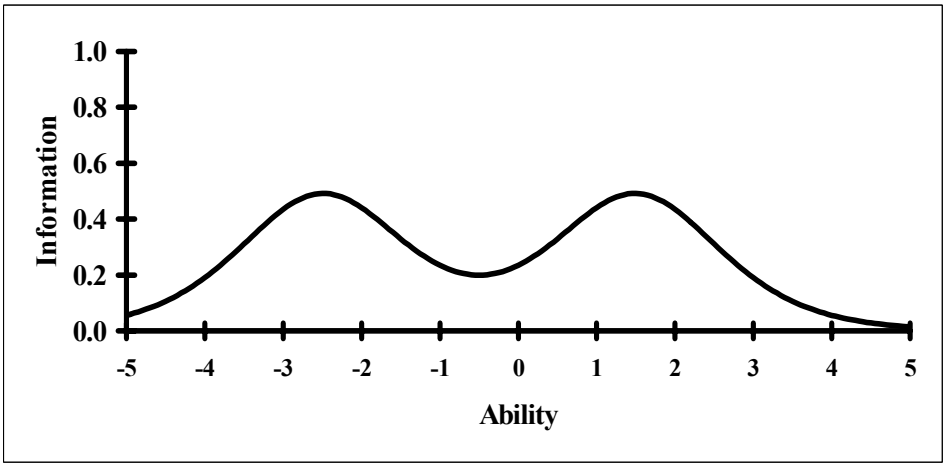

Figure 4. Item Information Function for 3-Category Item 3a-3cat

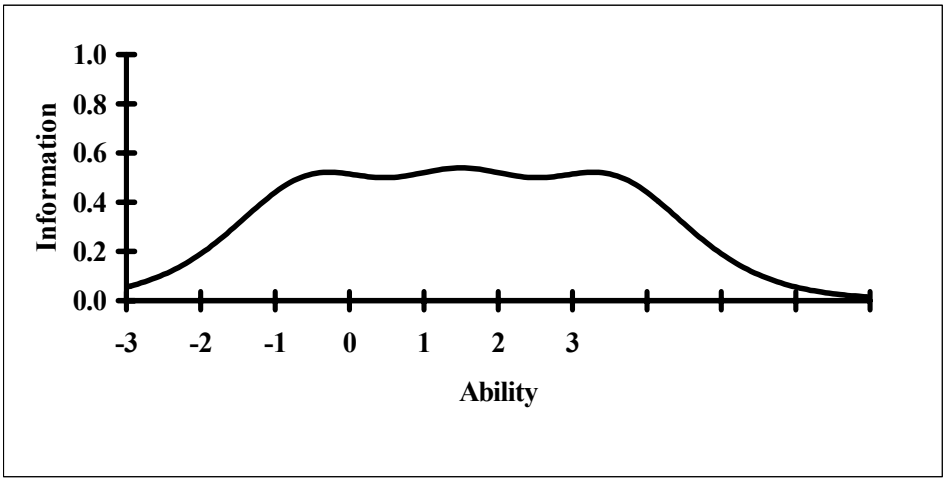

Figure 5. Item Information Function for 4-Category Item 3a-4cat

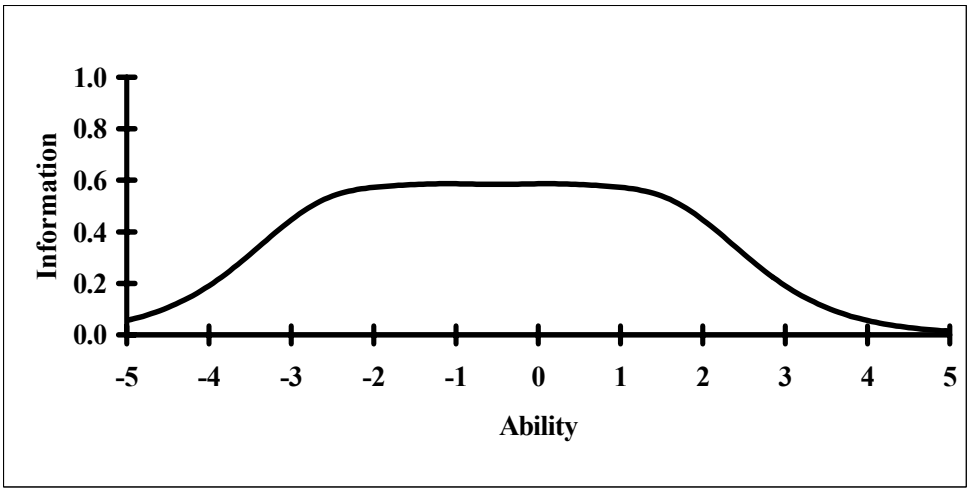

Figure 6. Item Information Function for 5-Category Item 3a-5Cat

The three 12 item tests at each item score category level were constructed to provide the bulk of their information centered on the ability value of zero. This is because in testing applications that use IRT, it is commonly assumed that examinees' abilities follow a normal distribution with a mean of zero and a standard deviation of one.

III. A. 3. Data Generation

Item response data for this study were simulated under the multidimensional Graded Response Model (MGRM) for each of the test configurations in the study. Data that fit the MGRM having one main and five additional nuisance dimensions were generated.

III. A. 3. a) Multidimensional Item Response Theory

Multidimensional Item Response Theory (MIRT) models have been developed to model the case where multiple traits are estimated by a single test. Many high-quality tests have been shown to measure a single, dominant ability, meeting the unidimensionality assumption in practice. Strict unidimensionality, however, is often not met even when evidence is found that tests measure a single trait, because these single ability dimensions typically account for less than half of the score variance (Davey et al., 1997). In simulation studies, when data are generated under the assumption of unidimensionality, the secondary factors present in real test data are ignored. Davey et al. argue that such differences between real and simulated test data "may lead to certain procedures evaluated as effective with simulated data performing much less well when applied to real data (p. 7)."

Item response data in the current study were generated to have one main dimension, and five additional minor dimensions. This data were intended to represent realistic item responses that in practice would be considered to have met the unidimensionality assumption. Including the additional minor dimensions allowed the complex structure of real item responses to be maintained, as suggested by Davey et al. (1997). Studies have shown that if a test exhibits only minor deviations from unidimensionality, unidimensional IRT models can be used appropriately for item analysis (Childs and Oppler, 2000). Further, Reckase (1985) stated that unidimensional

statistics were appropriate for items that measure one main dimension and only minor additional dimensions.

MIRT models can be classified as either compensatory or noncompensatory. In general, compensatory models have an item discrimination parameter and ability parameter for each of the multiple dimensions, but only one set of threshold parameters (assuming the polytomous case) across the dimensions. Noncompensatory models allow the threshold parameters as well as the discrimination parameters to vary across dimensions. The compensatory form of the MGRM was used in the current study.

III. A. 3. b) Multidimensional Graded Response Model

A multidimensional extension of Samejima's (1969) GRM was developed and follows directly from Equation 21 (Hambleton, 1993; Muraki & Carlson, 1993; Reckase, 1985; De Ayala, 1994). The compensatory form of the MGRM is expressed as

$$P_{ik}^{*}(\Theta) = \frac{1}{1 + exp\left[- D \sum_{h=1}^{r} a_{ih} \left(\theta_h - b_{ki} \right) \right]}, \tag{28}$$

where i is the item number,

$k = 0, 1, \ldots m_i$ is the response category,

$m_i + 1$ is the number of ordered response categories for item i,

D is 1 or a scaling constant of 1.702 that may be introduced,

Θ is a vector of latent trait (ability) parameters,

θ_h is the latent trait parameter on dimension h ($h = 1, \ldots, r$ dimensions),

a_{ih} is the discrimination parameter for item i, and

b_{ki} is the threshold parameter for category k of item i.

As with the GRM, in the MGRM the values of $P_{ik}^{*}(\Theta)$ represent the probability of a randomly selected examinee with latent traits Θ responding in category *k* or higher for item *i*. Also, in order to find the probability that an examinee responds to a specific category *k* of item *i*, one must compute the difference between the cumulative probabilities for two adjacent categories. If the total number of dimensions, *r*, is equal to one, the MGRM reduces to the GRM. For a single item under the compensatory MGRM, the values a_{ih} can vary across dimensions, but the values of b_{ki} remain constant across dimensions. There is one ability parameter for each dimension.

III. A. 3. c) Generation of Item Response Data

Item response data were generated under the MGRM with one main and five minor higher-order dimensions using code written in the statistical package SAS (Version 8.02). The program used to generate the item response data is found in Appendix A. The discrimination (*a*) parameters for only the single main dimension are listed in Tables 3 – 6. The *a* parameters for the additional minor dimensions for each item were chosen by setting an upper limit of 0.4 for each value, and randomly selecting the *a* values from a uniform distribution [U(0, 0.4)]. A maximum value of 0.4 was selected to represent low item discrimination parameters for each of the minor dimensions.

The empirical generation of item response data sets involved the following steps for each examinee:

1. Randomly generate a set of ability parameters from a Multivariate Normal MVN(0,1) ability distribution;

2. Use these ability values and the item parameters defined in Tables 3- 6 with the MGRM to calculate the probabilities of a response being classified into each of the possible score category levels, with the scaling constant D set equal to 1;
3. Compute the cumulative probabilities for each item score category, which are the probabilities that the response falls into a particular category or lower;
4. Generate a random number from a uniform U(0,1) distribution.
5. Assign the item response for item i (having $m_i + 1$ response categories labeled $k = 0$ to m_i), as k if the randomly generated uniform random variable was greater than or equal to the cumulative probability for category k, but smaller than the cumulative probability for category $k + 1$.

Data sets of examinee responses for each test configuration were generated in this manner.

III. A. 3. d) Validation of Item Response Data

Validation that the item response data generated for this study fit the intended model was necessary to ensure the accuracy of the conclusions of the study. This process involved validating that the item parameter estimates obtained from the simulated response data matched the item parameters used to generate the data sets, and also that data sets having one main dimension were generated.

(1) Validation of Item Parameter Estimates

Several steps were taken to validate the item response data generated in the study. Initially, two validation data sets were generated. The data sets consisted of responses to 12 2- and 5-category items, respectively, for 2000 subjects. The items chosen were randomly selected from the pool of 2- and 5- category items used in the study. The item response data sets were

generated under the GRM using the SAS program utilized in the study. The data sets were generated using the item parameters from the selected items and a constant ability value of 0. The item parameters for the two validation data sets are found in Tables 7 and 8.

Table 7. Item Parameters of the Two-Category Items Selected for Item Validation

Item Name	*a*	b_1	*Info*
2a-2Cat	1.0	-1.5	0.811
9a-2Cat	1.4	1.0	1.317
10a-2Cat	1.7	1.0	1.646
5b-2Cat	2.1	-0.5	2.088
6b-2Cat	2.4	-0.5	2.394
8b-2Cat	1.0	0.5	0.897
11b-2Cat	2.1	0.5	2.088
1c-2Cat	0.7	-1.0	0.524
4c-2Cat	1.7	-1.0	1.646
6c-2Cat	2.4	-1.0	2.381
9c-2Cat	1.4	1.5	1.250
12c-2Cat	2.4	1.5	2.340

Table 8. Item Parameters of the Five-Category Items Selected for Item Validation

Item Name	*a*	b_1	b_2	b_3	b_4	*Average b*	*Info*
5a-5cat	2.1	-2.5	-1.17	0.17	1.5	-0.5	6.117
10a-5cat	1.7	-1.5	-0.17	1.17	2.5	0.5	4.382
11a-5cat	2.1	-1.5	-0.17	1.17	2.5	0.5	6.177
12a-5cat	2.4	-1.5	-0.17	1.17	2.5	0.5	6.753
1b-5cat	0.7	-1.5	-0.50	0.50	1.5	0.0	0.821
5b-5cat	2.1	-1.5	-.50	0.50	1.5	0.0	5.787
8b-5cat	1.0	-2.0	-.67	0.67	2.0	0.0	1.728
10b-5cat	1.7	-2.0	-.67	0.67	2.0	0.0	4.490
2c-5cat	1.0	-2.0	-1.0	0.0	1.0	-0.5	1.585
5c-5cat	2.1	-2.0	-1.0	0.0	1.0	-0.5	5.703
7c-5cat	0.7	-1.0	0.0	1.0	2.0	0.5	0.812
9c-5cat	1.4	-1.0	0.0	1.0	2.0	0.5	2.893

To validate the data generation procedures, the proportion of examinees responding in each of the possible item score categories for each item was computed. These observed proportions were then compared to the model-based expected proportions, found using the original item parameters and the constant ability level of 0. This comparison showed how well the data generated under the GRM followed that model.

Comparisons between the observed and expected proportions were made across the possible item score categories for each item. Differences were expected to be small if the data generation procedures were valid. As outlined, this process was carried out 2 times, once for the lowest score category level, and once for the highest score category level, in order to confirm that the number of item categories had no effect on the quality of data generation in the current study.

This process was then repeated by generating data under the MGRM with one main and 5 minor dimensions. Comparisons between the observed and expected proportions across examinees for each multidimensional data set were examined. It was expected that more variation between the observed and expected proportions of examinees would be observed with the multidimensional data. The amount of additional variation was expected to be small because the multidimensional data had only one main dimension.

Table 9 presents the absolute differences between the observed and expected proportions of examinees answering each item correctly for the simulated unidimensional and multidimensional 2-category data sets. Table 9 shows that for the 2-category unidimensional data set, the largest absolute difference between the observed and expected proportions was 0.010. The average absolute difference was 0.006. For the multidimensional 2-category

validation data set, the largest absolute difference between the observed and expected proportions was .072, and the average was 0.060.

The small differences for the unidimensional data indicate that the data generation program generated simulated item response data that followed the GRM for 2-category items. The increased differences for the multidimensional data show that variation was introduced, in order to reflect real item responses.

Table 9. Absolute Differences between Observed and Expected Proportions For the 12-Item 2-Category Unidimensional and Multidimensional Validation Data Sets

Item Name	Unidimensional	Multidimensional
2a-2Cat	0.001	0.093
9a-2Cat	0.010	0.072
10a-2Cat	0.005	0.068
5b-2Cat	0.010	0.054
6b-2Cat	0.010	0.047
8b-2Cat	0.004	0.064
11b-2Cat	0.004	0.053
1c-2Cat	0.008	0.118
4c-2Cat	0.008	0.050
6c-2Cat	0.005	0.027
9c-2Cat	0.004	0.061
12c-2Cat	0.003	0.008

Tables 10 shows the absolute differences between the observed and expected proportions of examinees at each response category level for the simulated unidimensional and multidimensional 5-category validation data sets. For the unidimensional data, Table 10 shows

that the largest absolute difference between the observed and expected proportions was 0.013. The average absolute difference was 0.002. For the 5-category multidimensional validation data set, the largest absolute difference between the observed and expected proportions was 0.156, and the average difference was .023.

The small differences for the unidimensional 5-category data set indicate that the data generation program generated simulated item response data that followed the GRM for 5-category items. Introducing multidimensionality into the item response data introduced variation into the item response data, which was meant to reflect real item response data.

Table 10. Absolute Differences Between Observed and Expected Proportions For the 12-Item 5-Category Unidimensional and Multidimensional Validation Data Sets

	Unidimensional					Multidimensional				
Item Name	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5
5a-5cat	0.001	0.001	0.002	0.002	0.003	0.004	0.033	0.056	0.001	0.018
10a-5cat	0.002	0.002	0.006	0.000	0.002	0.019	0.004	0.065	0.037	0.005
11a-5cat	0.003	0.006	0.006	0.005	0.001	0.038	0.009	0.083	0.041	0.013
12a-5cat	0.000	0.004	0.006	0.001	0.001	0.011	0.008	0.053	0.031	0.002
1b-5cat	0.001	0.013	0.012	0.001	0.001	0.127	0.053	0.142	0.055	0.123
5b-5cat	0.002	0.006	0.004	0.007	0.003	0.019	0.023	0.099	0.030	0.027
8b-5cat	0.007	0.010	0.003	0.005	0.004	0.077	0.002	0.156	0.007	0.074
10b-5cat	0.001	0.010	0.006	0.007	0.003	0.018	0.040	0.128	0.045	0.024
2c-5cat	0.007	0.008	0.002	0.002	0.002	0.077	0.013	0.084	0.106	0.100
5c-5cat	0.002	0.011	0.003	0.005	0.006	0.006	0.035	0.055	0.037	0.050
7c-5cat	0.001	0.001	0.002	0.001	0.003	0.105	0.099	0.118	0.007	0.119
9c-5cat	0.013	0.002	0.011	0.002	0.002	0.074	0.087	0.063	0.034	0.042

(2) Validation of Factor Structure

As another type of data validation, the factor structures of two validation data sets generated under the MGRM were examined using MPLUS Version 1.04 (Muthen & Muthen, 1998). It was expected that the underlying factor structures of these data sets would have only one main dimension. Thus, it was expected that the eigenvalue for the single main dimension would be large, and eigenvalues for any additional dimensions would be small.

To investigate the factor structure of the simulated multidimensional data, one 12-item, 2-category data set and one 12-item 5-category data set consisting of item responses for 2000 examinees were simulated under the MGRM using the item parameters found in Tables 7 and 8. For the 2-category data set, the largest eigenvalue was 5.996, and all others were less than 1. For the 5-category data set, the largest eigenvalue was 5.918, and all other eigenvalues were less than 1. Thus, the incorporation of nuisance dimensions to the item response data did not affect the overall dimensionality of the data set.

III. A. 4. Empirical Sampling Distributions

Empirically generated sampling distributions of the fit statistics were generated for each test configuration. The generation of these distributions for each item involved replications of the following steps: (a) simulate item response data for 2000 examinees under the MGRM, using item parameters found in Tables 3 - 6 and randomly selected θ [MVN(0,1)]; (b) calibrate the simulated item response data under the GRM using MULTILOG; (c) calculate the χ^{2*} and G^{2*} fit statistic for each item; and (d) repeat steps (a) through (c) 1000 times to form the sampling distributions of the fit statistics.

III. A. 4. a) Selection of Pearson Versus Likelihood Ratio Fit Statistic

Both the Pearson form of the pseudocounts-based fit statistic, χ^{2*}, and the likelihood ratio form of the statistic, G^{2*}, were investigated in the current study. Several factors were considered in making the decision to investigate both forms of the statistic. Literature relating to chi-square tests is somewhat inconclusive in the determination of the best overall item fit statistic. Collins et al. (1993) state that in a variety of situations, the distribution of χ^2 is more like a chi-square random variable than is G^2. At the same time, the authors caution that in some cases involving latent class models, neither statistic follows the chi-square distribution. Yen (1981) reported adequate performance for the Q_1 statistic with respect to decisions of item fit, and concluded that χ^2_B would perform similarly. McKinley and Mills (1985) could not choose one statistic as being the best for making decisions of item fit based on the results of their simulation study. Finally, Stone and Hansen (2000) found similar results for Pearson and likelihood ratio based statistics in terms of their null sampling distributions, although they found slightly less consistent results for G^2.

For studies specifically involving the pseudocounts-based statistic, Stone (2003) found similar Type I error rates for G^{2*} and χ^{2*}. In addition, he found that the power of χ^{2*} was slightly higher than that for G^{2*}. Stone and Zhang (2003) found that χ^{2*} exhibited nominal level Type I error rates and adequate power for tests consisting of dichotomous items. The studies by Stone et al. (1994) and Ankenmann (1994) both attempted to predict the scaling corrections of the G^{2*} statistics. Other researchers such as Donoghue and Hombo (2001a, 2001b) have investigated use of the χ^{2*} form of the statistic in their research.

Using these studies as a base, it does not appear that one statistic, G^{2*} or χ^{2*}, is consistently best. In the current study, χ^{2*} and G^{2*} were used to assess item fit. Although χ^{2*}

was shown to perform better than or similarly to G^{2*} (Stone, 2003), preliminary results of the current study showed that the chi-square distribution provided a closer approximation to the distribution of G^{2*} than χ^{2*}. Because the prediction technique had not been investigated using χ^{2*}, both forms of the statistic were considered.

III. A. 4. b) Number of Replications

The choice of generating 1000 fit statistics for each sampling distribution was made after considering several factors. Harwell, et al. (1996) stated that if sampling distributions are generated, a large number of repetitions may be needed to assess the validity of the IRT technique being evaluated. Stone et al. (1994) reported some variability in the estimates of the scaling factor and degrees of freedom values when sampling distributions of 500 fit statistics were generated using the Rasch model, and Stone et al. (1993) reported variability in these estimates for sampling distributions consisting of 1000 fit statistics with the GRM. Ankenmann (1994) generated sampling distributions consisting of 4000 fit statistics.

Due to the variability in the number of replications in previous studies that investigated the pseudocounts-based fit statistic, several steps were taken to determine the most appropriate number of replications for the current study. Under a condition of the study, namely the 5-category case with 12 items, item response data sets were simulated for 2000 examinees. Then, sampling distributions consisting of 2000 χ^{2*} values were generated. This condition was selected because it was the test configuration that would provide the least accurate parameter estimates. The 12 items chosen for this preliminary step of the study were randomly selected from the pool of 36 5-category items to be used in the simulation study, and can be found in Table 8.

For each item, the means and variances of the sampling distributions were calculated for every addition of 200 replications of χ^{2*} fit statistics. The values were then examined in order to

determine the number of replications needed to stabilize these two statistics. Tables 11, 12, and 13 show the means, variances, and standard deviations of the sampling distributions of χ^{2*} fit statistics, respectively, after each addition of 200 replications to the sampling distributions. Table 11 shows that the values for the means appear to be fairly stable after the first 200 replications. The values for the variances for individual items require between 600 and 1000 replications to stabilize. Findings were similar for G^{2*}.

Inspection of Tables 11 - 13 shows that the generation of sampling distributions of χ^{2*} fit statistics consisting of 1000 values appears to stabilize the first two moments of the sampling distributions. For this reason, for all test configurations, 1000 fit statistics were generated for each sampling distribution.

Table 11. Means of the Sampling Distributions of χ^{2*} After Every 200 Replications

	Number of Replications									
	200	400	600	800	1000	1200	1400	1600	1800	2000
Item	$\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}$
1	13.5	13.6	13.4	13.5	13.5	13.5	13.5	13.4	13.5	13.5
2	15.0	14.9	14.9	14.8	14.8	14.7	14.8	14.7	14.8	14.7
3	13.2	13.1	13.3	13.5	13.5	13.5	13.5	13.5	13.5	13.5
4	12.7	12.4	12.6	12.7	12.7	12.8	12.8	12.8	12.8	12.7
5	16.4	16.4	16.5	16.5	16.5	16.4	16.5	16.6	16.6	16.6
6	13.3	13.3	13.3	13.3	13.2	13.2	13.3	13.3	13.3	13.4
7	17.0	16.7	16.8	17.0	16.8	16.7	16.8	16.7	16.7	16.7
8	14.4	14.7	14.8	14.9	14.8	14.8	14.7	14.7	14.7	14.7
9	15.8	16.2	16.3	16.2	16.2	16.3	16.3	16.2	16.2	16.2
10	13.5	13.4	13.4	13.2	13.2	13.2	13.2	13.2	13.2	13.3
11	16.5	16.3	16.3	16.4	16.4	16.4	16.4	16.5	16.5	16.5
12	16.0	15.6	15.5	15.6	15.5	15.5	15.4	15.4	15.4	15.5

Table 12. Variances of the Sampling Distributions of χ^{2*} After Every 200 Replications

	Number of Replications									
	200	400	600	800	1000	1200	1400	1600	1800	2000
Item	s^2	s^2	s^2	s^2	s^2	s^2	s^2	s^2	s^2	s^2
1	24.2	20.3	19.4	20.4	19.3	18.9	18.4	18.2	18.0	17.9
2	22.6	21.1	20.8	20.1	20.6	20.1	19.7	20.2	20.5	20.4
3	17.6	18.0	19.4	19.0	18.8	18.8	18.8	19.1	19.0	18.9
4	19.0	16.4	15.9	16.6	16.7	16.7	16.9	17.6	17.2	17.2
5	24.8	25.2	24.6	24.3	24.0	24.2	24.3	24.3	24.1	24.2
6	19.3	17.3	16.9	17.5	17.6	17.6	17.1	17.3	17.3	17.5
7	21.8	22.8	21.4	22.1	22.3	22.6	23.8	23.3	23.5	23.5
8	20.9	18.8	18.3	18.2	18.5	18.3	18.4	18.4	18.7	18.9
9	24.3	24.9	24.0	23.5	24.2	24.1	24.0	23.2	23.6	23.6
10	20.7	18.1	17.3	17.0	17.3	17.9	18.0	18.0	18.7	18.8
11	27.4	24.5	23.5	23.8	24.3	24.0	23.5	23.8	24.0	23.9
12	22.0	22.5	21.7	21.3	21.7	21.2	20.7	21.0	20.8	21.0

Table 13. Standard Deviations of the Sampling Distributions of χ^{2*} After Every 200 Replications

	Number of Replications									
	200	400	600	800	1000	1200	1400	1600	1800	2000
Item	s	s	s	s	s	s	s	s	s	s
1	4.9	4.5	4.4	4.5	4.4	4.3	4.3	4.3	4.2	4.2
2	4.8	4.6	4.6	4.5	4.5	4.5	4.4	4.5	4.5	4.5
3	4.2	4.2	4.4	4.4	4.3	4.3	4.3	4.4	4.4	4.3
4	4.4	4.1	4.0	4.1	4.1	4.1	4.1	4.2	4.2	4.1
5	5.0	5.0	5.0	4.9	4.9	4.9	4.9	4.9	4.9	4.9
6	4.4	4.2	4.1	4.2	4.2	4.2	4.1	4.2	4.2	4.2
7	4.7	4.8	4.6	4.7	4.7	4.8	4.9	4.8	4.8	4.9
8	4.6	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3
9	4.9	5.0	4.9	4.8	4.9	4.9	4.9	4.8	4.9	4.9
10	4.5	4.3	4.2	4.1	4.2	4.2	4.2	4.2	4.3	4.3
11	5.2	4.9	4.8	4.9	4.9	4.9	4.8	4.9	4.9	4.9
12	4.7	4.7	4.7	4.6	4.7	4.6	4.6	4.6	4.6	4.6

III. A. 4. c) Computation of χ^{2*} and G^{2*}

In the current study, the computational formula for the χ^{2*} fit statistic is given by Equation 12. The computational form of G^{2*} is given by Equation 13. As can be seen in these equations, a number of discrete ability levels, or quadrature points, are used to approximate the continuous ability scale. In the current study, 11 ability subgroups over the ability range ($-2 \leq \theta \leq 2$) were utilized in forming the item fit tables for each item. The fit statistics were computed over the ability range ($-2 \leq \theta \leq 2$), because of the expected sparseness in the pseudocounts and expected frequencies beyond this ability range (Ankenmann, 1994; Stone et al., 1994).

As was discussed, problems arise with chi-square fit statistics when the expected cell counts are small. Larger sample sizes may overcome the existence of small expected cell counts. If this is not the case, adjustments or corrections are often made to the expected frequencies in order to overcome computational errors. Stone and Hansen (2000), as was recommended by Agresti (1990), computed their goodness-of-fit statistics by adding a small constant, in their study 0.000001, to each expected cell frequency that was zero. Using this same adjustment, nominal Type I error rates for the χ^{2*} fit statistic have been found (Stone & Zhang, 2003). Orlando and Thissen (2000) handled this problem by requiring a minimum cell frequency of 1 in order for the cell to be used in the computation of the chi-square statistic. Ankenmann (1994) handled the problem by calculating a goodness-of-fit statistic over a limited range of the ability distribution ($-2 \leq \theta \leq 2$), and excluding cells in which the expected cell frequency was less than 0.01.

In the current study, any cells for which the expected values were less than 0.01, or for which the pseudocounts were equal to 0, were excluded from the computation of the particular χ^{2*} fit statistic. Elimination of these cells of the item fit tables was necessary in order to avoid computational errors.

Eliminating some cells of the item fit tables for some of the fit statistics would have caused discrepancies in the number of degrees of freedom associated with the item fit tables for a single item across replications**.** To ensure that all fit statistics for an item were based on the same number of degrees of freedom, some adjustments to the degrees of freedom were made. The degrees of freedom associated with the fit statistics that were affected by the exclusion criteria were adjusted by multiplying the fit statistic by the ratio of the number of cells in the fit table over the ability range ($-2 \leq \theta \leq 2$), divided by the number of cells in the item fit table for

which the particular statistics was formed considering the exclusion criteria. This adjustment is equivalent to substituting the mean of the observed values into the cells having missing values (Ankenmann, 1994; Stone et al., 1994).

Individual ability estimates in the current study were given by *EAP(θ)* and prior weights based on the N(0,1) distribution were used. It should be noted that standard IRT computer packages (e.g., MULTILOG) do not compute the χ^{2*} or G^{2*} item fit statistics. The program used for the current study that allows users to obtain these statistics is found in Appendix B.

III. A. 4. d) Calculation of Mean Posterior Variance

The mean posterior variance was computed using the SAS program found in Appendix B. One value for the mean posterior variance was obtained for each test configuration. The value was the mean of the posterior variance distribution calculated across examinees and replications of the study for the test configuration. In practice, the mean posterior variance will be obtained by computing VAR($\hat{\theta}$), given by Equation 18, for each examinee in the single sample of examinees taking the test items. The mean of these values will be obtained, and the mean for this single sample will be the value of the mean posterior variance that is used in the prediction equations.

III. A. 4. e) Data Obtained For Each Item and Test Configuration

At the completion of the data simulation component of this study, each item in each test configuration had associated with it an empirically generated sampling distribution of 1000 χ^{2*} fit statistics, and empirically derived scaling factor and degrees of freedom values based on the Pearson distributions. Each item also had associated with it an empirically generated sampling distribution of 1000 G^{2*} fit statistics, and empirically derived scaling factor and degrees of

freedom values based on these likelihood-ratio based distributions. Each test configuration had associated with it a single value of the mean posterior variance. The data obtained from the simulations for each item in each condition are found in Appendix C.

III. B. Examination of Rescaled Distributions

After sampling distributions of fit statistics for each item were generated as described above, the means and variances of these distributions were computed. The means and variances were then used with the method of moments as given by Equations 19 and 20 to determine the appropriate scaling corrections for each distribution. The scaling factor γ and degrees of freedom value ν were obtained for each item. Each fit statistic in the sampling distribution for the item was then divided by the scaling factor γ. Then the rescaled distribution was compared to the theoretical chi-square distribution with ν degrees of freedom. This was done for each item in each test configuration.

Comparisons between the rescaled empirical distributions and theoretical chi-square distributions were carried out through examination of Q-Q plots of the rescaled empirical versus theoretical distributions. If the distributions of the fit statistics followed scaled chi-square distributions, the Q-Q plots followed along the line $y = x$. To examine the match between the distributions, regression lines were fitted to each of the Q-Q plots, and the slopes and intercepts of the regression lines were collected. The closeness of the slopes and intercepts to the values of 1 and 0, respectively, were examined to determine the appropriateness of the scaling factor/degree of freedom combinations for each item.

In addition, the data in the tails of the distributions were examined further through the analysis of Type I error rates. Type I error rates indicate the proportion of times items that fit the model are flagged as misfitting. In this study, data were simulated under the assumption that the

items fit the GRM (apart from the error introduced to reflect real item responses). For a particular significance level, a certain number of Type I errors, or false rejections of the null hypothesis, were expected. Type I error rates were found by calculating the proportion of statistics that exceeded the 90^{th}, 95^{th}, and 99^{th} percentiles of the theoretical chi-square distribution with degrees of freedom ν. These percentiles correspond to Type I error rates of α = .10, α = .05, and α = .01, respectively. A close match between observed and nominal Type I error rates indicated a close match in the tails of the empirical and theoretical chi-square distributions.

To account for sampling error, 95% confidence intervals were computed using the formula defined in Equation 29.

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{1000}}, \tag{29}$$

Thus, across 1000 replications, Type I error rates of .018 to .119, .037 to .063, and .004 to .016, were expected at α = .10, α = .05, and α = .01, respectively.

After finding evidence through this analysis that the χ^{2*} and G^{2*} statistics followed scaled chi-square distributions, an attempt to estimate the appropriate scaling factors and degrees of freedom values for each item from sample data was made. Prediction equations were estimated for both the Pearson and likelihood ratio forms of the pseudocounts-based statistic.

III. C. Prediction of Scaling Corrections

The second phase of this study involved predicting the scaling corrections for each item from item and sample characteristics. Two sets of two prediction equations were formed, one set for predicting the scaling corrections for χ^{2*} and one for predicting the scaling corrections for G^{2*}. The first equation in each set used item and sample characteristics to predict the appropriate

value of the scaling factor $\hat{\gamma}$ for an item. The second equation used item and sample characteristics to predict the appropriate degrees of freedom $\hat{\nu}$ for the item. Thus, a total of four equations, two for χ^{2*} and two for G^{2*}, were estimated.

The results of Stone et al. (1994) and Ankenmann (1994) indicated that the mean posterior variance, root mean posterior variance, item information, and item discrimination parameter were useful in predicting the appropriate scaling corrections. In this study, item and test characteristics were included as independent variables so that two general equations to be used across items and tests could be found.

The mean posterior variance was included as a possible predictor variable in this study because it directly represented the effects of manipulating the test length factor. For shorter tests that yield less precise ability estimates, examinees' posterior distributions of ability are spread out over a wide range of ability. For longer tests, the posterior distributions of ability for examinees are concentrated over a smaller range on the ability scale.

As the variance of the posterior distribution of ability decreases, the dependence among pseudocounts in the item fit table also decreases, and less adjustment is needed for the pseudocounts-based statistic to follow theoretical chi-square distribution. Thus, the mean posterior variance was expected to serve as an important predictor variable for the scaling corrections. The square root of the mean posterior variance was also included as a possible predictor variable because it was included by Ankenmann (1994).

Item information was also included as a possible predictor variable. Test information is inversely related to the standard error of the ability estimates. The higher the test information, the lower the posterior variances, and vice versa. Test information at a specific level of ability is given by the sum of the item information functions at that ability level. Because item

information is related to the variance of the posterior distributions of ability, it was included as a possible predictor variable.

The item discrimination parameter *a* was also included as a possible predictor variable because item discrimination is related to the posterior variance. Items with high discrimination parameters provide more information, and are associated with smaller standard errors.

Finally, the number of item score category levels was included as a possible predictor variable because the degrees of freedom of the fit statistics are directly related to the number of categories for the item. It was thought that this variable would be a necessary predictor variable for the degrees of freedom values.

III. C. 1. Data Used to Estimate Prediction Equations

A total of four prediction equations were proposed to predict the scaling corrections needed to rescale the empirical χ^{2*} and G^{2*} fit statistic distributions. Item level data was used in estimating the equations. Associated with each item were the variables found in Table 14.

Table 14. Possible Item and Test Level Data to be Used For Prediction Equations

Discrimination (Main Dimension)	**Total Item Information**	**Number of Item Response Categories**	**Mean Posterior Variance**	**Root Mean Posterior Variance**
a	*info*	*ncat*	$\overline{p}$	$\sqrt{\overline{p}}$

A total of 384 sets of independent variables were used to estimate the regression equations. One set came from each item in each of the 20 test configurations. The set of

predictor variables for a single item consisted of the data found in Table 14. The raw data for each item in each condition are found in Appendix C.

The observations in each of the 384 data sets were not independent. Rather, the item level data was nested within each test configuration. That is, all items in a single test configuration had the same value of the mean posterior variance and root mean posterior variance. For this reason, a multilevel prediction model was fitted to the data.

Multilevel models are appropriate when data is provided at two levels within an organizational hierarchy, and interest lies in examining the behavior of a level-1 outcome as a function of both level-1 and level-2 predictors (Singer, 1998). Applied to this setting, the scaling factors were level-1 outcomes, and they were estimated using both level-1 (item level) and level-2 (test level) data.

III. C. 2. Fitting the Prediction Equations

In forming the equations to predict the scaling factor $\hat{\gamma}$ and degrees of freedom $\hat{\nu}$ for an item, the variables found in Table 14 served as the initial set of independent variables. Several models were estimated for each scaling correction, before the final prediction equations were obtained. The final models were chosen based on the amount of explained variation, and whether the contribution of the variable to the model was statistically significant. After prediction equations were estimated, they were evaluated to determine whether they predicted the appropriate scaling factor/degree of freedom combination for the items used in the current study.

III. D. Evaluation of Predicted Scaling Corrections

The multilevel prediction equations for predicting the scaling corrections for χ^{2*} and G^{2*} were assessed to determine if they yielded valid predictions of the scaling factor γ and degrees of freedom ν for each item. To assess the quality of the predicted scaling corrections, the empirically generated sampling distributions of the fit statistics for each item were rescaled by the predicted scaling factor $\hat{\gamma}$ for that item, by dividing each statistic in the sampling distribution by this value. The rescaled distribution was then compared to the chi-square distribution with predicted degrees of freedom $\hat{\nu}$. The comparisons for each item were again made using Q-Q plots and Type I error rates. Specifically, the linearity of the Q-Q plots was assessed, and the slope and intercept values of regression lines fitted to the Q-Q plots were calculated. Their deviations from 1 and 0, respectively, were examined. Type I error rates were examined to evaluate the behavior in the tails of the rescaled distributions.

To summarize, for each item, regression lines were fitted to Q-Q plots of the sampling distributions rescaled by $\hat{\gamma}$ versus the chi-square distribution with predicted degrees of freedom $\hat{\nu}$. The slopes and intercepts of the regression lines and Type I error rates associated with each item were examined. A match between the empirical and theoretical distributions indicated that the prediction equations worked adequately in determining the scaling corrections for the χ^{2*} and/or G^{2*} fit statistics.

III. E. Validation of the Use of χ^{2*} and G^2 With the Prediction Equations

Validation of the use of χ^{2*} and G^{2*} with the multilevel prediction equations found in the simulation study was carried out by examining the usefulness of the procedures for assessing the fit of real items having different item parameters than those used in the simulation aspect of this

study. Showing that the prediction equations could be applied to this set of new items would provide evidence that the equations would also generalize to other sets of test items. Then, when χ^{2*} or G^{2*} statistics are used in practice, the chi-square distribution, rather than empirically generated sampling distributions of the fit statistics, could be used for significance testing. The appropriate chi-square distribution for each item would be determined using the prediction equations.

III. E. 1. Application to Real Item Response Data

The prediction equations obtained in this study were useful to the extent that they generalized to real items having parameters different from those included in the study. Real data from two assessments were used to validate the procedures of this study.

First, data was obtained from the QUASAR Cognitive Assessment Instrument (QCAI) (Lane, 1993), administered during the 1991-1992 school year. The forms of the QCAI used in the validation procedures consisted of eight 5-category items. Data from Form A, administered during the Spring of 1991, and Forms A and B, administered during the Spring of 1992, were used.

The forms of the QCAI that were used for validation of the prediction equations in the current study were also used in the studies by Ankenmann (1994) and Stone et al. (1993). Item level results regarding decisions of fit were presented in Stone, et al. Therefore, the decision consistency regarding identification of item misfit using the procedures of this study and the procedure employed by Stone, et al. was examined.

The second set of validation data was obtained from the Reading domain of the 1994 NAEP assessment (Allen. Blocks of items on this NAEP assessment consisted of between 9 and

12 items having 2, 3, and 4 score categories. In the current study, Block 9M, consisting of 9 items, (4 two category, 4 three category, and 1 four category) was utilized.

The validation aspect of this study assessed whether the prediction equations generalized beyond the scope of this study. For the NAEP and QCAI validation data sets, the item parameters were estimated using MULTILOG. The pseudocounts-based item fit statistics χ^{2*} and G^2 were computed for each item. The predicted scaling factor $\hat{\gamma}$ and degrees of freedom value $\hat{\nu}$ were then calculated for each statistic using the appropriate set of prediction equations. The predictor variables used in the equations were *a*, and *ncat* obtained from the individual NAEP and QCAI items, and $\bar{p}$ and $\sqrt{\bar{p}}$, obtained from the single sample of subjects in the real data sets. Decisions regarding item fit were made through comparison of the fit statistic rescaled by $\hat{\gamma}$ to the chi-square distribution with predicted degrees of freedom $\hat{\nu}$.

III. E. 2. Decisions of Item Fit Using Several Methods

For the QCAI data, in order to determine if the correct decision of item fit was made using the prediction equations, the fit of the model was assessed in several ways. First, the observed and expected score distributions for each item score category level were plotted on the same graph. Graphs in which the observed and expected score distributions overlapped indicated fit, while graph showing large discrepancies between the observed and expected score distributions indicated misfit.

In addition, the resampling method described by Stone (2000) was also employed here. This method involved generating sampling distributions of the fit statistics for each item, but with a modification from the methodology used in the earlier stages of this project. The Stone (2000) method used the initial item parameter estimates with a randomly selected ability value

when computing each fit statistic. The item parameters were not re-estimated for each statistic, as they were in the simulation aspect of this study. After sampling distributions were formed in this manner, the scaling corrections for each item were obtained from the sampling distributions. Decisions of item fit were then evaluated by comparing the rescaled statistic to the appropriate chi-square distribution.

The decisions of item fit based on the prediction equations obtained in the current study were compared to decisions of item fit made by Stone et al. (1993) for the same QCAI data sets. The decision consistency between the methods was evaluated.

For the NAEP data, in order to determine if the correct decisions of item fit were made using the prediction equations, the fit of the item was also assessed using graphical displays of observed versus expected score distributions, and using the resampling method employed by Stone (2000).

The validation procedures assessed the utility of the estimated prediction equations for sets of real items that were not used to form the prediction equations. The validation procedures that were employed allowed for comparisons of decisions of item fit using several methods. Comparing the decisions of item fit made using these different techniques was useful in evaluating the utility and generalizability of the prediction equations.

CHAPTER IV

IV. RESULTS

Chapter IV presents the results from this study. Results are presented in three sections. The first contains information about the empirical sampling distributions of the fit statistics. Quantile-Quantile (Q-Q) plots of original and rescaled empirical distributions versus theoretical chi-square distributions, and tables summarizing the scaling corrections and Type I error rates associated with the rescaled sampling distributions are included.

The second section contains results that relate to the prediction of the scaling corrections. Included are summaries of the multilevel models that were fitted to the data. Descriptions of sampling distributions that were rescaled by the predicted scaling corrections are provided. Further, this section contains results related to additional prediction models that were estimated.

The third section presents results that relate to the application of the prediction equations to real items. The real items were from the QUASAR Cognitive Assessment Instrument (QCAI) (Lane, 1993), and the reading domain of the 1994 National Assessment of Educational Progress (NAEP) (Allen et al., 1994).

IV. A. Results from the Simulated Fit Statistic Distributions

Empirical sampling distributions of Pearson χ^{2*} and likelihood-ratio G^{2*} fit statistics were generated for each item in each test configuration in this study. The generation of sampling

distributions involved simulating item response data sets under the Multidimensional Graded Response Model (MGRM) having one main and five additional nuisance dimensions, calibrating the simulated response data under Samejima's (1969) Graded Response Model (GRM), and computing the Pearson and likelihood ratio forms of the pseudocounts-based fit statistics for each item. One thousand replications of these steps produced sampling distributions of the χ^{2*} and G^{2*} fit statistics.

IV. A. 1. Q-Q Plots of Empirical Sampling Distributions

For each item in each test configuration, the empirically generated sampling distribution was compared to the null chi-square distribution. The null chi-square distribution for each item was the chi-square distribution with $J * (k-1) - m$ degrees of freedom, where J was the number of ability subgroups, k the number of item score categories for the item, and m the number of estimated item parameters. In the current study, 11 ability subgroups were utilized. For the 5-category items, the empirical sampling distributions were compared to the theoretical chi-square distribution with $J * (k-1) - m = 11 * (5-1) - 5$, or 39 degrees of freedom.

Comparisons between the empirical and theoretical distributions were made by 1) evaluating the linearity of Q-Q plots of the empirical versus theoretical distributions, and 2) examining the slopes and intercepts of regression lines fitted to the Q-Q plots for their closeness to 1 and 0, respectively.

Q-Q plots that were linear with slopes of 1 and intercepts of 0 indicated a match between the empirical and theoretical distributions. Q-Q plots that were linear with slopes different from 1 indicated that the empirical and theoretical distributions were from the same family of distributions but differed in spread. Q-Q plots that were linear with intercepts different from 0 indicated that the empirical and theoretical distributions differed in location.

Generally, as was expected, the plots of the original empirical sampling distributions were linear, indicating that the empirical and theoretical distributions came from the same family of distributions. However, the data did not fall along the line $y = x$. Instead, the plots were linear and compressed, indicating that the sampling distributions may follow scaled chi-square distributions.

To illustrate, Figures 7 and 8 present Q-Q plots of empirical Pearson χ^{2*} versus theoretical chi-square distributions for one 2-category and one 3-category item, respectively. Figures 9 and 10 present Q-Q plots for empirical likelihood ratio G^{2*} distributions for one 4-category and one 5-category item, respectively. Q-Q plots of empirical χ^{2*} and G^{2*} distributions for the majority of the remaining items were similar, and are not presented in order to conserve space. However, additional Q-Q plots for items that showed deviations from linearity are discussed in Section IV. A. 3.

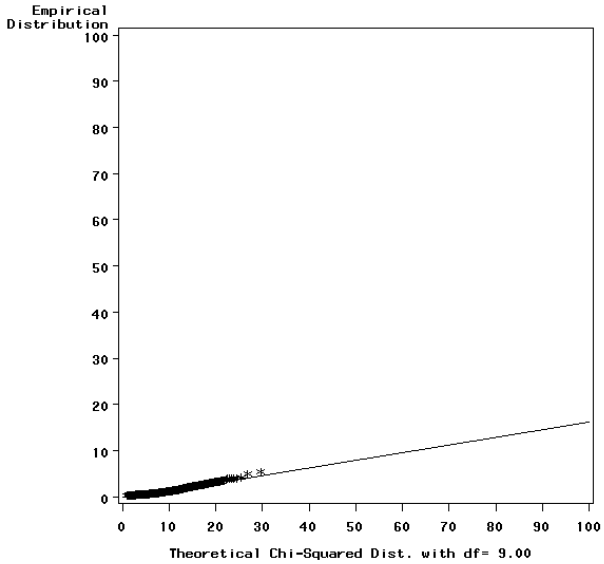

Figure 7. Q-Q Plot of Empirical χ^{2*} Distribution for Item 9b-2Cat From Test 2Cat12b

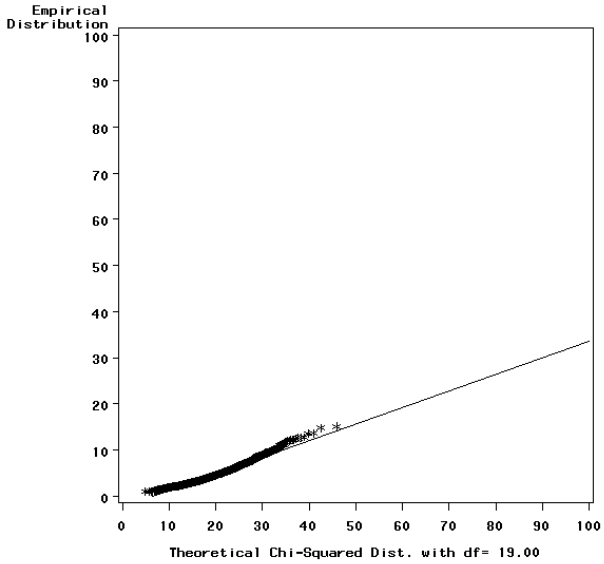

Figure 8. Q-Q Plot of Empirical χ^{2*} Distribution for Item 1b-3Cat From Test 3Cat24

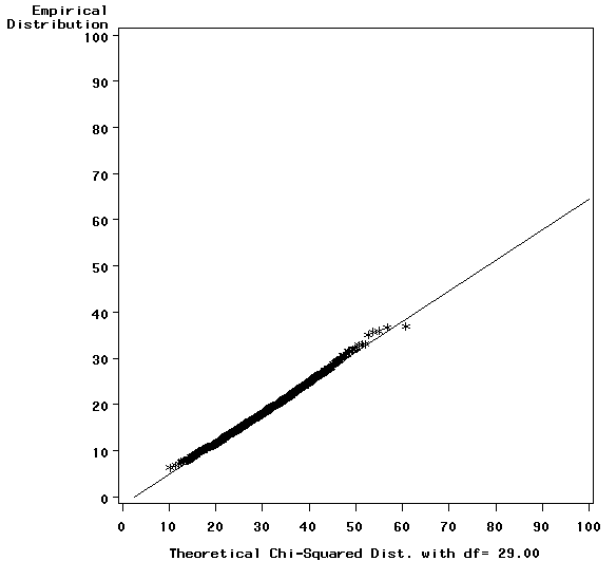

Figure 9. Q-Q Plot of Empirical G^{2*} Distribution for Item 6c-4Cat From Test 4Cat36

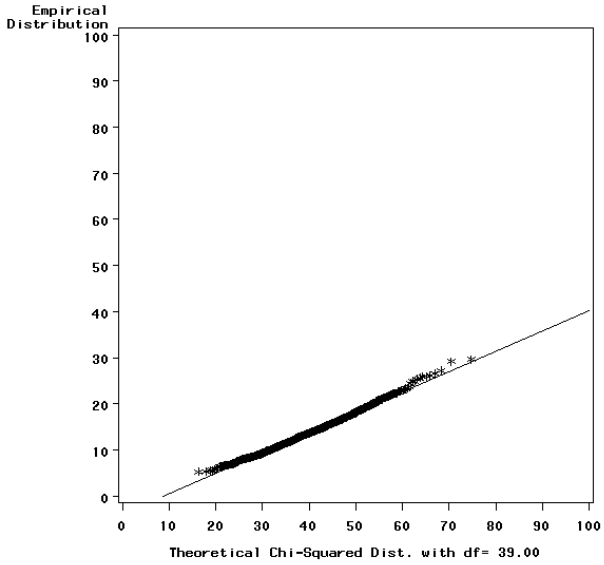

Figure 10. Q-Q Plot of Empirical G^{2*} Distribution 2a-5Cat From Test 5Cat12c

Having found evidence that the sampling distributions of the χ^{2*} and G^{2*} statistics may follow one of the family of chi-square distributions, the scaling corrections for each item in each test configuration were obtained. The scaling corrections were obtained using the method of moments with the means and variances of the empirical sampling distributions. The empirically generated sampling distributions were rescaled by the scaling factors γ (e.g. χ^{2*} / γ) obtained using Equation 19. The rescaled distributions were compared to the theoretical chi-square distributions with degrees of freedom ν, obtained using Equation 20. Comparisons between rescaled sampling distributions and theoretical chi-square distributions were made by assessing the linearity of Q-Q plots, the slopes and intercepts of regression lines fitted to Q-Q plots, and Type I error rates.

IV. A. 2. Q-Q Plots of Rescaled Sampling Distributions

Q-Q plots of rescaled Pearson χ^{2*} and likelihood ratio G^{2*} distributions versus theoretical chi-square distributions with degrees of freedom ν for the four items presented in Figures 7 to 10 are found in Figures 11 to 14. Plots of these rescaled distributions showed that the data were linear and fell close to the line $y = x$. This indicated that the sampling distributions of fit statistics for these items followed scaled chi-square distributions, where the scaling corrections were obtained using Equations 19 and 20. Plots for the majority of the remaining items were similar, and are not shown. Q-Q plots for items that showed deviations from linearity are discussed in Section IV. A. 3.

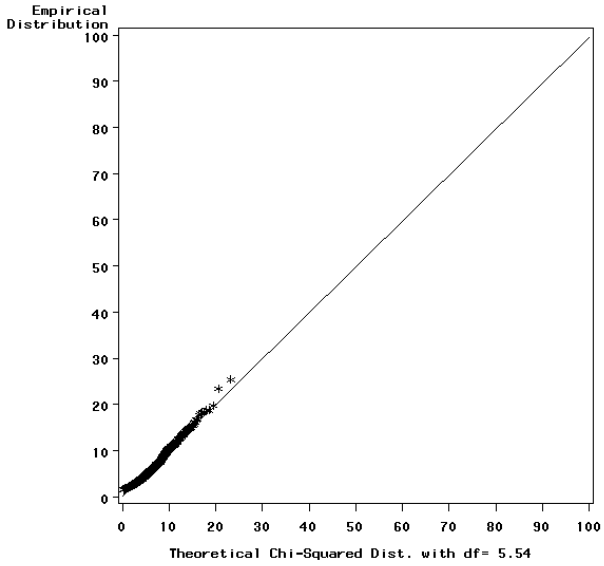

Figure 11. Q-Q Plot of Rescaled χ^{2*} Distribution for Item 9b-2Cat From Test 2cat12b

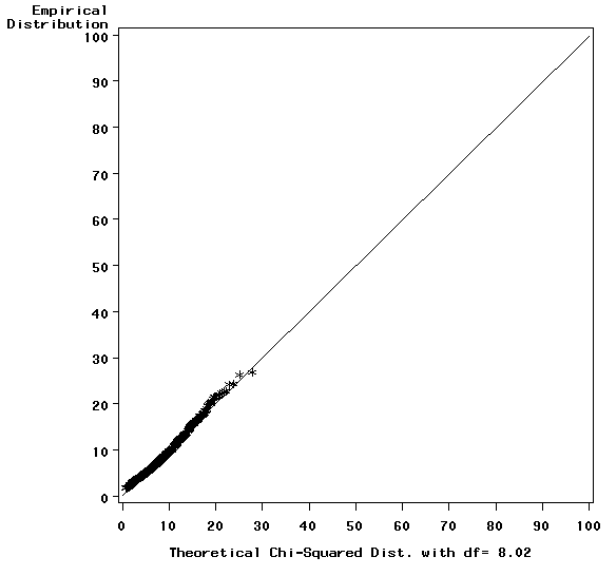

Figure 12. Q-Q Plot of Rescaled χ^{2*} Distribution for Item 1b-3Cat From Test 3Cat24

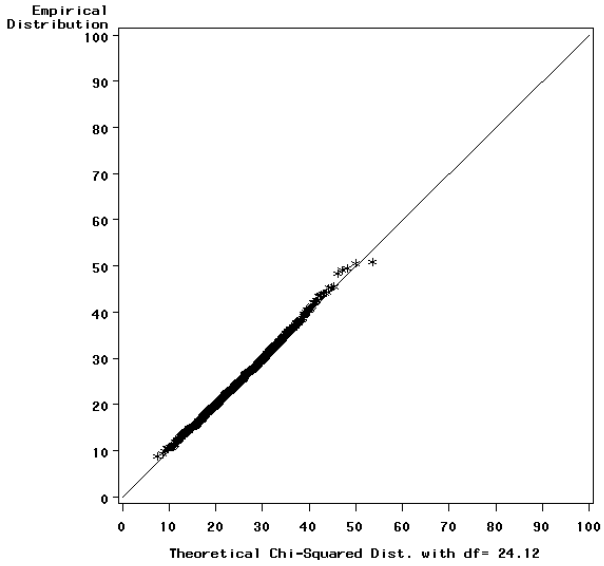

Figure 13. Q-Q Plot of Rescaled G^{2*} Distribution for Item 6c-4Cat From Test 4cat36

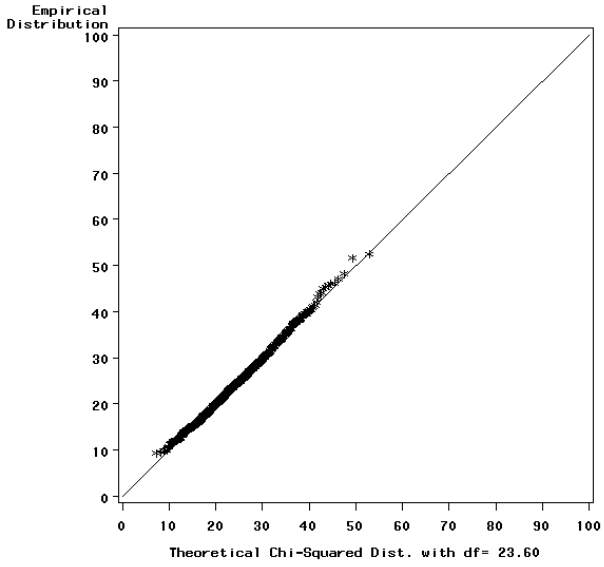

Figure 14. Q-Q Plot of Rescaled G^{2*} Distribution for Item 2a-5Cat From Test 5cat12c

IV. A. 3. Departures from Theoretical Chi-Square Distributions

For several items, the sampling distributions showed some deviation from linearity in the tails of the distributions. Examples from one such item are illustrated in Figures 15 and 16. Figure 15 presents a Q-Q plot of the rescaled Pearson χ^{2*} distribution for Item 7b-3Cat from Test 3cat12b. Figure 16 presents a Q-Q plot of the rescaled likelihood ratio G^{2*} sampling distribution for the same item.

Although the Q-Q plot analysis indicated that the majority of rescaled sampling distributions may follow chi-square distributions with ν degrees of freedom, the plots of some items, as illustrated by Figures 15 and 16, indicated that their sampling distributions deviated from theoretical chi-square distributions. Generally, deviations from linearity exhibited in the Q-Q plots were apparent only in the more extreme tails of the distributions.

More 2- and 3-category items than 4- and 5-category items showed deviations from linearity in the Q-Q plots. In addition, the likelihood ratio form of the statistic generally showed less deviation from linearity in the tails than the Pearson form of the statistic.

Deviations from linearity in the tails of the sampling distributions were the result of large values of the fit statistics, and were typically caused by small expected values in the item fit tables. Small expected cell counts can cause the distributions of Pearson and likelihood ratio goodness-of-fit statistics to deviate from theoretical chi-square distributions (see Section II. A.). In the current study, the Pearson χ^{2*} distributions were affected more by small cell expectations than the likelihood ratio G^{2*} distributions. Orlando and Thissen (2000) state that the G^2 statistic is more sensitive to small expected cell counts than χ^2. Results of this study, however, suggested that χ^{2*} was more sensitive to small cell expectations than G^{2*}. The behavior of the statistics in the tails of the distributions was also assessed further through analysis of Type I error rates.

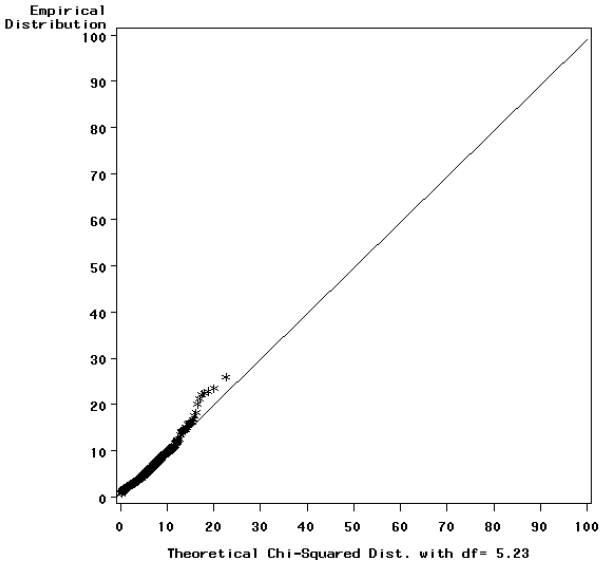

Figure 15. Q-Q Plot of Rescaled χ^{2*} Statistics For Item 7b-3Cat From Test 3cat12b

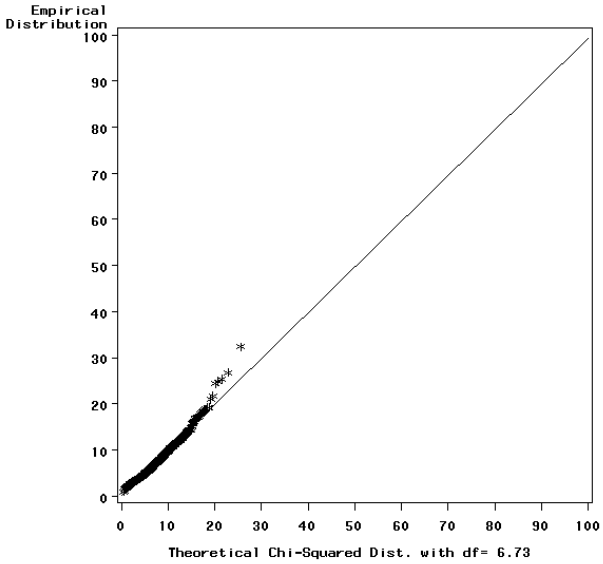

Figure 16. Q-Q Plot of Rescaled G^{2*} Statistics For Item 7b-3Cat From Test 3cat12b

IV. A. 4. Descriptions of Empirically Generated and Rescaled χ^{2*} and G^{2*} Sampling Distributions

Tables 15 to 18 present results from the analysis of Q-Q plots of the Pearson χ^{2*} distributions for the items in Tests 2Cat12A, 2Cat36, 5Cat12A, and 5Cat36, respectively. The tables contain item statistics including the item parameters and total item information (as given by Equation 26 with $\Delta\theta = 0.05$). In addition, the slopes and intercepts of regression lines fitted to original and rescaled fit statistic distributions are provided for each item. Tables 15 to 18 also contain the scaling corrections, a scaling factor γ and degrees of freedom value ν, obtained using the method of moments with Equations 19 and 20. Finally, the tables contain the Type I error rates associated with each rescaled distribution. Type I error rates that fall within the ranges expected with 95% confidence intervals are underlined. Tables 19 to 22 present results for the likelihood ratio G^{2*} statistics for the same four tests.

Tables for only this subset of tests are presented below to simplify the discussion, and at the same time allow for comparisons of the Pearson and likelihood ratio forms of the statistics across number of item score category levels and test lengths. Tables containing this information for all conditions can be obtained from the author.

Table 15. Summary of Empirical and Rescaled χ^{2*} Sampling Distributions for the Two Category Items in Test 2Cat12a

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	*a*	*Avg b*	*Info*	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
1a-2Cat	0.7	-1.5	0.5	0.181	-0.392	0.984	0.079	0.253	4.892	0.014	0.054	0.086
2a-2Cat	1.0	-1.5	0.8	0.219	-0.328	0.965	0.194	0.290	5.666	0.016	0.054	0.078
3a-2Cat	1.4	-1.5	1.2	0.245	0.066	0.928	0.593	0.278	8.163	0.015	0.038	0.065
4a-2Cat	1.7	-1.5	1.6	0.275	0.344	0.948	0.550	0.266	10.585	0.021	0.046	0.077
5a-2Cat	2.1	-1.5	2.0	0.294	0.816	0.963	0.539	0.238	14.537	0.016	0.041	0.080
6a-2Cat	2.4	-1.5	2.3	0.310	1.139	0.989	0.185	0.222	17.716	0.017	0.061	0.091
7a-2Cat	0.7	1.0	0.5	0.167	-0.361	0.970	0.139	0.243	4.696	0.014	0.045	0.082
8a-2Cat	1.0	1.0	0.9	0.169	-0.190	0.967	0.206	0.211	6.317	0.019	0.045	0.078
9a-2Cat	1.4	1.0	1.3	0.166	-0.029	0.948	0.404	0.191	7.690	0.017	0.051	0.077
10a-2Cat	1.7	1.0	1.6	0.176	-0.112	0.912	0.555	0.235	6.260	0.017	0.038	0.067
11a-2Cat	2.1	1.0	2.1	0.143	0.064	0.921	0.657	0.161	8.375	0.016	0.044	0.065
12a-2Cat	2.4	1.0	2.4	0.121	0.157	0.905	0.923	0.128	9.692	0.012	0.034	0.064

Table 16. Summary of Empirical and Rescaled χ^{2*} Sampling Distributions for the Two Category Items in Test 2Cat36

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	a	Avg b	Info	Slope	Intercept	Slope	Intercept	γ	ν	α = .01	α = .05	α = .10
1a-2Cat	0.7	-1.5	0.5	0.429	-0.876	0.990	0.052	0.578	5.160	0.013	0.057	0.086
2a-2Cat	1.0	-1.5	0.8	0.477	-0.940	0.980	0.100	0.654	5.127	0.018	0.048	0.090
3a-2Cat	1.4	-1.5	1.2	0.569	-1.103	0.935	0.299	0.870	4.616	0.014	0.034	0.059
4a-2Cat	1.7	-1.5	1.6	0.639	-1.139	0.956	0.222	0.904	5.099	0.016	0.034	0.078
5a-2Cat	2.1	-1.5	2.0	0.599	-0.301	0.984	0.125	0.659	7.718	0.015	0.053	0.088
6a-2Cat	2.4	-1.5	2.3	0.591	0.009	0.974	0.221	0.624	8.546	0.017	0.049	0.083
7a-2Cat	0.7	1.0	2.1	0.398	-0.660	0.999	0.007	0.493	5.931	0.014	0.047	0.103
8a-2Cat	1.0	1.0	0.5	0.414	-0.631	0.997	0.021	0.507	6.112	0.013	0.047	0.093
9a-2Cat	1.4	1.0	1.6	0.426	-0.684	0.994	0.038	0.532	5.920	0.014	0.046	0.100
10a-2Cat	1.7	1.0	2.4	0.506	-1.191	0.987	0.061	0.726	4.629	0.019	0.056	0.085
11a-2Cat	2.1	1.0	1.1	0.505	-1.551	0.960	0.134	0.898	3.330	0.020	0.039	0.065
12a-2Cat	2.4	1.0	2.2	0.495	-1.725	0.920	0.203	1.086	2.513	0.011	0.033	0.057
1b-2Cat	0.7	-0.5	0.5	0.408	-0.771	1.000	0.002	0.522	5.568	0.011	0.056	0.110
2b-2Cat	1.0	-0.5	0.9	0.391	-0.738	0.993	0.040	0.509	5.461	0.010	0.049	0.091
3b-2Cat	1.4	-0.5	1.4	0.338	-0.541	1.000	0.002	0.414	6.053	0.012	0.049	0.102
4b-2Cat	1.7	-0.5	1.7	0.316	-0.574	0.992	0.049	0.408	5.552	0.010	0.047	0.094
5b-2Cat	2.1	-0.5	2.1	0.314	-0.760	0.979	0.089	0.460	4.485	0.010	0.044	0.078
6b-2Cat	2.4	-0.5	2.4	0.324	-0.987	0.985	0.054	0.527	3.660	0.014	0.039	0.074

Table 16 (cont'd).

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	a	Avg b	Info	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
7b-2Cat	0.7	0.5	0.5	0.390	-0.670	1.000	0.005	0.484	5.854	0.010	0.061	0.100
8b-2Cat	1.0	0.5	0.9	0.393	-0.751	0.994	0.028	0.512	5.445	0.010	0.051	0.095
9b-2Cat	1.4	0.5	1.4	0.339	-0.579	1.000	0.006	0.422	5.855	0.011	0.052	0.097
10b-2Cat	1.7	0.5	1.7	0.354	-0.789	0.994	0.032	0.485	4.941	0.008	0.046	0.091
11b-2Cat	2.1	0.5	2.1	0.296	-0.657	0.993	0.036	0.407	4.928	0.012	0.045	0.087
12b-2Cat	2.4	0.5	2.4	0.314	-0.945	0.977	0.082	0.519	3.618	0.017	0.040	0.075
1c-2Cat	0.7	-1.0	0.5	0.389	-0.589	0.999	0.002	0.471	6.187	0.010	0.054	0.112
2c-2Cat	1.0	-1.0	0.9	0.388	-0.574	0.998	0.017	0.471	6.201	0.010	0.054	0.108
3c-2Cat	1.4	-1.0	1.3	0.460	-0.830	0.971	0.156	0.626	5.280	0.013	0.045	0.078
4c-2Cat	1.7	-1.0	1.6	0.440	-0.860	0.988	0.059	0.589	5.266	0.013	0.049	0.087
5c-2Cat	2.1	-1.0	2.1	0.531	-1.728	0.935	0.192	1.058	2.882	0.015	0.033	0.057
6c-2Cat	2.4	-1.0	2.4	0.456	-1.447	0.946	0.164	0.857	3.098	0.017	0.028	0.063
7c-2Cat	0.7	1.5	0.5	0.442	-1.012	0.983	0.084	0.631	4.704	0.009	0.046	0.086
8c-2Cat	1.0	1.5	0.8	0.435	-0.626	0.995	0.031	0.529	6.218	0.013	0.055	0.099
9c-2Cat	1.4	1.5	1.2	0.627	-1.444	0.947	0.221	0.986	4.256	0.013	0.043	0.080
10c-2Cat	1.7	1.5	1.6	0.632	-1.073	0.960	0.211	0.873	5.279	0.017	0.047	0.086
11c-2Cat	2.1	1.5	2.0	0.705	-1.213	0.916	0.398	1.095	4.691	0.013	0.038	0.065
12c-2Cat	2.4	1.5	2.3	0.606	-0.110	0.949	0.397	0.693	7.713	0.014	0.045	0.082

Table 17. Summary of Empirical and Rescaled χ^{2*} Sampling Distributions for the Five Category Items in Test 5Cat12a

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	*a*	*Avg b*	*Info*	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
1a-5Cat	0.7	-0.5	0.9	0.466	-5.728	0.993	0.134	0.700	17.763	0.018	0.052	0.094
2a-5Cat	1.0	-0.5	1.7	0.500	-6.419	0.993	0.118	0.767	17.068	0.015	0.057	0.099
3a-5Cat	1.4	-0.5	3.1	0.523	-6.422	0.992	0.152	0.787	17.762	0.020	0.048	0.089
4a-5Cat	1.7	-0.5	4.4	0.574	-7.269	0.994	0.103	0.872	17.326	0.017	0.057	0.092
5a-5Cat	2.1	-0.5	6.2	0.570	-6.180	0.989	0.220	0.818	19.627	0.015	0.051	0.097
6a-5Cat	2.4	-0.5	7.6	0.596	-5.676	0.998	0.051	0.797	22.056	0.014	0.059	0.109
7a-5Cat	0.7	0.5	0.8	0.460	-5.289	0.999	0.024	0.656	19.251	0.016	0.046	0.091
8a-5Cat	1.0	0.5	1.6	0.505	-6.763	0.950	0.736	0.874	14.791	0.013	0.039	0.067
9a-5Cat	1.4	0.5	2.9	0.495	-5.706	0.991	0.176	0.725	18.786	0.014	0.044	0.094
10a-5Cat	1.7	0.5	4.0	0.568	-7.401	0.990	0.161	0.884	16.676	0.017	0.049	0.088
11a-5Cat	2.1	0.5	5.5	0.553	-5.470	0.996	0.077	0.751	21.460	0.011	0.053	0.101
12a-5Cat	2.4	0.5	6.8	0.619	-6.481	0.998	0.032	0.853	20.699	0.009	0.058	0.108

Table 18. Summary of Empirical and Rescaled χ^{2*} Sampling Distributions for the Five Category Items in Test 5Cat36

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	a	Avg b	Info	Slope	Intercept	Slope	Intercept	γ	ν	α = .01	α = .05	α = .10
1a-5Cat	0.7	-0.5	0.9	0.689	-3.727	0.997	0.093	0.808	28.671	0.014	0.054	0.105
2a-5Cat	1.0	-0.5	1.7	0.747	-5.469	0.996	0.097	0.931	25.419	0.015	0.050	0.101
3a-5Cat	1.4	-0.5	3.1	0.839	-7.760	0.981	0.413	1.157	21.565	0.015	0.051	0.090
4a-5Cat	1.7	-0.5	4.4	0.974	-11.540	0.965	0.600	1.536	17.207	0.015	0.039	0.071
5a-5Cat	2.1	-0.5	6.2	1.017	-11.390	0.992	0.153	1.467	19.285	0.012	0.066	0.106
6a-5Cat	2.4	-0.5	7.6	0.958	-8.100	0.993	0.162	1.249	23.434	0.017	0.047	0.087
7a-5Cat	0.7	0.5	0.8	0.686	-3.679	0.996	0.123	0.805	28.668	0.014	0.058	0.108
8a-5Cat	1.0	0.5	1.6	0.782	-6.495	0.984	0.374	1.037	23.166	0.016	0.048	0.086
9a-5Cat	1.4	0.5	2.9	0.825	-7.175	0.993	0.169	1.086	23.030	0.017	0.049	0.087
10a-5Cat	1.7	0.5	4.0	0.969	-11.370	0.985	0.274	1.453	18.181	0.023	0.043	0.090
11a-5Cat	2.1	0.5	5.5	0.964	-9.738	0.990	0.200	1.341	20.778	0.017	0.052	0.086
12a-5Cat	2.4	0.5	6.8	0.976	-9.002	0.991	0.204	1.314	22.121	0.015	0.057	0.092
1b-5Cat	0.7	0.0	0.8	0.655	-2.513	0.998	0.051	0.729	31.561	0.013	0.051	0.096
2b-5Cat	1.0	0.0	1.6	0.672	-2.626	0.998	0.059	0.750	31.414	0.012	0.059	0.110
3b-5Cat	1.4	0.0	2.9	0.753	-4.653	0.985	0.406	0.929	26.618	0.013	0.052	0.092
4b-5Cat	1.7	0.0	4.1	0.812	-6.386	0.991	0.225	1.044	24.237	0.020	0.046	0.087
5b-5Cat	2.1	0.0	5.8	0.893	-8.031	0.992	0.193	1.191	22.501	0.020	0.055	0.090
6b-5Cat	2.4	0.0	7.1	0.950	-9.897	0.996	0.083	1.318	20.606	0.014	0.059	0.097

Table 18 (cont'd).

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	a	Avg b	Info	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
7b-5Cat	0.7	0.0	0.9	0.658	-2.919	1.000	0.015	0.743	30.589	0.009	0.056	0.104
8b-5Cat	1.0	0.0	1.7	0.702	-3.765	0.991	0.249	0.831	28.404	0.014	0.046	0.094
9b-5Cat	1.4	0.0	3.2	0.873	-8.759	0.982	0.361	1.236	20.479	0.018	0.051	0.082
10b-5Cat	1.7	0.0	4.5	0.876	-7.583	0.994	0.140	1.148	23.147	0.018	0.051	0.093
11b-5Cat	2.1	0.0	6.3	1.004	-9.028	0.994	0.135	1.330	22.644	0.017	0.053	0.102
12b-5Cat	2.4	0.0	7.8	1.125	-10.850	0.974	0.547	1.599	20.655	0.012	0.053	0.090
1c-5Cat	0.7	-0.5	0.8	0.628	-1.924	0.999	0.026	0.683	33.053	0.015	0.055	0.096
2c-5Cat	1.0	-0.5	1.6	0.672	-2.634	0.998	0.074	0.752	31.360	0.017	0.051	0.099
3c-5Cat	1.4	-0.5	2.9	0.817	-7.542	0.974	0.548	1.141	21.317	0.009	0.042	0.091
4c-5Cat	1.7	-0.5	4.0	0.803	-6.454	0.996	0.109	1.027	24.220	0.014	0.058	0.094
5c-5Cat	2.1	-0.5	5.7	0.864	-8.166	0.993	0.155	1.166	21.873	0.017	0.052	0.088
6c-5Cat	2.4	-0.5	7.0	0.974	-11.140	0.977	0.411	1.467	18.298	0.012	0.044	0.088
7c-5Cat	0.7	0.5	0.8	0.657	-3.075	0.998	0.057	0.750	30.038	0.013	0.063	0.100
8c-5Cat	1.0	0.5	1.6	0.676	-2.831	0.996	0.117	0.764	30.772	0.019	0.053	0.094
9c-5Cat	1.4	0.5	2.9	0.800	-7.053	0.983	0.371	1.080	22.366	0.014	0.048	0.083
10c-5Cat	1.7	0.5	4.0	0.807	-6.794	0.997	0.078	1.041	23.678	0.015	0.052	0.098
11c-5Cat	2.1	0.5	5.7	0.953	-11.420	0.980	0.352	1.460	17.649	0.018	0.049	0.087
12c-5Cat	2.4	0.5	7.0	0.887	-8.029	0.990	0.235	1.191	22.294	0.015	0.048	0.094

Table 19. Summary of Empirical and Rescaled G^{2*} Sampling Distributions for the Two Category Items in Test 2Cat 12a

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	*a*	*Avg b*	*Info*	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
1a-2Cat	0.7	-1.5	0.5	0.179	-0.355	0.992	0.044	0.238	5.283	0.014	0.057	0.094
2a-2Cat	1.0	-1.5	0.8	0.201	-0.170	0.978	0.155	0.234	6.982	0.018	0.056	0.086
3a-2Cat	1.4	-1.5	1.2	0.190	0.490	0.978	0.311	0.152	14.445	0.016	0.044	0.097
4a-2Cat	1.7	-1.5	1.6	0.212	0.803	0.985	0.267	0.151	17.954	0.019	0.060	0.098
5a-2Cat	2.1	-1.5	2.0	0.240	1.198	0.990	0.228	0.155	21.614	0.016	0.045	0.106
6a-2Cat	2.4	-1.5	2.3	0.265	1.432	0.994	0.147	0.166	22.970	0.013	0.061	0.093
7a-2Cat	0.7	1.0	0.5	0.169	-0.351	0.978	0.110	0.237	4.918	0.013	0.047	0.082
8a-2Cat	1.0	1.0	0.9	0.163	-0.112	0.981	0.144	0.186	7.311	0.015	0.047	0.093
9a-2Cat	1.4	1.0	1.3	0.150	0.130	0.974	0.269	0.144	10.320	0.020	0.055	0.089
10a-2Cat	1.7	1.0	1.6	0.148	0.142	0.962	0.391	0.144	10.256	0.018	0.054	0.088
11a-2Cat	2.1	1.0	2.1	0.121	0.271	0.974	0.352	0.100	13.530	0.023	0.052	0.095
12a-2Cat	2.4	1.0	2.4	0.104	0.334	0.983	0.284	0.078	16.291	0.021	0.057	0.104

Table 20. Summary of Empirical and Rescaled G^{2*} Sampling Distributions for the Two Category Items in Test 2Cat36

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	*a*	*Avg b*	*Info*	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
1a-2Cat	0.7	-1.5	0.5	0.414	-0.683	0.997	0.021	0.515	5.903	0.016	0.053	0.092
2a-2Cat	1.0	-1.5	0.8	0.427	-0.486	0.988	0.073	0.505	6.646	0.017	0.049	0.088
3a-2Cat	1.4	-1.5	1.2	0.420	0.073	0.988	0.111	0.422	9.130	0.013	0.047	0.088
4a-2Cat	1.7	-1.5	1.6	0.453	0.260	0.988	0.112	0.434	9.986	0.015	0.060	0.094
5a-2Cat	2.1	-1.5	2.0	0.461	0.713	0.993	0.091	0.397	12.250	0.012	0.054	0.092
6a-2Cat	2.4	-1.5	2.3	0.446	1.120	0.995	0.080	0.350	14.662	0.016	0.057	0.097
7a-2Cat	0.7	1.0	2.1	0.404	-0.634	0.999	0.003	0.493	6.089	0.014	0.053	0.096
8a-2Cat	1.0	1.0	0.5	0.415	-0.540	0.999	0.009	0.489	6.530	0.013	0.051	0.096
9a-2Cat	1.4	1.0	1.6	0.410	-0.429	0.996	0.024	0.470	6.937	0.011	0.040	0.114
10a-2Cat	1.7	1.0	2.4	0.433	-0.529	0.995	0.037	0.511	6.594	0.016	0.048	0.102
11a-2Cat	2.1	1.0	1.1	0.378	-0.464	0.995	0.037	0.447	6.581	0.015	0.056	0.089
12a-2Cat	2.4	1.0	2.2	0.355	-0.513	0.987	0.080	0.440	6.079	0.019	0.049	0.081
1b-2Cat	0.7	-0.5	0.5	0.431	-0.862	0.999	0.005	0.560	5.381	0.008	0.052	0.102
2b-2Cat	1.0	-0.5	0.9	0.413	-0.795	0.993	0.039	0.540	5.411	0.010	0.051	0.092
3b-2Cat	1.4	-0.5	1.4	0.361	-0.559	1.001	0.002	0.437	6.145	0.013	0.051	0.103
4b-2Cat	1.7	-0.5	1.7	0.328	-0.515	1.000	0.004	0.401	6.091	0.011	0.057	0.105
5b-2Cat	2.1	-0.5	2.1	0.301	-0.524	0.999	0.008	0.377	5.790	0.014	0.044	0.087
6b-2Cat	2.4	-0.5	2.4	0.279	-0.529	1.000	0.001	0.356	5.573	0.010	0.050	0.109

Table 20 (cont'd).

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	a	Avg b	Info	Slope	Intercept	Slope	Intercept	γ	ν	α = .01	α = .05	α = .10
7b-2Cat	0.7	0.5	0.5	0.410	-0.749	1.000	0.006	0.518	5.684	0.009	0.061	0.101
8b-2Cat	1.0	0.5	0.9	0.421	-0.841	0.997	0.019	0.552	5.344	0.012	0.056	0.087
9b-2Cat	1.4	0.5	1.4	0.361	-0.604	0.999	0.001	0.447	5.925	0.012	0.050	0.102
10b-2Cat	1.7	0.5	1.7	0.351	-0.623	1.000	-0.002	0.440	5.765	0.009	0.049	0.109
11b-2Cat	2.1	0.5	2.1	0.283	-0.421	1.000	0.006	0.341	6.243	0.010	0.057	0.106
12b-2Cat	2.4	0.5	2.4	0.276	-0.537	0.999	0.006	0.356	5.469	0.011	0.059	0.097
1c-2Cat	0.7	-1.0	0.5	0.400	-0.594	0.999	0.003	0.482	6.236	0.010	0.059	0.110
2c-2Cat	1.0	-1.0	0.9	0.389	-0.470	0.998	0.015	0.453	6.688	0.016	0.052	0.109
3c-2Cat	1.4	-1.0	1.3	0.419	-0.392	0.995	0.041	0.476	7.102	0.014	0.053	0.090
4c-2Cat	1.7	-1.0	1.6	0.385	-0.313	0.996	0.028	0.428	7.360	0.017	0.052	0.103
5c-2Cat	2.1	-1.0	2.1	0.369	-0.367	0.992	0.058	0.425	6.948	0.013	0.047	0.088
6c-2Cat	2.4	-1.0	2.4	0.332	-0.369	0.995	0.040	0.386	6.793	0.014	0.049	0.097
7c-2Cat	0.7	1.5	0.5	0.427	-0.828	0.986	0.072	0.569	5.299	0.010	0.041	0.086
8c-2Cat	1.0	1.5	0.8	0.400	-0.288	0.997	0.025	0.439	7.544	0.013	0.052	0.102
9c-2Cat	1.4	1.5	1.2	0.442	-0.020	0.996	0.037	0.448	8.828	0.015	0.057	0.104
10c-2Cat	1.7	1.5	1.6	0.445	0.294	0.990	0.100	0.421	10.196	0.015	0.050	0.094
11c-2Cat	2.1	1.5	2.0	0.483	0.489	0.985	0.167	0.445	10.861	0.014	0.050	0.088
12c-2Cat	2.4	1.5	2.3	0.456	1.048	0.987	0.186	0.369	13.950	0.014	0.051	0.095

Table 21. Summary of Empirical and Rescaled G^{2*} Sampling Distributions for the Five Category Items in Test 5Cat12a

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	*a*	*Avg b*	*Info*	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
1a-5Cat	0.7	-0.5	0.9	0.431	-4.336	0.998	0.038	0.586	21.271	0.014	0.059	0.103
2a-5Cat	1.0	-0.5	1.7	0.438	-4.159	0.998	0.037	0.584	22.136	0.014	0.056	0.107
3a-5Cat	1.4	-0.5	3.1	0.431	-3.310	0.998	0.050	0.540	24.950	0.014	0.048	0.094
4a-5Cat	1.7	-0.5	4.4	0.460	-3.658	0.999	0.030	0.581	24.584	0.014	0.051	0.098
5a-5Cat	2.1	-0.5	6.2	0.461	-3.131	0.993	0.183	0.570	26.087	0.014	0.060	0.093
6a-5Cat	2.4	-0.5	7.6	0.471	-2.490	0.998	0.046	0.548	29.001	0.012	0.050	0.099
7a-5Cat	0.7	0.5	0.8	0.432	-4.203	0.999	0.009	0.577	21.946	0.008	0.048	0.085
8a-5Cat	1.0	0.5	1.6	0.415	-3.436	0.999	0.023	0.530	24.090	0.014	0.053	0.096
9a-5Cat	1.4	0.5	2.9	0.419	-3.122	0.999	0.021	0.520	25.443	0.011	0.049	0.108
10a-5Cat	1.7	0.5	4.0	0.451	-3.636	0.998	0.047	0.573	24.356	0.015	0.058	0.101
11a-5Cat	2.1	0.5	5.5	0.440	-2.254	0.998	0.053	0.509	29.321	0.003	0.057	0.112
12a-5Cat	2.4	0.5	6.8	0.491	-3.228	0.999	0.042	0.594	26.834	0.013	0.060	0.109

Table 22. Summary of Empirical and Rescaled G^{2*} Sampling Distributions for the Five Category Items in Test 5Cat36

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	*a*	*Avg b*	*Info*	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
1a-5Cat	0.7	-0.5	0.9	0.654	-1.988	0.998	0.049	0.712	33.029	0.012	0.061	0.112
2a-5Cat	1.0	-0.5	1.7	0.667	-2.325	0.999	0.014	0.734	32.286	0.011	0.050	0.104
3a-5Cat	1.4	-0.5	3.1	0.642	-0.964	0.999	0.027	0.669	35.988	0.011	0.052	0.107
4a-5Cat	1.7	-0.5	4.4	0.671	-1.457	0.999	0.040	0.712	34.682	0.009	0.052	0.105
5a-5Cat	2.1	-0.5	6.2	0.716	-1.847	0.999	0.028	0.769	33.939	0.014	0.050	0.091
6a-5Cat	2.4	-0.5	7.6	0.681	0.149	0.999	0.055	0.679	39.330	0.009	0.041	0.108
7a-5Cat	0.7	0.5	0.8	0.644	-1.750	0.999	0.052	0.695	33.622	0.013	0.053	0.104
8a-5Cat	1.0	0.5	1.6	0.652	-1.583	0.998	0.086	0.699	34.091	0.012	0.049	0.095
9a-5Cat	1.4	0.5	2.9	0.644	-0.983	1.000	0.015	0.671	35.975	0.013	0.047	0.099
10a-5Cat	1.7	0.5	4.0	0.671	-1.486	0.999	0.030	0.713	34.622	0.014	0.044	0.110
11a-5Cat	2.1	0.5	5.5	0.707	-1.709	0.999	0.046	0.756	34.203	0.011	0.052	0.099
12a-5Cat	2.4	0.5	6.8	0.742	-2.139	0.999	0.040	0.804	33.336	0.012	0.056	0.093
1b-5Cat	0.7	0.0	0.8	0.649	-1.671	0.999	0.038	0.697	33.927	0.015	0.050	0.097
2b-5Cat	1.0	0.0	1.6	0.655	-1.331	0.998	0.054	0.694	34.916	0.014	0.062	0.101
3b-5Cat	1.4	0.0	2.9	0.640	-0.285	0.998	0.074	0.650	37.957	0.014	0.058	0.101
4b-5Cat	1.7	0.0	4.1	0.663	-0.830	0.999	0.038	0.687	36.452	0.013	0.052	0.095
5b-5Cat	2.1	0.0	5.8	0.706	-1.420	0.999	0.019	0.746	35.027	0.010	0.054	0.100
6b-5Cat	2.4	0.0	7.1	0.761	-3.125	0.998	0.062	0.855	31.048	0.015	0.059	0.093

Table 22 (cont'd).

	Item Statistics			Empirical		Rescaled		Scaling Corrections		Type I Error Rates		
Item	a	Avg b	Info	Slope	Intercept	Slope	Intercept	γ	ν	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
7b-5Cat	0.7	0.0	0.9	0.635	-1.625	0.999	0.021	0.681	34.022	0.008	0.051	0.105
8b-5Cat	1.0	0.0	1.7	0.626	-0.719	0.998	0.078	0.648	36.571	0.015	0.049	0.092
9b-5Cat	1.4	0.0	3.2	0.651	-1.040	0.999	0.043	0.681	35.770	0.013	0.053	0.099
10b-5Cat	1.7	0.0	4.5	0.670	-0.943	0.999	0.042	0.697	36.139	0.007	0.057	0.101
11b-5Cat	2.1	0.0	6.3	0.730	-1.126	0.999	0.024	0.761	35.925	0.009	0.055	0.108
12b-5Cat	2.4	0.0	7.8	0.761	-0.597	0.998	0.087	0.781	37.260	0.014	0.058	0.100
1c-5Cat	0.7	-0.5	0.8	0.634	-1.491	1.000	0.013	0.675	34.397	0.012	0.054	0.092
2c-5Cat	1.0	-0.5	1.6	0.625	-0.386	1.000	0.021	0.636	37.732	0.013	0.049	0.102
3c-5Cat	1.4	-0.5	2.9	0.656	-1.452	0.999	0.045	0.698	34.596	0.009	0.044	0.109
4c-5Cat	1.7	-0.5	4.0	0.648	-0.910	0.999	0.033	0.674	36.178	0.007	0.049	0.105
5c-5Cat	2.1	-0.5	5.7	0.631	-0.306	1.000	0.013	0.640	38.011	0.008	0.048	0.111
6c-5Cat	2.4	-0.5	7.0	0.670	-1.112	1.000	0.012	0.700	35.721	0.010	0.054	0.104
7c-5Cat	0.7	0.5	0.8	0.656	-2.449	0.999	0.016	0.727	31.838	0.011	0.055	0.102
8c-5Cat	1.0	0.5	1.6	0.626	-0.455	0.999	0.046	0.640	37.458	0.014	0.052	0.101
9c-5Cat	1.4	0.5	2.9	0.647	-1.280	0.999	0.029	0.683	35.066	0.010	0.058	0.105
10c-5Cat	1.7	0.5	4.0	0.633	-0.727	1.000	0.017	0.653	36.704	0.009	0.049	0.109
11c-5Cat	2.1	0.5	5.7	0.672	-1.870	1.000	0.009	0.724	33.599	0.012	0.055	0.096
12c-5Cat	2.4	0.5	7.0	0.666	-0.912	0.999	0.055	0.693	36.192	0.009	0.041	0.108

IV. A. 4. a) Slopes and Intercepts of Empirically Generated Distributions

The slopes and intercepts of regression lines fitted to Q-Q plots of the empirically generated sampling distributions were evaluated to assess whether the empirical distributions may follow the null chi-square distributions. The columns under the "Empirical" heading in Tables 15 to 22 show that across items and test configurations, the slopes and intercepts differed from 1 and 0, respectively, for both the Pearson χ^{2*} and likelihood ratio G^{2*} forms of the statistic. These differences indicated that the empirically generated sampling distributions did not follow the null chi-square distributions, and supported the results shown in Figures 7 to 10. The standard errors associated with the regression coefficients were small due to the large sample size of 1000 statistics. Generally, the standard errors for the slopes were less than 0.01, and the standard errors for the intercepts were less than 0.1.

Further analysis of these slopes and intercepts showed that within item score category levels, as test length increased, the slopes also increased. Further, across tests, the slopes and intercepts were closer to 1 and 0, respectively, for the higher score response category levels. That is, for tests consisting of 5-category items, the slopes were closer to 1 and intercepts closer to 0 than they were for tests at the lower score category levels. These results indicated that the null chi-square distributions were a better match to the empirically generated sampling distributions for items on longer tests and for items with more score category levels.

In comparing the Pearson to likelihood ratio forms of the statistics, the slopes were similar for items on the 2-category 12 items tests. However, as the number of item score categories increased, the differences between the slopes associated with χ^{2*} and G^{2*} also increased. For the 3-, 4-, and 5-category tests, across test length, the slopes associated with the Pearson statistics were larger than those associated with the likelihood ratio statistics. This

could be due to the fact that the Pearson form of the statistic showed more deviations in the tails of the distributions than the likelihood ratio form, which resulted in increased slope values for regression lines fitted to the plots. For the 2-category tests, the slopes for the Pearson and likelihood ratio distributions were more similar.

IV. A. 4. b) Slopes and Intercepts of Rescaled Distributions

The empirically generated χ^{2*} and G^{2*} distributions were rescaled by scaling corrections obtained using the method of moments with Equations 19 and 20. The slopes and intercepts of regression lines fitted to the rescaled empirical sampling distributions are found in the columns labeled "Rescaled" in Tables 15 to 22.

Generally, across item score category levels and tests, the slopes of regression lines fitted to Q-Q plots were close to 1, and the intercepts close to 0. The standard errors associated with the regression coefficients were small due to the large sample size of 1000 statistics. Generally, the standard errors for the slopes were less than 0.01, and the standard errors for the intercepts were less than 0.1. Even considering the small standard errors, Tables 15 to 22 show that for many of the items, the slopes and intercepts do not show significant deviations from 1 and 0, respectively.

The match between the rescaled distributions and theoretical chi-square distributions for both the Pearson and likelihood ratio statistics was greater for items with higher numbers of score response categories. For example, for the Pearson χ^{2*} statistics for 2-category items, Tables 15 and 16 show that the rescaled slopes and intercepts exhibited small differences from 1 and 0, respectively. Tables 19 and 20 show improvements in the slopes and intercepts for the G^{2*} distributions, but small differences from 1 and 0 were still present.

For the tests containing 5-category items, as seen in Tables, 17, 18, 21, and 22, the slopes and intercepts associated with both χ^{2*} and G^{2*} were close to 1 and 0, respectively. This indicated that the rescaled distributions for items with more score response categories matched the theoretical chi-square distributions more closely than the rescaled distributions of items with fewer score response categories. That is, the scaling corrections obtained worked better for items with more score response categories.

For the 2- and 3- item score category levels (tables available from the author) the slopes for the rescaled distributions were closer to 1, and the intercepts closer to 0, as test length increased. The slopes and intercepts of items on the shorter 12 item tests deviated more from 1 and 0, respectively, than the items on longer tests. This indicated that the scaling corrections worked slightly better for 2- and 3-category items on longer tests. For the 4- and 5-category items, at all test lengths, the slopes and intercepts were close to 1 and 0, respectively.

In general, across items and tests, the slopes and intercepts were closer to 1 and 0, respectively, for the rescaled likelihood ratio G^{2*} distributions than for the rescaled Pearson χ^{2*} distributions. This could be due to deviations from the theoretical chi-square distribution seen in the tails of the Pearson sampling distributions.

To summarize, the slopes and intercepts associated with the rescaled distributions were close to 1 and 0, respectively. This indicated that overall, the sampling distributions of χ^{2*} and G^{2*} statistics were consistent with scaled chi-square distributions, after rescaling the statistics using the scaling corrections obtained with the method of moments..

For the items with higher numbers of score response categories (4- and 5-category), the slopes were close to 1 and intercepts close to 0 for all test lengths. Small differences between 1 and 0 for the slopes and intercepts were seen at the lower score category levels (2- and 3-

category). Further, for the lower score category levels, the slopes and intercepts of the rescaled distributions were closer to 1 and 0, respectively, for items on longer tests. In addition, in general, the slopes and intercepts associated with the likelihood ratio distributions were slightly closer to the 1 and 0 than the slopes and intercepts associated with the Pearson statistics.

IV. A. 4. c) Type I Error Rates for Rescaled Distributions

The Type I error rates associated with the rescaled distributions were evaluated to further examine the data in the tails of the sampling distributions. Tables 15 to 22 present the observed proportion of Type I errors across 1000 replications for each item at $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$.

To account for sampling error, 95% confidence intervals for the expected proportion of Type I errors were considered. Across 1000 replications, Type I error rates of .004 to .016, .037 to .064, and .081 to .119, were expected for $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$, respectively. In Tables 15 to 22, Type I error rates that fell in these expected ranges were underlined.

For most items, across item score category levels and test lengths, nominal Type I error rates were observed for the Pearson and likelihood ratio forms of the statistic. At $\alpha = .01$, some items had inflated Type I error rates. However, those same items met the nominal levels at $\alpha = .05$ and $\alpha = .10$. This indicated that Type I error rates were affected only in the extreme tails of the distributions. Overall, nominal Type I error rates were observed.

IV. A. 4. d) Scaling Corrections Obtained Using the Method of Moments

The values of the scaling corrections associated with each item, the scaling factor γ and the degrees of freedom value ν, are presented in Tables 15 to 22. These values were obtained

using the means and variances of the empirical sampling distributions with the method of moments, and are given by Equations 19 and 20.

Within the tests, for both the Pearson and likelihood ratio distributions, the scaling factors were similar across items. However, there was significant variation in the degrees of freedom values within tests. For the likelihood ratio statistics, the degrees of freedom values tended to increase with larger *a* values. This pattern was not seen for the Pearson statistics.

Items with varying *a* parameters were included in the study because it was thought that this variable could aid in the prediction of the scaling corrections. While the influence of the *a* parameter on the scaling factors was not apparent through inspection of Tables 15 to 22, the results did indicate a relationship between the discrimination parameter and the degrees of freedom for the G^{2*} distributions.

Items with larger discrimination parameters had larger item fit statistics, and larger values for the means of their sampling distributions. Inspection of several item fit tables indicated that this could be due to the fact that for highly discriminating items, the score distributions across examinees were spread over fewer rows of the item fit tables, and the expectations in the cells at the extremes of the ability scale were small. Item fit tables that had several cells with small cell expectations resulted in larger test statistics. This could be related to the item discrimination parameter.

Within score category levels, for both the Pearson and likelihood ratio distributions, the scaling factors increased as test length increased. This indicated that less adjustment was required to transform the empirical sampling distributions so that they closely approximated theoretical chi-square distributions for items on longer tests. The increase in scaling factors as test length increased was expected, because longer tests are associated with smaller values of the

posterior variance of ability. As the variance of the posterior distribution of ability decreases, the dependence among pseudocounts in the item fit table also decreases, and less adjustment is needed for the pseudocounts-based statistic to follow theoretical chi-square distribution. Also, within score category levels, the degrees of freedom values became less variable as test length increased.

For the Pearson and likelihood ratio statistics, as the number of score categories increased, the scaling factors and degrees of freedom values also increased. It was expected that the degrees of freedom would be larger for items with more score category levels, because the number of item score response categories (which is the number of columns in item fit tables) is used in the computation of the degrees of freedom for item fit statistics.

The increase in the scaling factors for items with more score response categories indicated that less adjustment was required to transform the empirical sampling distributions for items with more score response categories. One reason for this relationship could be that items with more score response categories provide more information than items with fewer categories. Since test information is inversely related to the standard error of the ability estimates, the higher the test information, the lower the posterior variances, and vice versa. Therefore, the higher the test information, the greater the precision in estimating examinee ability on the test, and the less adjustment required in transforming the fit statistics for items on the test.

In general, the scaling factors and degrees of freedom values were higher for the Pearson than likelihood ratio statistics. Values associated with the Pearson statistics were also more variable than values associated with the likelihood ratio statistics.

To summarize, patterns in the scaling factors and degrees of freedom values obtained using the method of moments were more apparent for the likelihood ratio statistics than for the

Pearson statistics. In general, for both forms of the fit statistics, scaling factors were larger for longer tests, and for tests with higher numbers of score response categories. Degrees of freedom values associated with the likelihood ratio statistics tended to increase as *a* increased within tests, and became less variable as test length increased. The relationship between *a* and degrees of freedom was not apparent for the Pearson statistics. Patterns in the values of the scaling corrections were relevant, since the current study involved predicting these values from item and sample characteristics.

IV. A. 5. Mean Posterior Variance

Table 23 presents the values of the mean posterior variance, total test information over the range $-3 \leq \theta \leq 3$, and total test information over the range $-1 \leq \theta \leq 1$, for each test. Table 23 shows that generally, within each item category level, the value of the mean posterior variance decreased as test length increased. Total test information, computed over the range $-3 \leq \theta \leq 3$, increased as test length increased. These relationships were expected, because the spread of the posterior distributions of ability decreased as test length increased. The exceptions to this pattern were Test 2Cat12b and Test 3Cat12c.

Tests 2Cat12b and 3Cat12c resulted in smaller values of the mean posterior variance than the other two 12-items tests at the corresponding item category level. Test construction could be one possible reason for the differences in mean posterior variances for these two tests. Items with difficulty parameters that match the ability parameters of the examinees tend to have smaller standard errors, and therefore provide more information. Tests 2Cat12b and 3Cat12c consisted of items with difficulty parameters that were closer to 0, the mean of the underlying ability distribution, and provided more information near the $\theta = 0$ ability range than other tests.

To illustrate this point, the total test information over the range $-1 \leq \theta \leq 1$ is provided in Table 23. For tests 2Cat12b and 3Cat12c, the total information in this range is greater than it is for the other 12 item tests at the 2- and 3-category levels. Within the 4- and 5-category levels, values of the mean posterior variance and test information were similar for the three 12-item tests.

In addition, Table 23 shows that the values of the mean posterior variance were smaller, and test information larger, for items with more score response categories. This relationship was also expected, because polytomous items tend to be more informative than dichotomous items.

Table 23. Mean Posterior Variance for Each of the 20 Test Configurations

Test	$\overline{p}$	Test Information $-3 \leq \theta \leq 3$	Test Information $-1 \leq \theta \leq 1$
2cat12a	0.322	17.3	6.9
2cat12b	0.162	17.9	11.7
2cat12c	0.320	17.3	6.9
2cat24	0.119	35.2	18.6
2cat36	0.094	52.5	25.5
3cat12a	0.414	29.6	6.3
3cat12b	0.383	30.2	7.3
3cat12c	0.228	31.9	10.7
3cat24	0.269	59.8	13.6
3cat36	0.145	91.7	24.3
4cat12a	0.135	41.6	15.3
4cat12b	0.123	42.1	16.3
4cat12c	0.113	40.7	17.6
4cat24	0.069	83.7	31.6
4cat36	0.045	124.4	49.2
5cat12a	0.106	45.4	18.3
5cat12b	0.099	46.8	19.2
5cat12c	0.094	44.1	19.7
5cat24	0.054	92.2	37.5
5cat36	0.036	136.3	57.2

IV. A. 6. Summary of Empirically Generated and Rescaled χ^{2*} and G^{2*} Sampling Distributions

The empirically generated and rescaled sampling distributions of χ^{2*} and G^{2*} statistics were evaluated to provide evidence that the statistics may follow scaled chi-square distributions.

This involved analysis of 1) the linearity of Q-Q plots, 2) the values of slopes and intercepts of regression lines fitted to Q-Q plots, and 3) Type I error rates.

Q-Q plots of the original empirically generated distributions were linear, with slopes and intercepts different from 1 and 0, respectively, indicating that the sampling distributions followed one of the family of theoretical chi-square distributions. The scaling corrections were estimated using the method of moments, and the empirical sampling distributions were rescaled by the scaling corrections.

Generally, Q-Q plots of the rescaled sampling distributions of χ^{2*} and G^{2*} statistics versus theoretical chi-square distributions were linear and fell along the line $y = x$, suggesting that the statistics followed scaled chi-square distributions, where the scaling corrections could be obtained using the method of moments. The slopes and intercepts of regression lines fitted to Q-Q plots of rescaled empirical versus theoretical chi-square distributions were generally close to 1 and 0, respectively, and Type I error rates fell in the range expected with 95% confidence intervals.

For several items, the Q-Q plots showed deviations from the theoretical chi-square distribution in the tails of the distribution. More of the 2- and 3-category items, than 4- and 5-category items, showed these deviations. The analysis of Type I error rates provided further analysis of the data in the upper tails of the sampling distributions. While the Q-Q plots exhibited some deviations from linearity in the tails, and the slopes of regression lines fitted to Q-Q plots showed deviations from 1 and 0, respectively, Type I error rates were not affected.

Overall, the sampling distributions of both the Pearson and likelihood ratio pseudocounts based fit statistics were well approximated by theoretical chi-square distributions. In general, the rescaled likelihood ratio G^{2*} statistics provided a closer match to the theoretical chi-square

distributions than the Pearson χ^{2*} statistics. In addition, the sampling distributions of items with more score response categories and on longer tests provided a better match to theoretical chi-square distributions than items with fewer score response categories, and items on shorter tests.

Patterns in the scaling correction values were more apparent for the likelihood ratio statistics than for the Pearson statistics. Generally, scaling factors and degrees of freedom values were larger for longer tests, and for tests consisting of items with more score response categories. These tests also had smaller mean posterior variances.

IV. B. Results from the Predicted Fit Statistic Distributions

Having found evidence suggesting that the empirically generated sampling distributions of χ^{2*} and G^{2*} statistics may follow scaled chi-square distributions, steps were taken to predict the scaling corrections using information from the test items and the sample responses. To this end, the data obtained from the empirically generated sampling distributions were analyzed. The analyzed data set is found in Appendix C.

IV. B. 1. Fitting Multilevel Prediction Equations

The observations in the data file shown in Appendix C were not independent. For example, all items in a single test configuration had the same value of the mean posterior variance. For this reason, multilevel prediction models were fitted to the data. Multilevel models are appropriate when data is provided at two levels in a hierarchy, and interest lies in examining the behavior of an outcome as a function of both levels of the hierarchy. In this study, the outcomes (scaling factor and degrees of freedom values) were functions of both item-level and test-level data.

Several multilevel prediction equations were estimated from the data to determine the best set of independent variables for predicting the scaling factor and degrees of freedom values for both the χ^{2*} and G^{2*} distributions. Models that included the item level predictor variables item discrimination (*a)*, total item information (*info*), number of item score response categories (*ncat*), and test level predictor variables mean posterior variance ($\overline{p}$), and root mean posterior variance ($\sqrt{\overline{p}}$) were estimated.

Four final multilevel models were selected. One set of two equations predicted the scaling factors and degrees of freedom for χ^{2*}. The second set of two equations predicted the scaling factors and degrees of freedom for the G^{2*} distributions.

Determining the best multilevel prediction equations involved several steps. Initially, for each dependent variable, an unconditional means model was fitted to the data. This model examined the variation in the dependent variable across tests, and provided a baseline against which more complex models were compared. The unconditional means models indicated that there was more variation in the dependent variables between tests than within tests. For the four dependent variables, the unconditional means models indicated that the percentages of the total variation that occurred between tests were 67% for the χ^{2*} scaling factors, 75% for the χ^{2*} degrees of freedom, 88% for the G^{2*} scaling factors, and 84% for the G^{2*} degrees of freedom.

Additional models containing item and test level predictors were examined, and the proportion of the *explainable variation* that was explained by each model was computed. For multilevel models, the fraction of explainable variation that is explained by adding variables to the models can be computed. This interpretation for multilevel models is different than the interpretation of r^2 in traditional regression analyses.

For the current analyses, the proportion of explained variance was computed using several models. In addition, the statistical significance of the parameter estimates for each variable in each model was assessed. The final models were chosen because they contained the set of independent variables that explained the largest amount of variance, and at the same time contributed parameter estimates to the model that were statistically significant. The exception was the final model for predicting the scaling factor for the likelihood ratio form of the statistic. The final model for this variable explained the largest amount of variation, but included coefficients that were not all statistically significant due to high correlations between the variables. For consistency, the same predictors were used to predict the scaling factors for the Pearson and likelihood ratio forms of the statistic.

Table 24 provides a summary of the proportion of explained variance for the unconditional means model and final multilevel prediction models. The proportion of variance that occurred between tests for each unconditional means models is provided. Also, the proportion of variance that was explained by the final models, between the tests and within the tests, is shown. These proportions show that the final models for estimating the scaling factors and degrees of freedom values accounted for much of the between test variance (67.8% to 92.0%). The final models for predicting the scaling corrections accounted for less of the within test variance (30.3% to 53.4%).

Table 24. Explained Variance for the Multilevel Prediction Equations

	Pearson		Likelihood ratio	
	$\hat{\gamma}$	$\hat{\nu}$	$\hat{\gamma}$	$\hat{\nu}$
Unconditional Means Model Between Test Variance	0.666	0.762	0.876	0.837
Final Model Between Test Variance	0.678	0.828	0.892	0.920
Final Model Within Test Variance	0.303	0.452	0.384	0.534

The final prediction models contained both item level and test level predictors, and both fixed and random effects. For the four final models, the intercept and item discrimination parameter were included as random effects. As a result, regression coefficients for the intercept term and discrimination parameter were estimated for each category of the test level variable. That is, two random effects coefficients were estimated for each test configuration. A single coefficient for each fixed effect in each model was also estimated.

Table 25 presents the fixed and random effects coefficients for each of the final multilevel prediction equations. The coefficients for each of the final models are presented down the columns of the table.

Table 25. Estimated Coefficients for the Multilevel Prediction Equations

		Pearson				Likelihood ratio			
		$\hat{\gamma}$		$\hat{\nu}$		$\hat{\gamma}$		$\hat{\nu}$	
Fixed Effects	*Intercept*	-0.375		10.413		0.621		25.927	
	a	0. 114		1.607		-0.003		4.102	
	$\overline{p}$	-5.983		75.695		-1.044		146.700	
	$\sqrt{\overline{p}}$	4.904		-70.983		0.096		-147.190	
				4.336				5.197	
Random Effects	*ncat*	*int*	a_{int}	*int*	a_{int}	*int*	a_{int}	*int*	a_{int}
	2cat12a	-0.20	-0.15	-2.63	3.36	-0.06	-0.07	-1.56	4.84
	2cat12b	-0.27	-0.19	2.15	-1.65	-0.10	-0.09	2.83	-3.41
	2cat12c	-0.20	-0.15	-3.09	3.94	-0.06	-0.07	-1.48	4.82
	2cat24	-0.17	-0.12	1.81	-1.51	-0.05	-0.07	0.85	-1.94
	2cat36	-0.08	-0.01	1.75	-2.17	-0.01	-0.06	0.09	-2.57
	3cat12a	0.08	0.08	-5.00	1.64	0.04	0.05	-2.96	1.80
	3cat12b	0.02	0.01	-8.67	6.78	0.03	0.05	-4.29	5.71
	3cat12c	-0.09	-0.05	-1.65	1.58	-0.01	0.00	0.07	1.77
	3cat24	0.12	-0.01	-6.07	3.98	0.07	0.03	-2.32	3.43
	3cat36	0.15	0.05	-2.59	1.23	0.04	0.05	0.24	0.12
	4cat12a	-0.03	-0.03	-1.69	0.86	0.00	0.00	-0.34	-0.01
	4cat12b	-0.05	-0.03	-1.94	2.08	0.00	0.01	-1.28	0.81
	4cat12c	-0.05	-0.02	1.64	-1.13	0.00	0.00	0.59	-1.33
	4cat24	0.14	0.09	0.34	-1.02	0.02	0.04	-0.12	-1.93
	4cat36	0.19	0.18	5.19	-4.11	0.01	0.05	1.47	-3.35
	5cat12a	0.03	0.00	-1.48	0.51	0.01	0.01	-0.03	-0.51
	5cat12b	0.02	0.00	-0.07	0.23	0.01	0.01	-0.33	-0.12
	5cat12c	-0.03	0.00	5.76	-3.78	0.00	0.00	2.24	-2.70
	5cat24	0.18	0.13	5.10	-3.76	0.03	0.03	2.40	-2.53
	5cat36	0.24	0.22	11.16	-7.05	0.04	0.04	3.93	-2.89

To further illustrate the form of the prediction equations, Equations 30 and 31 present the multilevel equations for predicting the scaling factor and degrees of freedom for the Pearson χ^{2*} statistics. Equations 32 and 33 present the multilevel equations for predicting the scaling factor and degrees of freedom for the likelihood ratio G^{2*} statistics. In Equations 30 to 33, the terms 'a_{int}' and '*int*' would be replaced by the numeric values of the appropriate random intercept terms found in Table 25.

$$\hat{\gamma}_{\chi_2^*} = -0.375 + 0.0114 * a - 5.983 * \bar{p} + 4.904 * \sqrt{\bar{p}} + a_{int} * a + int \tag{30}$$

$$\hat{\nu}_{\chi_2^*} = 10.413 + 1.607 * a + 75.695 * \bar{p} - 70.983 * \sqrt{\bar{p}} + 4.336 * ncat + a_{int} * a + int \tag{31}$$

$$\hat{\gamma}_{G_2^*} = 0.621 - 0.003 * a - 1.044\bar{p} + .096 * \sqrt{\bar{p}} + a_{int} * a + int \tag{32}$$

$$\hat{\nu}_{G_2^*} = 25.927 + 4.102 * a + 146.70 * \bar{p} - 147.190 * \sqrt{\bar{p}} + 5.197 * ncat + a_{int} * a + int \tag{33}$$

Table 25 presents the multiple random effects coefficients for each of the final multilevel prediction equations. However, in applying the multilevel prediction equations, only one random effects coefficient would be used in each of Equations 30 to 33. Therefore, a single value of the random effect intercept term would have to be selected from Table 25 when applied to real test items. One suggestion employed in this study is to choose the coefficient from Table 25 for the test from the simulation study that is most similar to the real test, in terms of number of item score categories, test length, and total test information.

IV. B. 2. Results Based on Multilevel Prediction Equations

The multilevel models were fitted to the data found in Appendix C, and the predicted scaling corrections for χ^{2*} and G^{2*} for each item were obtained. The empirical sampling

distributions for each item were then rescaled by the predicted scaling corrections $\hat{\gamma}$, and compared to the theoretical chi-square distributions with predicted degrees of freedom $\hat{\nu}$.

The correspondence between the χ^{2*} and G^{2*} sampling distributions that were rescaled by the predicted scaling corrections $\hat{\gamma}$ and the theoretical chi-square distributions with predicted degrees of freedom $\hat{\nu}$ was assessed using Q-Q plots and Type I error rates. Specifically, the linearity of the Q-Q plots was assessed, and the slopes and intercepts of regression lines fitted to Q-Q plots were examined for their closeness to 1 and 0, respectively. Further, Type I error rates were computed to evaluate the behavior in the tails of the distributions.

Tables 26 to 33 present results from the analysis of Q-Q plots obtained using the predicted scaling corrections for the items in Tests 2Cat12a, 2Cat36, 5Cat12a, and 5Cat36, respectively. Tables 26 to 33 present the scaling corrections, slopes and intercepts of regression lines fitted to Q-Q plots, and Type I error rates associated with both the empirically generated and predicted distributions. The presentation of results in this format allows for the comparison of the results obtained using the prediction equations to the results obtained using the method of moments.

Tables 26 to 29 present the results of the Pearson χ^{2*} statistics for Tests 2Cat12a, 2Cat36, 5Cat12a, and 5Cat36, respectively. Tables 30 to 33 present results for the likelihood ratio G^{2*} statistics for the same four tests.

Tables for only this subset of tests are presented below to simplify the discussion, and at the same time allow for comparisons between the Pearson and likelihood ratio forms of the statistics across number of item score category levels and test lengths. Tables for the additional test configurations are available from the author.

Table 26. Comparison of Rescaled and Predicted χ^{2*} Sampling Distributions for the Two Category Items in Test 2Cat12a

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-2Cat	0.253	4.892	0.261	4.025	0.984	0.079	1.054	0.490	0.014	0.054	0.086	0.023	0.074	0.116
2a-2Cat	0.290	5.666	0.251	5.516	0.965	0.194	1.131	0.303	0.016	0.054	0.078	0.041	0.077	0.122
3a-2Cat	0.278	8.163	0.238	7.505	0.928	0.593	1.134	1.035	0.015	0.038	0.065	0.030	0.092	0.156
4a-2Cat	0.266	10.585	0.229	8.996	0.948	0.550	1.202	1.506	0.021	0.046	0.077	0.052	0.127	0.211
5a-2Cat	0.238	14.537	0.216	10.985	0.963	0.539	1.231	2.551	0.016	0.041	0.080	0.072	0.215	0.313
6a-2Cat	0.222	17.716	0.206	12.476	0.989	0.185	1.272	3.174	0.017	0.061	0.091	0.100	0.274	0.418
7a-2Cat	0.243	4.696	0.261	4.025	0.970	0.139	0.980	0.426	0.014	0.045	0.082	0.014	0.052	0.093
8a-2Cat	0.211	6.317	0.251	5.516	0.967	0.206	0.871	0.490	0.019	0.045	0.078	0.014	0.039	0.057
9a-2Cat	0.191	7.690	0.238	7.505	0.948	0.404	0.769	0.392	0.017	0.051	0.077	0.009	0.023	0.044
10a-2Cat	0.235	6.260	0.229	8.996	0.912	0.555	0.771	-0.491	0.017	0.038	0.067	0.008	0.021	0.027
11a-2Cat	0.161	8.375	0.216	10.985	0.921	0.657	0.595	-0.281	0.016	0.044	0.065	0.003	0.007	0.014
12a-2Cat	0.128	9.692	0.206	12.476	0.905	0.923	0.494	-0.115	0.012	0.034	0.064	0.003	0.003	0.004

Table 27. Comparison of Rescaled and Predicted χ^{2*} Sampling Distributions for the Two Category Items in Test 2Cat36

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-2Cat	0.578	5.160	0.557	5.792	0.990	0.052	0.968	-0.248	0.013	0.057	0.086	0.012	0.047	0.079
2a-2Cat	0.654	5.127	0.589	5.624	0.980	0.100	1.037	-0.136	0.018	0.048	0.090	0.021	0.054	0.098
3a-2Cat	0.870	4.616	0.630	5.399	0.935	0.299	1.187	-0.035	0.014	0.034	0.059	0.026	0.064	0.121
4a-2Cat	0.904	5.099	0.661	5.231	0.956	0.222	1.289	0.225	0.016	0.034	0.078	0.033	0.112	0.176
5a-2Cat	0.659	7.718	0.703	5.007	0.984	0.125	1.154	1.458	0.015	0.053	0.088	0.035	0.128	0.225
6a-2Cat	0.624	8.546	0.734	4.838	0.974	0.221	1.112	1.878	0.017	0.049	0.083	0.040	0.121	0.215
7a-2Cat	0.493	5.931	0.557	5.792	0.999	0.007	0.894	0.067	0.014	0.047	0.103	0.007	0.033	0.064
8a-2Cat	0.507	6.112	0.589	5.624	0.997	0.021	0.895	0.229	0.013	0.047	0.093	0.007	0.034	0.069
9a-2Cat	0.532	5.920	0.630	5.399	0.994	0.038	0.880	0.250	0.014	0.046	0.100	0.010	0.031	0.060
10a-2Cat	0.726	4.629	0.661	5.231	0.987	0.061	1.016	-0.234	0.019	0.056	0.085	0.021	0.056	0.085
11a-2Cat	0.898	3.330	0.703	5.007	0.960	0.134	0.984	-0.675	0.020	0.039	0.065	0.021	0.037	0.056
12a-2Cat	1.086	2.513	0.734	4.838	0.920	0.203	0.948	-0.868	0.011	0.033	0.057	0.013	0.030	0.041
1b-2Cat	0.522	5.568	0.557	5.792	1.000	0.002	0.917	-0.099	0.011	0.056	0.110	0.006	0.033	0.075
2b-2Cat	0.509	5.461	0.589	5.624	0.993	0.040	0.845	-0.032	0.010	0.049	0.091	0.006	0.019	0.050
3b-2Cat	0.414	6.053	0.630	5.399	1.000	0.002	0.695	0.219	0.012	0.049	0.102	0.001	0.007	0.022
4b-2Cat	0.408	5.552	0.661	5.231	0.992	0.049	0.631	0.127	0.010	0.047	0.094	0.002	0.006	0.013
5b-2Cat	0.460	4.485	0.703	5.007	0.979	0.089	0.605	-0.098	0.010	0.044	0.078	0.001	0.006	0.010
6b-2Cat	0.527	3.660	0.734	4.838	0.985	0.054	0.611	-0.331	0.014	0.039	0.074	0.001	0.009	0.014

Table 27 (cont'd).

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
7b-2Cat	0.484	5.854	0.557	5.792	1.000	0.005	0.873	0.030	0.010	0.061	0.100	0.004	0.030	0.072
8b-2Cat	0.512	5.445	0.589	5.624	0.994	0.028	0.851	-0.047	0.010	0.051	0.095	0.006	0.025	0.051
9b-2Cat	0.422	5.855	0.630	5.399	1.000	0.006	0.697	0.156	0.011	0.052	0.097	0.001	0.008	0.025
10b-2Cat	0.485	4.941	0.661	5.231	0.994	0.032	0.708	-0.078	0.008	0.046	0.091	0.003	0.008	0.024
11b-2Cat	0.407	4.928	0.703	5.007	0.993	0.036	0.570	-0.001	0.012	0.045	0.087	0.000	0.005	0.008
12b-2Cat	0.519	3.618	0.734	4.838	0.977	0.082	0.593	-0.311	0.017	0.040	0.075	0.004	0.005	0.017
1c-2Cat	0.471	6.187	0.557	5.792	0.999	0.002	0.873	0.172	0.010	0.054	0.112	0.004	0.027	0.070
2c-2Cat	0.471	6.201	0.589	5.624	0.998	0.017	0.839	0.246	0.010	0.054	0.108	0.004	0.030	0.058
3c-2Cat	0.626	5.280	0.630	5.399	0.971	0.156	0.954	0.100	0.013	0.045	0.078	0.013	0.038	0.073
4c-2Cat	0.589	5.266	0.661	5.231	0.988	0.059	0.883	0.070	0.013	0.049	0.087	0.008	0.037	0.060
5c-2Cat	1.058	2.882	0.703	5.007	0.935	0.192	1.039	-0.867	0.015	0.033	0.057	0.018	0.037	0.054
6c-2Cat	0.857	3.098	0.734	4.838	0.946	0.164	0.868	-0.585	0.017	0.028	0.063	0.013	0.024	0.033
7c-2Cat	0.631	4.704	0.557	5.792	0.983	0.084	0.999	-0.456	0.009	0.046	0.086	0.009	0.042	0.076
8c-2Cat	0.529	6.218	0.589	5.624	0.995	0.031	0.941	0.296	0.013	0.055	0.099	0.012	0.050	0.094
9c-2Cat	0.986	4.256	0.630	5.399	0.947	0.221	1.306	-0.393	0.013	0.043	0.080	0.040	0.104	0.151
10c-2Cat	0.873	5.279	0.661	5.231	0.960	0.211	1.274	0.309	0.017	0.047	0.086	0.043	0.106	0.173
11c-2Cat	1.095	4.691	0.703	5.007	0.916	0.398	1.376	0.413	0.013	0.038	0.065	0.053	0.111	0.207
12c-2Cat	0.693	7.713	0.734	4.838	0.949	0.397	1.144	1.742	0.014	0.045	0.082	0.042	0.121	0.206

Table 28. Comparison of Rescaled and Predicted χ^{2*} Sampling Distributions for the Five Category Items in Test 5Cat12a

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-5Cat	0.700	17.763	0.690	17.004	0.993	0.134	1.028	0.516	0.018	0.052	0.094	0.022	0.067	0.116
2a-5Cat	0.767	17.068	0.723	17.638	0.993	0.118	1.035	-0.168	0.015	0.057	0.099	0.023	0.065	0.112
3a-5Cat	0.787	17.762	0.767	18.483	0.992	0.152	0.997	-0.200	0.020	0.048	0.089	0.019	0.048	0.084
4a-5Cat	0.872	17.326	0.800	19.118	0.994	0.103	1.030	-0.817	0.017	0.057	0.092	0.017	0.058	0.095
5a-5Cat	0.818	19.627	0.844	19.963	0.989	0.220	0.950	0.059	0.015	0.051	0.097	0.009	0.039	0.074
6a-5Cat	0.797	22.056	0.877	20.597	0.998	0.051	0.938	0.713	0.014	0.059	0.109	0.010	0.044	0.089
7a-5Cat	0.656	19.251	0.690	17.004	0.999	0.024	1.011	1.121	0.016	0.046	0.091	0.022	0.069	0.127
8a-5Cat	0.874	14.791	0.723	17.638	0.950	0.736	1.049	-0.629	0.013	0.039	0.067	0.018	0.049	0.093
9a-5Cat	0.725	18.786	0.767	18.483	0.991	0.176	0.943	0.309	0.014	0.044	0.094	0.010	0.035	0.074
10a-5Cat	0.884	16.676	0.800	19.118	0.990	0.161	1.020	-1.084	0.017	0.049	0.088	0.017	0.048	0.082
11a-5Cat	0.751	21.460	0.844	19.963	0.996	0.077	0.919	0.739	0.011	0.053	0.101	0.005	0.030	0.070
12a-5Cat	0.853	20.699	0.877	20.597	0.998	0.032	0.973	0.080	0.009	0.058	0.108	0.007	0.048	0.098

Table 29. Comparison of Rescaled and Predicted χ^{2*} Sampling Distributions for the Five Category Items in Test 5Cat36

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-5Cat	0.808	28.671	0.812	28.696	0.997	0.093	0.991	0.078	0.014	0.054	0.105	0.012	0.050	0.100
2a-5Cat	0.931	25.419	0.914	27.062	0.996	0.097	0.984	-0.712	0.015	0.050	0.101	0.014	0.042	0.080
3a-5Cat	1.157	21.565	1.049	24.884	0.981	0.413	1.006	-1.237	0.015	0.051	0.090	0.013	0.046	0.082
4a-5Cat	1.536	17.207	1.150	23.251	0.965	0.600	1.104	-2.686	0.015	0.039	0.071	0.027	0.054	0.083
5a-5Cat	1.467	19.285	1.286	21.073	0.992	0.153	1.082	-0.799	0.012	0.066	0.106	0.024	0.090	0.132
6a-5Cat	1.249	23.434	1.387	19.440	0.993	0.162	0.983	1.989	0.017	0.047	0.087	0.020	0.062	0.136
7a-5Cat	0.805	28.668	0.812	28.696	0.996	0.123	0.987	0.107	0.014	0.058	0.108	0.014	0.054	0.105
8a-5Cat	1.037	23.166	0.914	27.062	0.984	0.374	1.031	-1.622	0.016	0.048	0.086	0.016	0.048	0.083
9a-5Cat	1.086	23.030	1.049	24.884	0.993	0.169	0.988	-0.743	0.017	0.049	0.087	0.013	0.043	0.073
10a-5Cat	1.453	18.181	1.150	23.251	0.985	0.274	1.097	-2.542	0.023	0.043	0.090	0.027	0.058	0.102
11a-5Cat	1.341	20.778	1.286	21.073	0.990	0.200	1.026	0.062	0.017	0.052	0.086	0.020	0.061	0.106
12a-5Cat	1.314	22.121	1.387	19.440	0.991	0.204	1.002	1.471	0.015	0.057	0.092	0.023	0.076	0.125
1b-5Cat	0.729	31.561	0.812	28.696	0.998	0.051	0.941	1.352	0.013	0.051	0.096	0.010	0.035	0.081
2b-5Cat	0.750	31.414	0.914	27.062	0.998	0.059	0.884	1.888	0.012	0.059	0.110	0.003	0.033	0.062
3b-5Cat	0.929	26.618	1.049	24.884	0.985	0.406	0.903	1.118	0.013	0.052	0.092	0.007	0.031	0.057
4b-5Cat	1.044	24.237	1.150	23.251	0.991	0.225	0.918	0.647	0.020	0.046	0.087	0.014	0.028	0.056
5b-5Cat	1.191	22.501	1.286	21.073	0.992	0.193	0.950	0.836	0.020	0.055	0.090	0.018	0.047	0.076
6b-5Cat	1.318	20.606	1.387	19.440	0.996	0.083	0.975	0.634	0.014	0.059	0.097	0.012	0.057	0.094

Table 29 (cont'd).

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
7b-5Cat	0.743	30.589	0.812	28.696	1.000	0.015	0.944	0.892	0.009	0.056	0.104	0.005	0.036	0.083
8b-5Cat	0.831	28.404	0.914	27.062	0.991	0.249	0.924	0.832	0.014	0.046	0.094	0.013	0.030	0.064
9b-5Cat	1.236	20.479	1.049	24.884	0.982	0.361	1.048	-1.940	0.018	0.051	0.082	0.019	0.052	0.082
10b-5Cat	1.148	23.147	1.150	23.251	0.994	0.140	0.989	0.091	0.018	0.051	0.093	0.018	0.048	0.092
11b-5Cat	1.330	22.644	1.286	21.073	0.994	0.135	1.067	0.951	0.017	0.053	0.102	0.029	0.102	0.153
12b-5Cat	1.599	20.655	1.387	19.440	0.974	0.547	1.158	1.304	0.012	0.053	0.090	0.057	0.144	0.227
1c-5Cat	0.683	33.053	0.812	28.696	0.999	0.026	0.902	1.911	0.015	0.055	0.096	0.007	0.028	0.070
2c-5Cat	0.752	31.360	0.914	27.062	0.998	0.074	0.884	1.879	0.017	0.051	0.099	0.004	0.032	0.057
3c-5Cat	1.141	21.317	1.049	24.884	0.974	0.548	0.979	-1.180	0.009	0.042	0.091	0.008	0.033	0.064
4c-5Cat	1.027	24.220	1.150	23.251	0.996	0.109	0.907	0.527	0.014	0.058	0.094	0.008	0.029	0.065
5c-5Cat	1.166	21.873	1.286	21.073	0.993	0.155	0.918	0.499	0.017	0.052	0.088	0.010	0.035	0.061
6c-5Cat	1.467	18.298	1.387	19.440	0.977	0.411	1.002	-0.132	0.012	0.044	0.088	0.015	0.045	0.090
7c-5Cat	0.750	30.038	0.812	28.696	0.998	0.057	0.944	0.675	0.013	0.063	0.100	0.009	0.040	0.079
8c-5Cat	0.764	30.772	0.914	27.062	0.996	0.117	0.889	1.682	0.019	0.053	0.094	0.008	0.031	0.057
9c-5Cat	1.080	22.366	1.049	24.884	0.983	0.371	0.959	-0.831	0.014	0.048	0.083	0.011	0.029	0.056
10c-5Cat	1.041	23.678	1.150	23.251	0.997	0.078	0.911	0.263	0.015	0.052	0.098	0.007	0.030	0.057
11c-5Cat	1.460	17.649	1.286	21.073	0.980	0.352	1.016	-1.371	0.018	0.049	0.087	0.017	0.044	0.076
12c-5Cat	1.191	22.294	1.387	19.440	0.990	0.235	0.911	1.433	0.015	0.048	0.094	0.012	0.034	0.075

Table 30. Comparison of Rescaled and Predicted G^{2*} Sampling Distributions for the Two Category Items in Test 2Cat12a

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-2Cat	0.238	5.283	0.228	4.735	0.992	0.044	1.095	0.329	0.014	0.057	0.094	0.025	0.077	0.127
2a-2Cat	0.234	6.982	0.207	7.418	0.978	0.155	1.072	-0.053	0.018	0.056	0.086	0.027	0.073	0.106
3a-2Cat	0.152	14.445	0.179	10.996	0.978	0.311	0.958	1.756	0.016	0.044	0.097	0.018	0.062	0.124
4a-2Cat	0.151	17.954	0.158	13.679	0.985	0.267	1.081	2.349	0.019	0.060	0.098	0.045	0.123	0.200
5a-2Cat	0.155	21.614	0.130	17.256	0.990	0.228	1.326	2.942	0.016	0.045	0.106	0.157	0.324	0.441
6a-2Cat	0.166	22.970	0.109	19.939	0.994	0.147	1.629	2.574	0.013	0.061	0.093	0.349	0.590	0.721
7a-2Cat	0.237	4.918	0.228	4.735	0.978	0.110	1.038	0.206	0.013	0.047	0.082	0.016	0.060	0.098
8a-2Cat	0.186	7.311	0.207	7.418	0.981	0.144	0.873	0.083	0.015	0.047	0.093	0.011	0.031	0.051
9a-2Cat	0.144	10.320	0.179	10.996	0.974	0.269	0.756	-0.034	0.020	0.055	0.089	0.005	0.017	0.033
10a-2Cat	0.144	10.256	0.158	13.679	0.962	0.391	0.753	-0.960	0.018	0.054	0.088	0.004	0.014	0.021
11a-2Cat	0.100	13.530	0.130	17.256	0.974	0.352	0.664	-0.995	0.023	0.052	0.095	0.000	0.006	0.013
12a-2Cat	0.078	16.291	0.109	19.939	0.983	0.284	0.632	-0.977	0.021	0.057	0.104	0.001	0.002	0.003

Table 31. Comparison of Rescaled and Predicted G^{2*} Sampling Distributions for the Two Category Items in Test 2Cat36

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-2Cat	0.515	5.903	0.499	6.143	0.997	0.021	1.008	-0.098	0.016	0.053	0.092	0.017	0.054	0.092
2a-2Cat	0.505	6.646	0.480	6.601	0.988	0.073	1.044	0.102	0.017	0.049	0.088	0.019	0.060	0.103
3a-2Cat	0.422	9.130	0.455	7.213	0.988	0.111	1.033	1.009	0.013	0.047	0.088	0.024	0.069	0.141
4a-2Cat	0.434	9.986	0.436	7.671	0.988	0.112	1.126	1.296	0.015	0.060	0.094	0.038	0.112	0.197
5a-2Cat	0.397	12.250	0.411	8.282	0.993	0.091	1.170	2.139	0.012	0.054	0.092	0.056	0.178	0.268
6a-2Cat	0.350	14.662	0.393	8.740	0.995	0.080	1.154	2.999	0.016	0.057	0.097	0.066	0.197	0.323
7a-2Cat	0.493	6.089	0.499	6.143	0.999	0.003	0.982	-0.021	0.014	0.053	0.096	0.011	0.049	0.093
8a-2Cat	0.489	6.530	0.480	6.601	0.999	0.009	1.012	-0.026	0.013	0.051	0.096	0.013	0.053	0.098
9a-2Cat	0.470	6.937	0.455	7.213	0.996	0.024	1.009	-0.110	0.011	0.040	0.114	0.012	0.040	0.114
10a-2Cat	0.511	6.594	0.436	7.671	0.995	0.037	1.078	-0.545	0.016	0.048	0.102	0.021	0.059	0.114
11a-2Cat	0.447	6.581	0.411	8.282	0.995	0.037	0.959	-0.799	0.015	0.056	0.089	0.012	0.037	0.068
12a-2Cat	0.440	6.079	0.393	8.740	0.987	0.080	0.917	-1.194	0.019	0.049	0.081	0.009	0.030	0.047
1b-2Cat	0.560	5.381	0.499	6.143	0.999	0.005	1.048	-0.401	0.008	0.052	0.102	0.013	0.056	0.103
2b-2Cat	0.540	5.411	0.480	6.601	0.993	0.039	1.008	-0.572	0.010	0.051	0.092	0.009	0.042	0.073
3b-2Cat	0.437	6.145	0.455	7.213	1.001	0.002	0.887	-0.488	0.013	0.051	0.103	0.002	0.026	0.051
4b-2Cat	0.401	6.091	0.436	7.671	1.000	0.004	0.816	-0.668	0.011	0.057	0.105	0.002	0.014	0.035
5b-2Cat	0.377	5.790	0.411	8.282	0.999	0.008	0.763	-1.010	0.014	0.044	0.087	0.002	0.010	0.024
6b-2Cat	0.356	5.573	0.393	8.740	1.000	0.001	0.721	-1.256	0.010	0.050	0.109	0.000	0.004	0.010

Table 31 (cont'd).

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
7b-2Cat	0.518	5.684	0.499	6.143	1.000	0.006	0.998	-0.227	0.009	0.061	0.101	0.008	0.057	0.094
8b-2Cat	0.552	5.344	0.480	6.601	0.997	0.019	1.028	-0.645	0.012	0.056	0.087	0.013	0.053	0.082
9b-2Cat	0.447	5.925	0.455	7.213	0.999	0.001	0.888	-0.589	0.012	0.050	0.102	0.002	0.025	0.049
10b-2Cat	0.440	5.765	0.436	7.671	1.000	-0.002	0.873	-0.880	0.009	0.049	0.109	0.003	0.017	0.040
11b-2Cat	0.341	6.243	0.411	8.282	1.000	0.006	0.719	-0.773	0.010	0.057	0.106	0.000	0.002	0.015
12b-2Cat	0.356	5.469	0.393	8.740	0.999	0.006	0.714	-1.277	0.011	0.059	0.097	0.002	0.003	0.012
1c-2Cat	0.482	6.236	0.499	6.143	0.999	0.003	0.972	0.051	0.010	0.059	0.110	0.006	0.051	0.097
2c-2Cat	0.453	6.688	0.480	6.601	0.998	0.015	0.948	0.057	0.016	0.052	0.109	0.008	0.046	0.086
3c-2Cat	0.476	7.102	0.455	7.213	0.995	0.041	1.031	-0.013	0.014	0.053	0.090	0.016	0.058	0.103
4c-2Cat	0.428	7.360	0.436	7.671	0.996	0.028	0.957	-0.119	0.017	0.052	0.103	0.012	0.041	0.084
5c-2Cat	0.425	6.948	0.411	8.282	0.992	0.058	0.936	-0.572	0.013	0.047	0.088	0.007	0.033	0.064
6c-2Cat	0.386	6.793	0.393	8.740	0.995	0.040	0.860	-0.832	0.014	0.049	0.097	0.003	0.021	0.039
7c-2Cat	0.569	5.299	0.499	6.143	0.986	0.072	1.043	-0.359	0.010	0.041	0.086	0.015	0.048	0.091
8c-2Cat	0.439	7.544	0.480	6.601	0.997	0.025	0.975	0.458	0.013	0.052	0.102	0.013	0.053	0.106
9c-2Cat	0.448	8.828	0.455	7.213	0.996	0.037	1.087	0.855	0.015	0.057	0.104	0.030	0.100	0.159
10c-2Cat	0.421	10.196	0.436	7.671	0.990	0.100	1.107	1.360	0.015	0.050	0.094	0.039	0.114	0.202
11c-2Cat	0.445	10.861	0.411	8.282	0.985	0.167	1.225	1.609	0.014	0.050	0.088	0.054	0.163	0.271
12c-2Cat	0.369	13.950	0.393	8.740	0.987	0.186	1.180	2.820	0.014	0.051	0.095	0.068	0.190	0.309

Table 32. Comparison of Rescaled and Predicted G^{2*} Sampling Distributions for the Five Category Items in Test 5Cat12a

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-5Cat	0.586	21.271	0.555	22.026	0.998	0.038	1.035	-0.353	0.014	0.059	0.103	0.019	0.067	0.110
2a-5Cat	0.584	22.136	0.556	23.102	0.998	0.037	1.026	-0.455	0.014	0.056	0.107	0.017	0.062	0.115
3a-5Cat	0.540	24.950	0.557	24.537	0.998	0.050	0.976	0.247	0.014	0.048	0.094	0.011	0.042	0.088
4a-5Cat	0.581	24.584	0.558	25.614	0.999	0.030	1.019	-0.495	0.014	0.051	0.098	0.015	0.054	0.103
5a-5Cat	0.570	26.087	0.559	27.049	0.993	0.183	0.993	-0.288	0.014	0.060	0.093	0.013	0.055	0.089
6a-5Cat	0.548	29.001	0.560	28.126	0.998	0.046	0.992	0.470	0.012	0.050	0.099	0.012	0.050	0.104
7a-5Cat	0.577	21.946	0.555	22.026	0.999	0.009	1.036	-0.032	0.008	0.048	0.085	0.014	0.065	0.115
8a-5Cat	0.530	24.090	0.556	23.102	0.999	0.023	0.972	0.495	0.014	0.053	0.096	0.011	0.045	0.088
9a-5Cat	0.520	25.443	0.557	24.537	0.999	0.021	0.950	0.442	0.011	0.049	0.108	0.007	0.035	0.087
10a-5Cat	0.573	24.356	0.558	25.614	0.998	0.047	0.998	-0.579	0.015	0.058	0.101	0.014	0.051	0.095
11a-5Cat	0.509	29.321	0.559	27.049	0.998	0.053	0.945	1.100	0.003	0.057	0.112	0.003	0.045	0.095
12a-5Cat	0.594	26.834	0.560	28.126	0.999	0.042	1.034	-0.632	0.013	0.060	0.109	0.019	0.067	0.120

Table 33. Comparison of Rescaled and Predicted G^{2*} Sampling Distributions for the Five Category Items in Test 5Cat36

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-5Cat	0.712	33.029	0.670	34.039	0.998	0.049	1.045	-0.477	0.012	0.061	0.112	0.021	0.081	0.130
2a-5Cat	0.734	32.286	0.681	34.402	0.999	0.014	1.043	-1.098	0.011	0.050	0.104	0.015	0.065	0.115
3a-5Cat	0.669	35.988	0.697	34.886	0.999	0.027	0.974	0.556	0.011	0.052	0.107	0.010	0.043	0.099
4a-5Cat	0.712	34.682	0.709	35.249	0.999	0.040	0.996	-0.244	0.009	0.052	0.105	0.007	0.048	0.102
5a-5Cat	0.769	33.939	0.724	35.733	0.999	0.028	1.033	-0.904	0.014	0.050	0.091	0.016	0.058	0.103
6a-5Cat	0.679	39.330	0.736	36.096	0.999	0.055	0.962	1.567	0.009	0.041	0.108	0.007	0.039	0.103
7a-5Cat	0.695	33.622	0.670	34.039	0.999	0.052	1.030	-0.162	0.013	0.053	0.104	0.016	0.068	0.123
8a-5Cat	0.699	34.091	0.681	34.402	0.998	0.086	1.019	-0.069	0.012	0.049	0.095	0.016	0.054	0.110
9a-5Cat	0.671	35.975	0.697	34.886	1.000	0.015	0.977	0.537	0.013	0.047	0.099	0.011	0.043	0.089
10a-5Cat	0.713	34.622	0.709	35.249	0.999	0.030	0.996	-0.284	0.014	0.044	0.110	0.013	0.042	0.094
11a-5Cat	0.756	34.203	0.724	35.733	0.999	0.046	1.020	-0.737	0.011	0.052	0.099	0.012	0.054	0.106
12a-5Cat	0.804	33.336	0.736	36.096	0.999	0.040	1.048	-1.424	0.012	0.056	0.093	0.017	0.062	0.105
1b-5Cat	0.697	33.927	0.670	34.039	0.999	0.038	1.038	-0.018	0.015	0.050	0.097	0.020	0.074	0.120
2b-5Cat	0.694	34.916	0.681	34.402	0.998	0.054	1.024	0.319	0.014	0.062	0.101	0.023	0.078	0.129
3b-5Cat	0.650	37.957	0.697	34.886	0.998	0.074	0.971	1.518	0.014	0.058	0.101	0.013	0.058	0.106
4b-5Cat	0.687	36.452	0.709	35.249	0.999	0.038	0.984	0.620	0.013	0.052	0.095	0.011	0.052	0.093
5b-5Cat	0.746	35.027	0.724	35.733	0.999	0.019	1.019	-0.338	0.010	0.054	0.100	0.011	0.060	0.106
6b-5Cat	0.855	31.048	0.736	36.096	0.998	0.062	1.074	-2.727	0.015	0.059	0.093	0.025	0.064	0.104

Table 33 (cont'd).

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
7b-5Cat	0.681	34.022	0.670	34.039	0.999	0.021	1.016	0.011	0.008	0.051	0.105	0.012	0.059	0.121
8b-5Cat	0.648	36.571	0.681	34.402	0.998	0.078	0.978	1.115	0.015	0.049	0.092	0.015	0.049	0.095
9b-5Cat	0.681	35.770	0.697	34.886	0.999	0.043	0.988	0.472	0.013	0.053	0.099	0.013	0.050	0.099
10b-5Cat	0.697	36.139	0.709	35.249	0.999	0.042	0.995	0.479	0.007	0.057	0.101	0.007	0.057	0.104
11b-5Cat	0.761	35.925	0.724	35.733	0.999	0.024	1.053	0.125	0.009	0.055	0.108	0.018	0.086	0.152
12b-5Cat	0.781	37.260	0.736	36.096	0.998	0.087	1.075	0.706	0.014	0.058	0.100	0.032	0.112	0.177
1c-5Cat	0.675	34.397	0.670	34.039	1.000	0.013	1.013	0.195	0.012	0.054	0.092	0.014	0.062	0.104
2c-5Cat	0.636	37.732	0.681	34.402	1.000	0.021	0.976	1.605	0.013	0.049	0.102	0.013	0.053	0.112
3c-5Cat	0.698	34.596	0.697	34.886	0.999	0.045	0.995	-0.099	0.009	0.044	0.109	0.009	0.041	0.103
4c-5Cat	0.674	36.178	0.709	35.249	0.999	0.033	0.962	0.477	0.007	0.049	0.105	0.003	0.038	0.079
5c-5Cat	0.640	38.011	0.724	35.733	1.000	0.013	0.910	1.034	0.008	0.048	0.111	0.002	0.023	0.053
6c-5Cat	0.700	35.721	0.736	36.096	1.000	0.012	0.945	-0.168	0.010	0.054	0.104	0.005	0.030	0.065
7c-5Cat	0.727	31.838	0.670	34.039	0.999	0.016	1.049	-1.152	0.011	0.055	0.102	0.019	0.073	0.116
8c-5Cat	0.640	37.458	0.681	34.402	0.999	0.046	0.979	1.502	0.014	0.052	0.101	0.014	0.057	0.107
9c-5Cat	0.683	35.066	0.697	34.886	0.999	0.029	0.981	0.116	0.010	0.058	0.105	0.007	0.048	0.097
10c-5Cat	0.653	36.704	0.709	35.249	1.000	0.017	0.940	0.688	0.009	0.049	0.109	0.006	0.031	0.067
11c-5Cat	0.724	33.599	0.724	35.733	1.000	0.009	0.969	-1.036	0.012	0.055	0.096	0.007	0.036	0.065
12c-5Cat	0.693	36.192	0.736	36.096	0.999	0.055	0.941	0.094	0.009	0.041	0.108	0.005	0.027	0.056

Table 26 presents results comparing the rescaled and predicted sampling distributions for the χ^{2*} statistics for Test 2Cat12a. For many items, the scaling factors obtained using the method of moments (found under the heading 'Rescaled') were similar to the scaling factors obtained using the prediction equations. The degrees of freedom values obtained using the prediction equations followed the same general trend as the values obtained by the method of moments (increasing as *a* increased), but some differences in the values obtained using the two methods were apparent. For some items, the predicted degrees of freedom were higher than the degrees of freedom found using the method of moments, while the direction of the differences was reversed for other items.

Under the heading 'Q-Q Plot Regression Coefficients', Table 26 shows variation in the slopes associated with the predicted scaling corrections. The slopes associated with items with high *a* values differed more from 1 than the slopes of items with low *a* values. Further, the intercepts also differed from 0. This indicated that the distributions of Pearson statistics rescaled by the predicted scaling factors did not follow the theoretical chi-square distributions with predicted degrees of freedom for the two category items on Test 2cat12a.

The Type I error rates associated with the predicted scaling corrections further indicated that the distributions for most items did not match. At $\alpha = .05$, only two of the items on Test 2Cat12a had Type I error rates that fell in the range of .037 to .063, which was expected with a 95% Confidence Interval. Further, some of the items had increased Type I error rates, while the error rates for other items were lower than expected. A pattern in the Type I error rates across the items was not evident.

In considering the three 2-category, 12 item tests, Tests 2Cat12a, 2Cat12b, and 2Cat12c, (tables available from the author), it was found that the slopes, intercepts, and Type I error rates

associated with the prediction equations were closer to the expected values for items having less extreme *b* values. That is, when applied to the 2-category items with *b* values closer to 0, namely the items on Test 2Cat12b that had threshold values of –0.5 and 0.5, the prediction equations resulted in most items having Type I error rates at their expected levels. This was not the case, however, for most of the 2-category items on Tests 2Cat12a and 2Cat12c, which had *b* values further from 0.

Table 27 provides results comparing the rescaled and predicted sampling distributions for the χ^{2*} statistics for Test 2Cat36. For this longer test, the predicted scaling factors and predicted degrees of freedom were less variable than the values obtained using the method of moments. Therefore, between-item differences in the scaling factors and degrees of freedom values that were observed using the method of moments were not as pronounced using the prediction equations. Differences were observed in the scaling corrections obtained using the two methods.

Generally, the slopes were closer to 1 than for Test 2Cat12a. However, the slopes and intercepts still showed differences from 1 and 0, respectively. At the $\alpha = .05$ level of significance, only 8 of the 36 items exhibited nominal Type I error rates. This again provided evidence that the predicted scaling corrections did not rescale the χ^{2*} distributions adequately for the two category items, even for the longer test length of 36 items.

Table 28 presents results comparing the rescaled and predicted sampling distributions for the χ^{2*} statistics for the 5-category items on Test 5Cat12a. Table 28 shows that the predicted scaling factors and degrees of freedom values were similar to the values obtained using the method of moments. Further, the slopes and intercepts associated with the predicted scaling corrections were close to 1 and 0, respectively. At the $\alpha = .05$ level of significance, the Type I error rates for most of the 5-category items on Test 5Cat12a fell in or close to the expected

ranges. The results for the items on Test 5Cat12c were similar, while the results for the items on Test 5Cat12b were less consistent.

Table 29 presents results comparing the rescaled and predicted sampling distributions for the χ^{2*} statistics for the 5-category items on Test 5Cat36. For most of the items, the predicted scaling factors were similar to the values obtained using the method of moments. However, for several items, there were differences in the degrees of freedom values obtained using the method of moments and the prediction equations. For many of the items, the slopes and intercepts associated with the predicted scaling corrections were close to 1 and 0, respectively. At the α = .05 level of significance, the Type I error rates for approximately half of the 5-category items fell in the expected range. The Type I error rates for most of the other items were lower than expected.

The results in Tables 28 and 29 indicated that, for slightly more than half of the 5-category items, the sampling distributions of χ^{2*} statistics rescaled by the predicted scaling factors closely matched the theoretical chi-square distributions with predicted degrees of freedom. However, differences from 1 and 0 for the slopes and intercepts, respectively, and Type I error rates outside of the expected range, were found for several of the 5-category items.

In comparing results associated with the χ^{2*} statistics across the number of score category levels, the slopes of the regression lines fitted to Q-Q plots of predicted chi-square distributions were closer to 1 as the number of score categories increased. In addition, as the number of score categories increased, the Type I error rates for more of the items fell in the expected range. Thus, the match between the sampling distributions of χ^{2*} statistics rescaled by the predicted scaling factors and the theoretical chi-square distributions with predicted degrees of freedom was closer for items with more score response category levels. Still, the prediction equations

adequately rescaled the distributions of χ^{2*} statistics for only approximately half of the items, even on the 5 category tests.

Tables 30 to 33 present results for the likelihood ratio G^{2*} statistics for Tests 2Cat12a, 2Cat36, 5Cat12a, and 5Cat36, respectively. The predicted distributions based on the likelihood ratio form of the statistic more closely followed theoretical chi-square distributions than the predicted distributions of Pearson statistics.

The results in Table 30 show differences in the predicted degrees of freedom and the degrees of freedom values found using the method of moments for the items on Test 2Cat12a. Further, differences in the slopes and intercepts from 1 and 0, respectively, were apparent. The results in Table 30 indicate that the predicted scaling corrections did not adequately rescale the G^{2*} distributions for the 2-category items on Test 2Cat12a. The results for the other 2 category tests (tables available from the author) were similar to the results found using the Pearson statistics. For example, the items on Test 2Cat12c were not adequately rescaled by the prediction equations, but the Type I error rates for most of the items on Test 2Cat12b, which had threshold values of –0.5 and 0.5, were at their expected levels.

As seen in Table 31, which presents results associated with the G^{2*} statistics for the 2-category items on Test 2Cat36, the slopes and intercepts were close to 1 and 0, respectively. The Type I error rates fell at the nominal levels for approximately half of the items. These results were an improvement over the results found with the Pearson statistic, where less than 25% of the 2-category items had Type I error rates at the nominal level.

For the likelihood ratio statistics, the match between the G^{2*} distributions rescaled by the predicted scaling factor and the theoretical chi-square distributions with predicted degrees of freedom improved as the number of score category levels increased. For the 5-category tests, the

match between the predicted rescaled G^{2*} distributions and the theoretical chi-square distributions with predicted degrees of freedom was evident. This can be seen in Tables 32 and 33 by the similarities of rescaled and predicted scaling corrections, the closeness of the predicted slopes and intercepts to 1 and 0, respectively, and the Type I error rates that fell within the expected ranges.

Overall, considering both the Pearson and likelihood ratio statistics, Tables 26 to 33 showed that the scaling factors were fairly well estimated across tests by the multilevel prediction equations. However, especially for tests at the lower score response category levels, the predicted degrees of freedom values differed from the degrees of freedom values obtained using the method of moments both forms of the statistics. This could be due to the variability found in the degrees of freedom values that were obtained using the method of moments.

For both the Pearson and likelihood ratio forms of the statistics, the 2- and 3-category items were not adequately rescaled by the predicted scaling corrections. However, for the 4- and 5-category tests, the distribution of likelihood ratio G^{2*} statistics were well estimated using the prediction equations. For the Pearson statistics, even for the 5-category items, the prediction equations adequately rescaled only slightly more than half of the items. Thus, when rescaled by the predicted scaling factors, the likelihood ratio G^{2*} distributions for the 4- and 5-category items approximated the theoretical chi-square distributions with predicted degrees of freedom more closely than the Pearson χ^{2*} distributions.

IV. B. 3. Prediction Equations for Item Subsets

The results discussed in Section IV. B. 2. indicated that the multilevel prediction equations did not adequately predict the sampling distributions of χ^{2*} and G^{2*} statistics for all items. While the results in Section IV. B. 2. suggested that the prediction equations worked

fairly well for items on tests with higher numbers of score response categories for the G^{2*} statistics, equations that would adequately rescale the G^{2*} statistics at the lower score category levels, and the Pearson χ^{2*} statistics at all score category levels, were desired. As a result, additional regression analyses were conducted as an attempt to adequately predict the scaling corrections for items having fewer numbers of score response categories.

The multilevel prediction equations utilized data from all of the items in the simulation study as an attempt to estimate a single set of prediction equations that would work across all items. Since the multilevel prediction equations did not work adequately for all items, an attempt to estimate the scaling corrections within subsets of items, rather than across all items, was made.

Ankenmann (1994) found that item level, rather than test level, prediction equations were needed to find the appropriate scaling corrections for the 5-category items in his study. It was not possible to conduct item level regression analyses in the current study, because each item appeared at most 3 times in the simulations. For example, items on tests 12a and 12b at each score category level were also found on the 24 and 36 items tests at that score category level. These items appeared 3 times in the study. Items on tests 12c at each score category level appeared twice, once on the 12c test, and again on the 36 item tests. In order to predict the scaling corrections at the item level, more than two or three data points were needed. As a result, item level equations could not be predicted.

However, it was possible to estimate several sets of prediction equations for subsets of similar items from the overall data set. That is, it was possible to group items together based on similar characteristics (i.e., *a* value), and perform regression analyses to predict the scaling corrections within each subset of items. This resulted in a number of prediction equations that

were appropriate for items having specific characteristics, rather than a single set of prediction equations for all items, as was obtained using the multilevel prediction models. Prediction in this manner could have resulted in equations that were more applicable for χ^{2*}, and for the 2- and 3-category tests for G^{2*}.

Regression models were fitted to several subsets of the data found in Appendix C. The final regression subset-level equations for predicting the scaling corrections were analyzed for all 64 items with the same *a* value. The final subset-level regression equation for predicting the degrees of freedom values were analyzed for all subsets of items with the same *a* and number of item score response categories. The data subsets for predicting the degrees of freedom values consisted of 16 sets of independent variables. The final models for the subset analysis were selected because they accounted for more of the variation in the dependent variables than other estimated models.

After the regression models were estimated for the subsets of data, the empirical sampling distributions were rescaled by the new predicted scaling factors, and compared to the theoretical chi-square distributions with new predicted degrees of freedom. The similarities between the distributions were assessed using Q-Q plots and Type I error rates. Specifically, the linearity of the Q-Q plots was examined, the slopes and intercepts of regression lines fitted to Q-Q plots of the rescaled empirical distributions and theoretical chi-square distributions with predicted degrees of freedom were examined for their closeness to 1 and 0, respectively, and Type I error rates were examined.

Tables 34 to 41 present results from the analysis of Q-Q plots obtained using the second set of predicted scaling corrections, those estimated within data subsets, for the items in Tests 2Cat12A, 2Cat36, 5Cat12A, and 5Cat36, respectively. Tables 34 to 41 present the scaling

corrections, slopes and intercepts of regression lines fitted to Q-Q plots, and Type I error rates associated with both the empirically generated and predicted distributions. The presentation of results in this format allows for the comparison of the results obtained using the subset level prediction equations to the results obtained using the method of moments. Type I error rates that fell in the ranges expected with 95% confidence intervals are underlined.

Tables 34 to 37 present the results for the Pearson χ^{2*} statistics for Tests 2Cat12A, 2Cat36, 5Cat12A, and 5Cat36, respectively. Tables 38 to 41 present results for the likelihood ratio G^{2*} statistics for the same four tests.

Tables for only this subset of tests are presented below to simplify the discussion, and at the same time allow for comparisons between the Pearson and likelihood ratio forms of the statistics across number of item score category levels and test lengths. Tables containing similar information for all tests are available from the author.

Tables 34 to 37 show that for the Pearson statistic, for 2- and 5-category tests consisting of 12 and 36 items, the slopes and intercepts of regression lines fitted to the Q-Q plots of the rescaled empirical versus theoretical chi-square distributions with predicted degrees of freedom were different from 1 and 0, respectively. Further, the Type I error rates were generally above the expected levels. This indicated that the prediction equations based on the item subsets did not adequately rescale the Pearson distributions for 2- or 5-category items. Similar results were observed for the 3- and 4-category tests.

Tables 38 and 39 indicate for the likelihood ratio distributions that the predicted chi-square distributions did not approximate theoretical chi-square distributions for the 2-category or 5-category tests. Again, the slopes and intercepts of regression lines fitted to Q-Q plots differed

from 1 and 0, respectively, and Type I error rates for many items were outside of their expected ranges.

Overall, Tables 36 to 41 indicate that the prediction equations analyzed for item subsets provided scaling corrections that did not adequately rescale the empirical distributions either χ^{2*} and G^{2*} statistics so that they follow theoretical chi-square distributions.

Table 34. Comparison of Rescaled and Predicted χ^{2*} Sampling Distributions for the Two Category Items in Test 2Cat12a (Subsets)

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-2Cat	0.253	4.892	0.215	1.991	0.984	0.079	1.836	2.099	0.014	0.054	0.086	0.124	0.359	0.541
2a-2Cat	0.290	5.666	0.229	2.708	0.965	0.194	1.802	2.279	0.016	0.054	0.078	0.131	0.337	0.539
3a-2Cat	0.278	8.163	0.184	5.323	0.928	0.593	1.764	2.972	0.015	0.038	0.065	0.178	0.434	0.651
4a-2Cat	0.266	10.585	0.211	6.395	0.948	0.550	1.556	3.373	0.021	0.046	0.077	0.152	0.390	0.595
5a-2Cat	0.238	14.537	0.139	9.644	0.963	0.539	2.045	5.237	0.016	0.041	0.080	0.509	0.808	0.925
6a-2Cat	0.222	17.716	0.111	10.617	0.989	0.185	2.568	8.146	0.017	0.061	0.091	0.861	0.972	0.992
7a-2Cat	0.243	4.696	0.215	1.991	0.970	0.139	1.720	1.886	0.014	0.045	0.082	0.095	0.305	0.467
8a-2Cat	0.211	6.317	0.229	2.708	0.967	0.206	1.385	2.043	0.019	0.045	0.078	0.059	0.210	0.368
9a-2Cat	0.191	7.690	0.184	5.323	0.948	0.404	1.196	1.613	0.017	0.051	0.077	0.051	0.128	0.210
10a-2Cat	0.235	6.260	0.211	6.395	0.912	0.555	1.003	0.554	0.017	0.038	0.067	0.022	0.054	0.083
11a-2Cat	0.161	8.375	0.139	9.644	0.921	0.657	0.990	0.159	0.016	0.044	0.065	0.018	0.049	0.077
12a-2Cat	0.128	9.692	0.111	10.617	0.905	0.923	0.999	0.623	0.012	0.034	0.064	0.017	0.044	0.092

Table 35. Comparison of Rescaled and Predicted χ^{2*} Sampling Distributions for the Two Category Items in Test 2Cat36 (Subsets)

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-2Cat	0.578	5.160	0.476	6.777	0.990	0.052	1.044	-0.809	0.013	0.057	0.086	0.015	0.057	0.084
2a-2Cat	0.654	5.127	0.483	6.462	0.980	0.100	1.176	-0.653	0.018	0.048	0.090	0.028	0.080	0.113
3a-2Cat	0.870	4.616	0.554	5.676	0.935	0.299	1.314	-0.212	0.014	0.034	0.059	0.036	0.089	0.154
4a-2Cat	0.904	5.099	0.547	5.647	0.956	0.222	1.497	-0.028	0.016	0.034	0.078	0.066	0.156	0.245
5a-2Cat	0.659	7.718	0.617	5.324	0.984	0.125	1.274	1.466	0.015	0.053	0.088	0.059	0.176	0.268
6a-2Cat	0.624	8.546	0.589	5.999	0.974	0.221	1.239	1.610	0.017	0.049	0.083	0.059	0.149	0.254
7a-2Cat	0.493	5.931	0.476	6.777	0.999	0.007	0.967	-0.411	0.014	0.047	0.103	0.012	0.038	0.076
8a-2Cat	0.507	6.112	0.483	6.462	0.997	0.021	1.017	-0.154	0.013	0.047	0.093	0.015	0.049	0.094
9a-2Cat	0.532	5.920	0.554	5.676	0.994	0.038	0.975	0.149	0.014	0.046	0.100	0.012	0.045	0.096
10a-2Cat	0.726	4.629	0.547	5.647	0.987	0.061	1.181	-0.524	0.019	0.056	0.085	0.031	0.079	0.117
11a-2Cat	0.898	3.330	0.617	5.324	0.960	0.134	1.085	-0.926	0.020	0.039	0.065	0.026	0.044	0.067
12a-2Cat	1.086	2.513	0.589	5.999	0.920	0.203	1.049	-1.662	0.011	0.033	0.057	0.015	0.033	0.048
1b-2Cat	0.522	5.568	0.476	6.777	1.000	0.002	0.992	-0.619	0.011	0.056	0.110	0.009	0.044	0.089
2b-2Cat	0.509	5.461	0.483	6.462	0.993	0.040	0.960	-0.446	0.010	0.049	0.091	0.009	0.027	0.070
3b-2Cat	0.414	6.053	0.554	5.676	1.000	0.002	0.771	0.141	0.012	0.049	0.102	0.001	0.018	0.040
4b-2Cat	0.408	5.552	0.547	5.647	0.992	0.049	0.734	0.001	0.010	0.047	0.094	0.002	0.009	0.026
5b-2Cat	0.460	4.485	0.617	5.324	0.979	0.089	0.668	-0.212	0.010	0.044	0.078	0.002	0.006	0.018
6b-2Cat	0.527	3.660	0.589	5.999	0.985	0.054	0.681	-0.808	0.014	0.039	0.074	0.004	0.010	0.020

Table 35 (cont'd).

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
7b-2Cat	0.484	5.854	0.476	6.777	1.000	0.005	0.945	-0.447	0.010	0.061	0.100	0.005	0.043	0.076
8b-2Cat	0.512	5.445	0.483	6.462	0.994	0.028	0.966	-0.467	0.010	0.051	0.095	0.009	0.042	0.067
9b-2Cat	0.422	5.855	0.554	5.676	1.000	0.006	0.773	0.069	0.011	0.052	0.097	0.002	0.015	0.035
10b-2Cat	0.485	4.941	0.547	5.647	0.994	0.032	0.823	-0.266	0.008	0.046	0.091	0.006	0.019	0.044
11b-2Cat	0.407	4.928	0.617	5.324	0.993	0.036	0.629	-0.098	0.012	0.045	0.087	0.002	0.007	0.015
12b-2Cat	0.519	3.618	0.589	5.999	0.979	0.082	0.660	-0.770	0.017	0.040	0.075	0.004	0.007	0.018
1c-2Cat	0.471	6.187	0.476	6.777	0.999	0.002	0.945	-0.279	0.010	0.054	0.112	0.007	0.036	0.076
2c-2Cat	0.471	6.201	0.483	6.462	0.998	0.017	0.953	-0.106	0.010	0.054	0.108	0.008	0.045	0.085
3c-2Cat	0.626	5.280	0.554	5.676	0.971	0.156	1.056	-0.029	0.013	0.045	0.078	0.016	0.057	0.090
4c-2Cat	0.589	5.266	0.547	5.647	0.988	0.059	1.026	-0.127	0.013	0.049	0.087	0.017	0.053	0.089
5c-2Cat	1.058	2.882	0.617	5.324	0.935	0.192	1.145	-1.151	0.015	0.033	0.057	0.022	0.048	0.068
6c-2Cat	0.857	3.098	0.589	5.999	0.946	0.164	0.963	-1.272	0.017	0.028	0.063	0.017	0.025	0.044
7c-2Cat	0.631	4.704	0.476	6.777	0.983	0.084	1.078	-1.066	0.009	0.046	0.086	0.013	0.048	0.086
8c-2Cat	0.529	6.218	0.483	6.462	0.995	0.031	1.069	-0.092	0.013	0.055	0.099	0.020	0.062	0.126
9c-2Cat	0.986	4.256	0.554	5.676	0.947	0.221	1.445	-0.638	0.013	0.043	0.080	0.059	0.123	0.187
10c-2Cat	0.873	5.279	0.547	5.647	0.960	0.211	1.479	0.077	0.017	0.047	0.086	0.076	0.154	0.236
11c-2Cat	1.095	4.691	0.617	5.324	0.916	0.398	1.516	0.250	0.013	0.038	0.065	0.063	0.157	0.250
12c-2Cat	0.693	7.713	0.589	5.999	0.949	0.397	1.273	1.431	0.014	0.045	0.082	0.058	0.149	0.237

Table 36. Comparison of Rescaled and Predicted χ^{2*} Sampling Distributions for the Five Category Items in Test 5Cat12a (Subsets)

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-5Cat	0.700	17.763	0.614	20.416	0.993	0.134	1.054	-1.274	0.018	0.052	0.094	0.021	0.056	0.102
2a-5Cat	0.767	17.068	0.694	18.649	0.993	0.118	1.049	-0.702	0.015	0.057	0.099	0.023	0.063	0.109
3a-5Cat	0.787	17.762	0.806	16.778	0.992	0.152	0.996	0.622	0.020	0.048	0.089	0.021	0.053	0.099
4a-5Cat	0.872	17.326	0.789	17.813	0.994	0.103	1.084	-0.143	0.017	0.057	0.092	0.029	0.079	0.138
5a-5Cat	0.818	19.627	0.730	20.936	0.989	0.220	1.072	-0.454	0.015	0.051	0.097	0.024	0.079	0.127
6a-5Cat	0.797	22.056	0.747	22.551	0.998	0.051	1.053	-0.203	0.014	0.059	0.109	0.022	0.083	0.140
7a-5Cat	0.656	19.251	0.614	20.416	0.999	0.024	1.037	-0.587	0.016	0.046	0.091	0.018	0.059	0.100
8a-5Cat	0.874	14.791	0.694	18.649	0.950	0.736	1.062	-1.184	0.013	0.039	0.067	0.018	0.048	0.088
9a-5Cat	0.725	18.786	0.806	16.778	0.991	0.176	0.943	1.065	0.014	0.044	0.094	0.012	0.039	0.093
10a-5Cat	0.884	16.676	0.789	17.813	0.990	0.161	1.074	-0.424	0.017	0.049	0.088	0.025	0.070	0.125
11a-5Cat	0.751	21.460	0.730	20.936	0.996	0.077	1.038	0.345	0.011	0.053	0.101	0.016	0.071	0.139
12a-5Cat	0.853	20.699	0.747	22.551	0.998	0.032	1.093	-0.989	0.009	0.058	0.108	0.024	0.095	0.142

Table 37. Comparison of Rescaled and Predicted χ^{2*} Sampling Distributions for the Five Category Items in Test 5Cat36 (Subsets)

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-5Cat	0.808	28.671	0.759	28.392	0.997	0.093	1.066	0.247	0.014	0.054	0.105	0.026	0.094	0.150
2a-5Cat	0.931	25.419	0.851	26.167	0.996	0.097	1.074	-0.295	0.015	0.050	0.101	0.027	0.085	0.143
3a-5Cat	1.157	21.565	1.106	21.469	0.981	0.413	1.029	0.478	0.015	0.051	0.090	0.019	0.065	0.126
4a-5Cat	1.536	17.207	1.223	20.710	0.965	0.600	1.102	-1.207	0.015	0.039	0.071	0.027	0.067	0.112
5a-5Cat	1.467	19.285	1.295	20.630	0.992	0.153	1.086	-0.559	0.012	0.066	0.106	0.025	0.096	0.138
6a-5Cat	1.249	23.434	1.333	21.198	0.993	0.162	0.979	1.202	0.017	0.047	0.087	0.018	0.051	0.106
7a-5Cat	0.805	28.668	0.759	28.392	0.996	0.123	1.062	0.277	0.014	0.058	0.108	0.024	0.097	0.154
8a-5Cat	1.037	23.166	0.851	26.167	0.984	0.374	1.126	-1.254	0.016	0.048	0.086	0.030	0.088	0.149
9a-5Cat	1.086	23.030	1.106	21.469	0.993	0.169	1.010	0.930	0.017	0.049	0.087	0.023	0.069	0.115
10a-5Cat	1.453	18.181	1.223	20.710	0.985	0.274	1.095	-1.072	0.023	0.043	0.090	0.027	0.077	0.122
11a-5Cat	1.341	20.778	1.295	20.630	0.990	0.200	1.029	0.283	0.017	0.052	0.086	0.021	0.064	0.113
12a-5Cat	1.314	22.121	1.333	21.198	0.991	0.204	0.999	0.648	0.015	0.057	0.092	0.018	0.068	0.106
1b-5Cat	0.729	31.561	0.759	28.392	0.998	0.051	1.012	1.602	0.013	0.051	0.096	0.021	0.077	0.140
2b-5Cat	0.750	31.414	0.851	26.167	0.998	0.059	0.964	2.450	0.012	0.059	0.110	0.016	0.072	0.126
3b-5Cat	0.929	26.618	1.106	21.469	0.985	0.406	0.923	2.550	0.013	0.052	0.092	0.012	0.052	0.099
4b-5Cat	1.044	24.237	1.223	20.710	0.991	0.225	0.916	1.721	0.020	0.046	0.087	0.015	0.034	0.078
5b-5Cat	1.191	22.501	1.295	20.630	0.992	0.193	0.953	1.035	0.020	0.055	0.090	0.020	0.050	0.085
6b-5Cat	1.318	20.606	1.333	21.198	0.996	0.083	0.971	-0.203	0.014	0.059	0.097	0.010	0.041	0.081

Table 37 (cont'd).

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
7b-5Cat	0.743	30.589	0.759	28.392	1.000	0.015	1.016	1.111	0.009	0.056	0.104	0.015	0.077	0.138
8b-5Cat	0.831	28.404	0.851	26.167	0.991	0.249	1.009	1.333	0.014	0.046	0.094	0.020	0.067	0.122
9b-5Cat	1.236	20.479	1.106	21.469	0.982	0.361	1.072	-0.121	0.018	0.051	0.082	0.032	0.074	0.119
10b-5Cat	1.148	23.147	1.223	20.710	0.994	0.140	0.987	1.285	0.018	0.051	0.093	0.022	0.058	0.115
11b-5Cat	1.330	22.644	1.295	20.630	0.994	0.135	1.070	1.175	0.017	0.053	0.102	0.030	0.108	0.162
12b-5Cat	1.599	20.655	1.333	21.198	0.974	0.547	1.153	0.344	0.012	0.053	0.090	0.048	0.123	0.199
1c-5Cat	0.683	33.053	0.759	28.392	0.999	0.026	0.970	2.194	0.015	0.055	0.096	0.016	0.062	0.114
2c-5Cat	0.752	31.360	0.851	26.167	0.998	0.074	0.965	2.441	0.017	0.051	0.099	0.019	0.061	0.126
3c-5Cat	1.141	21.317	1.106	21.469	0.974	0.548	1.002	0.492	0.009	0.042	0.091	0.011	0.052	0.102
4c-5Cat	1.027	24.220	1.223	20.710	0.996	0.109	0.904	1.598	0.014	0.058	0.094	0.009	0.037	0.076
5c-5Cat	1.166	21.873	1.295	20.630	0.993	0.155	0.921	0.694	0.017	0.052	0.088	0.010	0.038	0.065
6c-5Cat	1.467	18.298	1.333	21.198	0.977	0.411	0.998	-1.014	0.012	0.044	0.088	0.012	0.036	0.076
7c-5Cat	0.750	30.038	0.759	28.392	0.998	0.057	1.015	0.878	0.013	0.063	0.100	0.016	0.075	0.136
8c-5Cat	0.764	30.772	0.851	26.167	0.996	0.117	0.971	2.230	0.019	0.053	0.094	0.021	0.061	0.117
9c-5Cat	1.080	22.366	1.106	21.469	0.983	0.371	0.981	0.791	0.014	0.048	0.083	0.014	0.048	0.089
10c-5Cat	1.041	23.678	1.223	20.710	0.997	0.078	0.908	1.355	0.015	0.052	0.098	0.010	0.040	0.081
11c-5Cat	1.460	17.649	1.295	20.630	0.980	0.352	1.020	-1.144	0.018	0.049	0.087	0.018	0.047	0.080
12c-5Cat	1.191	22.294	1.333	21.198	0.989	0.235	0.907	0.690	0.015	0.048	0.094	0.010	0.028	0.057

Table 38. Comparison of Rescaled and Predicted G^{2*} Sampling Distributions for the Two Category Items in Test 2Cat12a (Subsets)

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-2Cat	0.238	5.283	0.207	2.716	0.992	0.044	1.598	1.716	0.014	0.057	0.094	0.098	0.268	0.433
2a-2Cat	0.234	6.982	0.195	4.096	0.978	0.155	1.553	2.034	0.018	0.056	0.086	0.102	0.270	0.429
3a-2Cat	0.152	14.445	0.132	8.368	0.978	0.311	1.499	4.169	0.016	0.044	0.097	0.203	0.454	0.633
4a-2Cat	0.151	17.954	0.132	10.283	0.985	0.267	1.501	5.128	0.019	0.060	0.098	0.240	0.544	0.715
5a-2Cat	0.155	21.614	0.106	13.698	0.990	0.228	1.827	6.610	0.016	0.045	0.106	0.547	0.821	0.924
6a-2Cat	0.166	22.970	0.104	14.935	0.994	0.147	1.982	7.246	0.013	0.061	0.093	0.704	0.902	0.955
7a-2Cat	0.237	4.918	0.207	2.716	0.978	0.110	1.524	1.480	0.013	0.047	0.082	0.078	0.219	0.357
8a-2Cat	0.186	7.311	0.195	4.096	0.981	0.144	1.263	1.795	0.015	0.047	0.093	0.050	0.161	0.288
9a-2Cat	0.144	10.320	0.132	8.368	0.974	0.269	1.185	1.337	0.020	0.055	0.089	0.055	0.134	0.210
10a-2Cat	0.144	10.256	0.132	10.283	0.962	0.391	1.050	0.418	0.018	0.054	0.088	0.025	0.074	0.119
11a-2Cat	0.100	13.530	0.106	13.698	0.974	0.352	0.917	0.258	0.023	0.052	0.095	0.017	0.039	0.073
12a-2Cat	0.078	16.291	0.104	14.935	0.983	0.284	0.771	0.707	0.021	0.057	0.104	0.003	0.019	0.028

Table 39. Comparison of Rescaled and Predicted G^{2*} Sampling Distributions for the Two Category Items in Test 2Cat36 (Subsets)

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-2Cat	0.515	5.903	0.477	7.296	0.997	0.021	0.966	-0.675	0.016	0.053	0.092	0.013	0.038	0.079
2a-2Cat	0.505	6.646	0.460	7.694	0.988	0.073	1.007	-0.445	0.017	0.049	0.088	0.017	0.049	0.082
3a-2Cat	0.422	9.130	0.427	8.179	0.988	0.111	1.032	0.573	0.013	0.047	0.088	0.020	0.064	0.124
4a-2Cat	0.434	9.986	0.399	8.742	0.988	0.112	1.153	0.798	0.015	0.060	0.094	0.040	0.109	0.191
5a-2Cat	0.397	12.250	0.363	9.239	0.993	0.091	1.253	1.815	0.012	0.054	0.092	0.070	0.205	0.307
6a-2Cat	0.350	14.662	0.337	9.925	0.995	0.080	1.259	2.727	0.016	0.057	0.097	0.093	0.243	0.367
7a-2Cat	0.493	6.089	0.477	7.296	0.999	0.003	0.942	-0.585	0.014	0.053	0.096	0.007	0.034	0.068
8a-2Cat	0.489	6.530	0.460	7.694	0.999	0.009	0.978	-0.571	0.013	0.051	0.096	0.010	0.039	0.075
9a-2Cat	0.470	6.937	0.427	8.179	0.996	0.024	1.008	-0.611	0.011	0.040	0.114	0.011	0.033	0.099
10a-2Cat	0.511	6.594	0.399	8.742	0.995	0.037	1.103	-1.189	0.016	0.048	0.102	0.021	0.057	0.109
11a-2Cat	0.447	6.581	0.363	9.239	0.995	0.037	1.027	-1.397	0.015	0.056	0.089	0.015	0.043	0.074
12a-2Cat	0.440	6.079	0.337	9.925	0.987	0.080	0.999	-1.979	0.019	0.049	0.081	0.017	0.034	0.056
1b-2Cat	0.560	5.381	0.477	7.296	0.999	0.005	1.005	-1.017	0.008	0.052	0.102	0.008	0.039	0.080
2b-2Cat	0.540	5.411	0.460	7.694	0.993	0.039	0.973	-1.134	0.010	0.051	0.092	0.009	0.025	0.057
3b-2Cat	0.437	6.145	0.427	8.179	1.001	0.002	0.887	-0.960	0.013	0.051	0.103	0.002	0.021	0.043
4b-2Cat	0.401	6.091	0.399	8.742	1.000	0.004	0.836	-1.185	0.011	0.057	0.105	0.002	0.014	0.031
5b-2Cat	0.377	5.790	0.363	9.239	0.999	0.008	0.816	-1.538	0.014	0.044	0.087	0.003	0.013	0.025
6b-2Cat	0.356	5.573	0.337	9.925	1.000	0.001	0.787	-1.937	0.010	0.050	0.109	0.000	0.006	0.016

Table 39 (cont'd).

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
7b-2Cat	0.518	5.684	0.477	7.296	1.000	0.006	0.957	-0.811	0.009	0.061	0.101	0.005	0.039	0.078
8b-2Cat	0.552	5.344	0.460	7.694	0.997	0.019	0.993	-1.222	0.012	0.056	0.087	0.009	0.041	0.065
9b-2Cat	0.447	5.925	0.427	8.179	0.999	0.001	0.888	-1.065	0.012	0.050	0.102	0.002	0.019	0.042
10b-2Cat	0.440	5.765	0.399	8.742	1.000	-0.002	0.894	-1.450	0.009	0.049	0.109	0.004	0.017	0.040
11b-2Cat	0.341	6.243	0.363	9.239	1.000	0.006	0.770	-1.249	0.010	0.057	0.106	0.000	0.007	0.018
12b-2Cat	0.356	5.469	0.337	9.925	0.999	0.006	0.779	-1.954	0.011	0.059	0.097	0.002	0.005	0.018
1c-2Cat	0.482	6.236	0.477	7.296	0.999	0.003	0.933	-0.506	0.010	0.059	0.110	0.005	0.028	0.073
2c-2Cat	0.453	6.688	0.460	7.694	0.998	0.015	0.916	-0.454	0.016	0.052	0.109	0.003	0.031	0.063
3c-2Cat	0.476	7.102	0.427	8.179	0.995	0.041	1.030	-0.516	0.014	0.053	0.090	0.015	0.053	0.090
4c-2Cat	0.428	7.360	0.399	8.742	0.996	0.028	0.980	-0.661	0.017	0.052	0.103	0.012	0.041	0.080
5c-2Cat	0.425	6.948	0.363	9.239	0.992	0.058	1.002	-1.128	0.013	0.047	0.088	0.012	0.039	0.073
6c-2Cat	0.386	6.793	0.337	9.925	0.995	0.040	0.937	-1.529	0.014	0.049	0.097	0.007	0.027	0.047
7c-2Cat	0.569	5.299	0.477	7.296	0.986	0.072	0.999	-0.961	0.010	0.041	0.086	0.010	0.030	0.067
8c-2Cat	0.439	7.544	0.460	7.694	0.997	0.025	0.942	-0.045	0.013	0.052	0.102	0.011	0.043	0.084
9c-2Cat	0.448	8.828	0.427	8.179	0.996	0.037	1.087	0.378	0.015	0.057	0.104	0.027	0.087	0.141
10c-2Cat	0.421	10.196	0.399	8.742	0.990	0.100	1.133	0.879	0.015	0.050	0.094	0.040	0.111	0.194
11c-2Cat	0.445	10.861	0.363	9.239	0.985	0.167	1.312	1.193	0.014	0.050	0.088	0.073	0.195	0.306
12c-2Cat	0.369	13.950	0.337	9.925	0.988	0.186	1.286	2.509	0.014	0.051	0.095	0.094	0.231	0.373

Table 40. Comparison of Rescaled and Predicted G^{2*} Sampling Distributions for the Five Category Items in Test 5Cat12a (Subsets)

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-5Cat	0.586	21.271	0.556	23.036	0.998	0.038	1.011	-0.865	0.014	0.059	0.103	0.013	0.052	0.091
2a-5Cat	0.584	22.136	0.537	24.338	0.998	0.037	1.034	-1.115	0.014	0.056	0.107	0.015	0.058	0.107
3a-5Cat	0.540	24.950	0.540	24.853	0.998	0.050	1.001	0.100	0.014	0.048	0.094	0.014	0.049	0.098
4a-5Cat	0.581	24.584	0.546	25.716	0.999	0.030	1.039	-0.563	0.014	0.051	0.098	0.016	0.061	0.117
5a-5Cat	0.570	26.087	0.527	28.331	0.993	0.183	1.030	-0.964	0.014	0.060	0.093	0.017	0.060	0.100
6a-5Cat	0.548	29.001	0.554	29.453	0.998	0.046	0.980	-0.175	0.012	0.050	0.099	0.009	0.043	0.087
7a-5Cat	0.577	21.946	0.556	23.036	0.999	0.009	1.012	-0.549	0.008	0.048	0.085	0.008	0.047	0.083
8a-5Cat	0.530	24.090	0.537	24.338	0.999	0.023	0.979	-0.100	0.014	0.053	0.096	0.011	0.044	0.078
9a-5Cat	0.520	25.443	0.540	24.853	0.999	0.021	0.974	0.305	0.011	0.049	0.108	0.008	0.044	0.097
10a-5Cat	0.573	24.356	0.546	25.716	0.998	0.047	1.018	-0.648	0.015	0.058	0.101	0.017	0.058	0.101
11a-5Cat	0.509	29.321	0.527	28.331	0.998	0.053	0.981	0.533	0.003	0.057	0.112	0.003	0.055	0.105
12a-5Cat	0.594	26.834	0.554	29.453	0.999	0.042	1.022	-1.318	0.013	0.060	0.109	0.013	0.056	0.104

Table 41. Comparison of Rescaled and Predicted G^{2*} Sampling Distributions for the Five Category Items in Test 5Cat36 (Subsets)

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted (Subset)		Rescaled			Predicted (Subset)		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
1a-5Cat	0.712	33.029	0.696	31.574	0.998	0.049	1.045	0.801	0.012	0.061	0.112	0.027	0.095	0.156
2a-5Cat	0.734	32.286	0.670	33.343	0.999	0.014	1.079	-0.553	0.011	0.050	0.104	0.025	0.092	0.159
3a-5Cat	0.669	35.988	0.674	33.473	0.999	0.027	1.029	1.296	0.011	0.052	0.107	0.020	0.093	0.163
4a-5Cat	0.712	34.682	0.691	33.537	0.999	0.040	1.047	0.631	0.009	0.052	0.105	0.023	0.094	0.157
5a-5Cat	0.769	33.939	0.729	33.232	0.999	0.028	1.064	0.404	0.014	0.050	0.091	0.031	0.087	0.162
6a-5Cat	0.679	39.330	0.755	33.205	0.999	0.055	0.978	2.909	0.009	0.041	0.108	0.012	0.071	0.146
7a-5Cat	0.695	33.622	0.696	31.574	0.999	0.052	1.030	1.085	0.013	0.053	0.104	0.021	0.086	0.150
8a-5Cat	0.699	34.091	0.670	33.343	0.998	0.086	1.053	0.480	0.012	0.049	0.095	0.028	0.084	0.148
9a-5Cat	0.671	35.975	0.674	33.473	1.000	0.015	1.033	1.279	0.013	0.047	0.099	0.019	0.081	0.159
10a-5Cat	0.713	34.622	0.691	33.537	0.999	0.030	1.047	0.590	0.014	0.044	0.110	0.020	0.088	0.162
11a-5Cat	0.756	34.203	0.729	33.232	0.999	0.046	1.050	0.551	0.011	0.052	0.099	0.018	0.088	0.155
12a-5Cat	0.804	33.336	0.755	33.205	0.999	0.040	1.066	0.112	0.012	0.056	0.093	0.027	0.086	0.150
1b-5Cat	0.697	33.927	0.696	31.574	0.999	0.038	1.038	1.237	0.015	0.050	0.097	0.024	0.088	0.146
2b-5Cat	0.694	34.916	0.670	33.343	0.998	0.054	1.059	0.879	0.014	0.062	0.101	0.030	0.097	0.162
3b-5Cat	0.650	37.957	0.674	33.473	0.998	0.074	1.026	2.287	0.014	0.058	0.101	0.029	0.098	0.173
4b-5Cat	0.687	36.452	0.691	33.537	0.999	0.038	1.036	1.508	0.013	0.052	0.095	0.025	0.086	0.155
5b-5Cat	0.746	35.027	0.729	33.232	0.999	0.019	1.049	0.948	0.010	0.054	0.100	0.021	0.093	0.167
6b-5Cat	0.855	31.048	0.755	33.205	0.998	0.062	1.092	-1.123	0.015	0.059	0.093	0.034	0.088	0.138

Table 41 (cont'd).

	Scaling Corrections				Q-Q Plot Regression Coefficients				Type I Error Rates					
	Rescaled		Predicted (Subset)		Rescaled		Predicted		Rescaled			Predicted		
Item	γ	ν	$\hat{\gamma}$	$\hat{\nu}$	Slope	Intercept	Slope	Intercept	.01	.05	.10	.01	.05	.10
7b-5Cat	0.681	34.022	0.696	31.574	0.999	0.021	1.015	1.244	0.008	0.051	0.105	0.014	0.078	0.145
8b-5Cat	0.648	36.571	0.670	33.343	0.998	0.078	1.011	1.665	0.015	0.049	0.092	0.019	0.071	0.132
9b-5Cat	0.681	35.770	0.674	33.473	0.999	0.043	1.044	1.218	0.013	0.053	0.099	0.022	0.094	0.161
10b-5Cat	0.697	36.139	0.691	33.537	0.999	0.042	1.047	1.371	0.007	0.057	0.101	0.026	0.095	0.169
11b-5Cat	0.761	35.925	0.729	33.232	0.999	0.024	1.085	1.451	0.009	0.055	0.108	0.038	0.126	0.216
12b-5Cat	0.781	37.260	0.755	33.205	0.998	0.087	1.093	2.225	0.014	0.058	0.100	0.050	0.145	0.240
1c-5Cat	0.675	34.397	0.696	31.574	1.000	0.013	1.013	1.413	0.012	0.054	0.092	0.017	0.072	0.134
2c-5Cat	0.636	37.732	0.670	33.343	1.000	0.021	1.009	2.164	0.013	0.049	0.102	0.022	0.086	0.148
3c-5Cat	0.698	34.596	0.674	33.473	0.999	0.045	1.051	0.635	0.009	0.044	0.109	0.015	0.092	0.180
4c-5Cat	0.674	36.178	0.691	33.537	0.999	0.033	1.012	1.345	0.007	0.049	0.105	0.014	0.073	0.140
5c-5Cat	0.640	38.011	0.729	33.232	1.000	0.013	0.937	2.179	0.008	0.048	0.111	0.007	0.037	0.096
6c-5Cat	0.700	35.721	0.755	33.205	1.000	0.012	0.961	1.194	0.010	0.054	0.104	0.008	0.046	0.094
7c-5Cat	0.727	31.838	0.696	31.574	0.999	0.016	1.049	0.158	0.011	0.055	0.102	0.020	0.089	0.145
8c-5Cat	0.640	37.458	0.670	33.343	0.999	0.046	1.012	2.058	0.014	0.052	0.101	0.025	0.084	0.145
9c-5Cat	0.683	35.066	0.674	33.473	0.999	0.029	1.037	0.846	0.010	0.058	0.105	0.020	0.093	0.148
10c-5Cat	0.653	36.704	0.691	33.537	1.000	0.017	0.988	1.540	0.009	0.049	0.109	0.011	0.055	0.124
11c-5Cat	0.724	33.599	0.729	33.232	1.000	0.009	0.998	0.192	0.012	0.055	0.096	0.012	0.055	0.096
12c-5Cat	0.693	36.192	0.755	33.205	0.999	0.055	0.956	1.440	0.009	0.041	0.108	0.007	0.037	0.096

IV. C. Comparison of Results Based on the Prediction Equations and Other Methods of Assessing Fit

The utility of the prediction equations to real test data was investigated in the current study. The data used in the validation procedures came from two assessments. The first set of real data were obtained from the 1991-1992 administrations of the QCAI (Lane, 1993). The second data set consisted of student responses to the 1994 administration of the NAEP reading assessment (Allen et al., 1994). Investigation of the utility of the prediction equations was necessary to determine if they could generalize beyond the scope of this study.

If the prediction equations did generalize beyond the scope of this study, then the chi-square distribution, rather than empirically generated sampling distributions of the fit statistics, could be used for significance testing of χ^{2*} and G^{2*}. The appropriate chi-square distribution for significance testing of each item would be determined using the prediction equations.

IV. C. 1. Validation Data Sets

For the QCAI, data from Form A, administered during the Spring of 1991, and Forms A and B, administered during the Spring of 1992, were used. The data sets for each form of the test consisted of eight 5-category items. The sample sizes associated with the three forms were n = 399 for Form A, Spring 1991; n = 459 for Form A, Spring 1992; and n = 466 for Form B, Spring 1992.

For the NAEP data, block 9M from the 1994 Reading Assessment was used. This data set consisted of item responses for n = 1847 examinees to nine items. The nine items consisted of four 2-category items, four 3-category items, and one 4-category item.

The item parameters for each test were estimated using MULTILOG (Thissen, 1991). The item parameter estimates for the three administrations of the QCAI are provided in Table 42. Table 43 presents the estimated item parameters for the items on Block 9M of the 1994 NAEP Reading Assessment. The total information provided by each item, given by Equation 26 with $\Delta\theta = 0.05$, is also provided in Tables 42 and 43.

Table 42. Item Parameters for the Three QCAI Validation Data Sets

Item	a	b_1	b_2	b_3	b_4	Average b	Info
Items from the QCAI, Form A, Spring 1991 Administration							
RNS3	0.913	1.095	1.870	2.184	2.680	1.957	0.932
PRP2	1.594	-1.019	-0.075	0.707	1.490	0.276	3.408
PNS2	1.297	0.008	1.211	1.450	1.724	1.098	1.980
PPA1	1.389	0.123	0.818	1.222	1.921	1.021	2.251
RPG1	1.101	-1.150	-0.114	1.786	3.195	0.929	1.951
PGE1	1.395	-0.740	0.101	0.714	3.126	0.800	2.643
PST1	1.316	-1.761	0.589	1.283	2.098	0.552	2.683
PST2	1.675	0.811	1.436	1.695	2.029	1.493	2.520
Items from the QCAI, Form A, Spring 1992 Administration							
RNS3	1.142	0.986	1.598	2.416	3.093	2.023	1.407
PRP2	1.553	-1.064	-0.108	0.608	1.339	0.194	3.210
PNS2	1.168	0.148	1.138	1.525	1.649	1.115	1.607
PPA1	0.798	-0.296	0.661	0.929	3.455	1.187	0.961
RPG1	0.821	-1.596	-0.523	1.881	3.622	0.846	1.166
PGE1	1.223	-1.112	-0.117	0.672	3.186	0.657	2.249
PST1	1.367	-1.557	0.246	1.004	1.686	0.345	2.821
PST2	1.452	1.129	2.030	2.295	2.830	2.071	1.961
Items from the QCAI, Form B, Spring 1992 Administration							
PNS3	0.955	-1.677	-0.132	1.336	3.093	0.655	1.585
PCO2	1.578	-0.460	0.566	1.019	1.285	0.602	2.829
PCO4	1.768	-0.289	0.573	1.101	1.535	0.730	3.448
PRP1	1.891	0.605	1.227	1.348	1.675	1.214	2.990
RPN1	1.327	-2.134	-0.268	0.500	1.653	-0.063	2.806
PGE3	1.207	-0.477	0.312	0.764	1.410	0.502	1.900
PST4	1.698	0.301	0.976	1.128	1.458	0.966	2.662
PPP1	0.934	-1.553	-0.724	-0.294	0.564	-0.502	1.264

Table 43. Item Parameters for Block 9M of the 1994 NAEP Reading Assessment

Item	*a*	b_1	b_2	b_3	*Average b*	*Info*
1	1.783	-1.487	--	--	-1.487	1.675
2	0.845	-0.881	--	--	-0.881	0.696
3	1.236	-1.552	1.432	--	-0.060	1.972
4	1.527	-1.249	0.922	2.880	0.851	3.173
5	0.814	-0.535	--	--	-0.535	0.677
6	1.214	-1.017	1.629	--	0.306	1.903
7	1.314	-1.185	1.064	--	-0.060	2.122
8	1.113	-1.679	--	--	-0.881	0.904
9	1.160	-1.353	1.642	--	0.145	1.788

IV. C. 2. Validation of the Multilevel Prediction Equations

Initially, the QCAI data sets were analyzed to evaluate the utility of the multilevel prediction equations. For each item on each form of the QCAI, the pseudocounts-based item fit statistics χ^{2*} and G^{2*} were computed using the SAS program found in Appendix B. The value of the mean posterior variance $\bar{p}$ was also obtained for each of the three QCAI tests using the program listed in Appendix B.

Then, the predicted scaling factor $\hat{\gamma}$ and degrees of freedom value $\hat{\nu}$ were calculated for each item using the multilevel prediction equations. The multilevel coefficients for predicting the scaling factors $\hat{\gamma}$ and degrees of freedom values $\hat{\nu}$ for the Pearson and likelihood ratio statistics are provided in Table 24.

One set of two equations for estimating $\hat{\gamma}$ and $\hat{\nu}$ for the Pearson χ^{2*} statistics, and a second set of two equations for estimating $\hat{\gamma}$ and $\hat{\nu}$ for the likelihood ratio G^{2*} statistics, were evaluated. In applying the multilevel prediction equations to the QCAI data, a single value for

the random effects coefficient had to be selected from Table 24 for each QCAI test. Because all forms of the QCAI contained eight 5-category items, it was decided that the random effects intercept term would be chosen from one of the 5-category 12 items tests, 5Cat12a, 5cat12b, or 5cat12c.

The data in Table 24 shows that the random effects terms for each of the four multilevel prediction models were similar for tests 5Cat12a, 5cat12b, and 5cat12c. The coefficient from the simulation study whose total test information was most similar to the total test information for each QCAI form was selected. In each case for the QCAI data, Test 5Cat12c was the test from the study most similar total test information. Therefore, the random effects coefficient for each test came from Test 5Cat12c.

The final four multilevel equations for predicting the scaling corrections that were applied to the QCAI assessments are given in Equations 34 to 37. Equations 34 and 35 were applied to the Pearson statistics, and Equations 36 and 37 were applied to the likelihood ratio form of the statistic for each item on each form of the QCAI.

$$\hat{\gamma}_{\chi_2^*} = -0.375 + 0.114 * a - 5.983 * \overline{p} + 4.904 * \sqrt{\overline{p}} + .000 * a - .027 \quad (34)$$

$$\hat{\nu}_{\chi_2^*} = 10.413 + 1.607 * a + 75.695 * \overline{p} - 70.983 * \sqrt{\overline{p}} + 4.336 * ncat - 3.783 * a + 5.763 \quad (35)$$

$$\hat{\gamma}_{G_2^*} = 0.621 - 0.003 * a - 1.044\overline{p} + .096 * \sqrt{\overline{p}} + .003 * a - .005 \quad (36)$$

$$\hat{\nu}_{G^*} = 25.927 + 4.102 * a + 146.700 * \overline{p} - 147.190 * \sqrt{\overline{p}} + 5.197 * ncat - 2.697 * a + 2.242 \quad (37)$$

The values for the predictor variables used in the equations were *a* and *ncat* obtained from the individual QCAI items, and $\overline{p}$ and $\sqrt{\overline{p}}$, obtained from the single sample of subjects on

each of the three tests. Decisions regarding item fit were made by comparing the observed fit statistic rescaled by $\hat{\gamma}$ to the chi-square distribution with predicted degrees of freedom $\hat{\nu}$.

The fit of the NAEP items were evaluated using the multilevel prediction models. Block 9M of the NAEP assessment consisted primarily of 2- and 3-category items. The scaling corrections predicted from the multilevel models did not adequately rescale the fit statistic distributions for 2- and 3–category items from the simulated fit statistic distributions. Therefore, it was not expected that the prediction equations would have practical utility when applied to 2- and 3–category items.

IV. C. 3. Decisions of Fit For the Three Forms of the QCAI Assessment

Decisions regarding item fit for the QCAI items were made several ways. Using the multilevel prediction equations, scaling corrections for each item on each form of the QCAI were predicted using Equations 34 to 37. Then, the observed fit statistic for each item was rescaled by $\hat{\gamma}$. The rescaled statistic was tested for significance by comparing it to the chi-square distribution with predicted degrees of freedom $\hat{\nu}$.

These results were compared to the results obtained by Stone et al. (1993). In that study, the researchers assessed the fit of the items on these same forms of the QCAI. The researchers generated empirical sampling distributions of the statistics and obtained critical values for decisions of item fit from the percentiles of the empirically generated sampling distributions.

Decisions of item fit were also carried out using the resampling method described by Stone (2000). The Stone (2000) method used the initial item parameter estimates with a randomly selected ability value when computing each fit statistic. The item parameters were not re-estimated for each statistic, as they were in the simulation aspect of this study. After sampling distributions were formed in this manner, the scaling corrections for each item were obtained.

Decisions of item fit were then evaluated by comparing the rescaled statistic to the appropriate chi-square distribution.

Further, graphical displays of the observed versus expected score distributions for each item score response category were plotted on the same graph. Graphs in which the observed and expected score distributions overlapped indicated fit, while graphs showing large discrepancies between the observed and expected score distributions indicated item misfit.

Table 44 compares the decisions regarding item fit for the Pearson χ^{2*} form of the statistic for items on the QCAI. The table summarizes the decisions of fit made using three methods: 1) the multilevel prediction equations given by Equations 34 to 37; 2) the method employed by Stone et al. (1993); and 3) the resampling method employed by Stone (2000), based on n = 200 resamples. Table 45 compares the decisions regarding item fit for the likelihood ratio G^{2*} form of the statistic using these same four methods for items on the QCAI. All decisions of item fit in Tables 44 and 45 were made at the $\alpha = .05$ level of significance. The asterisks denote the items identified as misfitting.

Table 44. Decisions of Fit Based on χ^{2*} Statistics for the QCAI Items

		Multilevel Prediction Equations				Stone et al., 1993	Stone, 2000	
Item	χ^{2*}	$\hat{\gamma}$	$\hat{\upsilon}$	p	Decision	Decision	p	Decision
Items from the QCAI, Form A, Spring 1991 Administration								
RNS3	14.21	0.47	18.30	0.035		*	0.021	*
PRP2	2.61	0.57	18.83	0.999			0.749	
PNS2	10.88	0.52	18.60	0.293		*	0.057	
PPA1	4.71	0.54	18.67	0.965			0.632	
RPG1	12.94	0.50	18.45	0.098		*	0.032	*
PGE1	4.92	0.54	18.68	0.957			0.363	
PST1	3.66	0.53	18.62	0.991			0.783	
PST2	12.18	0.58	18.90	0.278		*	0.038	*
Items from the QCAI, Form A, Spring 1992 Administration								
RNS3	4.47	0.46	18.65	0.943			0.561	
PRP2	7.25	0.52	18.97	0.737		*	0.129	
PNS2	11.04	0.47	18.67	0.166		*	0.032	*
PPA1	5.24	0.41	18.38	0.810			0.693	
RPG1	5.37	0.42	18.40	0.798			0.504	
PGE1	5.69	0.47	18.71	0.848			0.383	
PST1	4.37	0.50	18.82	0.964			0.521	
PST2	5.85	0.51	18.89	0.871			0.187	
Items from the QCAI, Form B, Spring 1992 Administration								
PNS3	16.24	0.50	18.30	0.019		*	0.012	*
PCO2	14.22	0.59	18.79	0.149		*	0.010	*
PCO4	13.42	0.62	18.93	0.241		*	0.011	*
PRP1	4.72	0.63	19.03	0.991			0.614	
RPN1	8.27	0.55	18.59	0.663			0.229	
PGE3	16.39	0.53	18.50	0.031		*	0.006	*
PST4	12.25	0.61	18.88	0.320		*	0.024	*
PPP1	7.25	0.49	18.28	0.685			0.369	

Note. The column labeled 'Stone et al., 1993' contains decisions of fit based on G^{2*}, because the Pearson form was not computed in that study.

Table 45. Decisions of Fit Based on G^{2*} Statistics for the QCAI Items

		Multilevel Prediction Equations				Stone et al., 1993	Stone, 2000	
Item	G^{2*}	$\hat{\gamma}$	$\hat{\upsilon}$	P	Decision	Decision	p	Decision
Items from the QCAI, Form A, Spring 1991 Administration								
RNS3	13.27	0.44	18.74	0.034	*	*	0.037	*
PRP2	2.71	0.44	19.70	0.997			0.742	
PNS2	12.43	0.44	19.28	0.076		*	0.034	*
PPA1	4.71	0.44	19.41	0.932			0.633	
RPG1	13.31	0.44	19.01	0.047	*	*	0.029	*
PGE1	5.21	0.44	19.42	0.890			0.419	
PST1	3.9	0.44	19.31	0.975			0.806	
PST2	13.28	0.44	19.81	0.048	*	*	0.028	*
Items from the QCAI, Form A, Spring 1992 Administration								
RNS3	4.55	0.41	18.85	0.889			0.624	
PRP2	7.89	0.41	19.42	0.438		*	0.102	
PNS2	11.87	0.41	18.88	0.048	*	*	0.024	*
PPA1	5.52	0.41	18.36	0.761			0.672	
RPG1	5.82	0.41	18.39	0.713			0.485	
PGE1	5.80	0.41	18.96	0.717			0.443	
PST1	4.64	0.41	19.16	0.911			0.491	
PST2	5.43	0.41	19.28	0.824			0.260	
Items from the QCAI, Form B, Spring 1992 Administration								
PNS3	17.36	0.45	19.09	0.006	*	*	0.008	*
PCO2	14.57	0.45	19.96	0.031	*	*	0.010	*
PCO4	14.51	0.45	20.23	0.044	*	*	0.008	*
PRP1	3.96	0.45	20.40	0.986			0.791	
RPN1	8.29	0.45	19.61	0.506			0.268	
PGE3	16.2	0.45	19.44	0.012	*	*	0.009	*
PST4	12.98	0.45	20.13	0.097		*	0.021	*
PPP1	8.24	0.45	19.06	0.513			0.309	

Tables 44 and 45 present the decision consistency for three methods of assessing item fit for the items on three forms of the QCAI assessment. Graphical displays of the observed versus expected score distributions indicated that the results by Stone et al. (1993) most accurately identified item misfit.

The results in Tables 44 show that using the multilevel prediction equations resulted in no misfitting items for the Pearson form of the statistic. Thus, for the Pearson form of the pseudocounts-based statistic, the prediction technique did not correctly identify item misfit for 5-category items.

However, for the likelihood-ratio form of the statistics, the results in Table 45 show that the prediction technique resulted in eight items being classified as misfitting items at the $\alpha = .05$ level of significance. These eight items were also classified as misfitting by Stone (2000) and Stone et al. (1993).

Three additional items were classified as misfitting by Stone et al. (1993), which obtained critical values from the percentile points of the empirically generated sampling distributions. Two of these three additional items were classified as misfitting by Stone (2000), which used a resampling method, and obtained critical values from a chi-square distribution. The p-values for the same two items classified as misfitting by Stone (2000) were significant at $\alpha = .10$ using the prediction equations.

Overall, the prediction equations were not adequate for identifying item misfit for the Pearson χ^{2*} form of the pseudocounts-based item fit statistics. However, for the likelihood-ratio form of the pseudocounts-based statistic, use of the prediction equations resulted in accurate identification of item misfit for 5-category items. Results suggested that a higher level of significance of $\alpha = .10$ should be employed.

IV. C. 4. Decisions of Fit For the NAEP Assessment

Table 46 compares the decisions regarding item fit for the Pearson and likelihood ratio forms of the statistic for items for the NAEP data. The table summarizes the decisions of fit made using two methods: 1) the multilevel prediction equations given in Table 25, and 2) the resampling method employed by Stone (2000), based on n = 200 resamples.

The random effects coefficients from Table 25 for the NAEP items came from Test 2Cat12b for the 2-category items, because in the simulation phase of this study, the items on this test were most promising in terms of following scaled chi-square distributions. The random effects coefficients came from Test 3Cat12a for the 3-category items, and Test 4Cat12c for the 4-category item. These were the tests from the study at the 3 and 4 item score category levels whose total test information were most similar to the total test information of this NAEP assessment.

Based on the simulation phase of this study, there were questions about the appropriateness of the prediction equations for many 2- and 3-category items. Therefore, the reliability of the results found in Table 46 are questionable.

Table 46 shows that more items were flagged as misfitting using the prediction equations than were identified as misfitting using the resampling method of Stone, (2000). Graphical displays of observed and expected score distributions at each score category level indicated that items 2, 3, 6, 7, and 9 showed small deviations between the observed and expected score distributions. The results based on the multilevel prediction equations for χ^{2*} and G^{2*}, not surprisingly, did not identify these same items.

The results based on the Pearson χ^{2*} statistics identified five items as misfitting, but only two items that were identified by Stone (2000). Using the prediction equations for the likelihood

ratio G^{2*} statistic, seven items were identified as misfitting. Three of these items were also identified by Stone (2000).

The results found in Table 46 suggest that the multilevel prediction equations did not adequately identify item misfit when either the χ^{2*} or G^{2*} statistics were used for the 2- and 3-category NAEP items. This was anticipated, because the empirically generated sampling distributions of 2- and 3-category items were not adequately rescaled by the predicted scaling corrections.

Item fit was not evaluated using the prediction equations estimated for subsets of the overall data set. These additional equations had been estimated because the multilevel equations could not be validly applied to 2- and 3- category items. However, analysis of the additional regression equations for item subsets showed that they also yielded inadequate scaling corrections, and so their utility for real items was not investigated.

Table 46. Decisions of Fit Based on χ^{2*} and G^{2*} Fit Statistics for the Items in Block 9M of the 1994 NAEP Reading Assessment

		Statistic	Multilevel Prediction Equations				Stone 2000	
Item	*ncat*	χ^{2*}	$\hat{\gamma}$	$\hat{\upsilon}$	p	Decision	p	Decision
1	2	4.5	0.32	7.62	0.000	*	0.12	
2	2	11.12	0.18	6.89	0.000	*	0.00	*
3	3	14.8	0.42	7.41	0.008	*	0.00	*
4	4	7.01	0.82	14.75	0.595		0.47	
5	2	4.71	0.18	6.86	0.000	*	0.66	
6	3	7.67	0.78	7.39	0.169		0.19	
7	3	9.02	0.79	7.47	0.107		0.00	*
8	2	3.12	0.22	7.10	0.002	*	0.44	
9	3	6.75	0.77	7.35	0.151		0.17	
Item	*ncat*	G^{2*}	$\hat{\gamma}$	$\hat{\upsilon}$	p	Decision	p	Decision
1	2	4.14	0.20	10.03	0.000	*	0.19	
2	2	11.32	0.20	6.78	0.000	*	0.00	*
3	3	15.05	0.49	8.26	0.000	*	0.00	*
4	4	7.36	0.35	18.69	0.119		0.43	
5	2	4.67	0.20	6.67	0.000	*	0.66	
6	3	7.93	0.49	8.19	0.034	*	0.18	
7	3	9.09	0.49	8.53	0.026	*	0.00	*
8	2	3.09	0.20	7.70	0.000	*	0.47	
9	3	6.87	0.49	8.00	0.069		0.18	

CHAPTER V

V. SUMMARY AND CONCLUSIONS

V. A. Summary

Item response theory (IRT) offers several advantages over other testing theories that are attained when the item response model fits the data. One aspect of evaluating model-data fit entails assessing item fit, or the match between the item response data and the IRT model being implemented. This most often involves the computation of a chi-square test statistic.

Chi-square goodness-of-fit statistics in the context of IRT are formed by first rank-ordering examinees according to a point estimate of their ability, and then dividing them into some number of subgroups based on this ability ranking. Then, within each of the subgroups, the observed and expected score distributions are computed, and compared through the computation of a chi-square test statistic. Statistical significance is determined using the chi-square distribution with the appropriate degrees of freedom.

'Traditional' goodness-of-fit statistics such as this group examinees according to a point estimate of their ability. The computed value of the goodness-of-fit statistic depends greatly on this single point estimate. For performance-based assessments, the number of items on the assessments is often small. Shorter test lengths can result in a lack of precision in individual examinee ability estimates for performance-based assessments. The imprecision in the

estimation of ability can cause classification errors in the item fit tables, which may result in increased Type I error rates and decreased power for significance tests of fit.

In the current study, a method of detecting item misfit that accounted for imprecision in ability estimation found on shorter tests was investigated. The statistic utilized a pseudo-observed score distribution as a discrete representation of the entire posterior distribution of ability. Due to dependencies in the pseudo-observed score distribution, or pseudocounts, the statistic could not be tested for significance using the theoretical chi-square distribution.

Past research has indicated that the Pearson χ^{2*} and likelihood ratio G^{2*} forms of the pseudocounts-based statistic may follow scaled chi-square distributions. This research indicated that the scaling corrections could be obtained using the method of moments with the means and variances of empirical sampling distributions. Further, studies had suggested that the scaling corrections could be predicted from item and sample characteristics so that the theoretical chi-square distribution could be used for significance testing, without requiring the generation of empirical sampling distributions,

The purpose of this study was to determine whether item and sample characteristics could be used to predict the scaling corrections needed for significance testing of the pseudocounts-based fit statistics χ^{2*} and G^{2*}, for tests of different lengths consisting of 2, 3, 4, and 5-category items. The main goal was to provide a set of two prediction equations for predicting the scaling factor and degrees of freedom values needed for significance testing of the statistics versus theoretical chi-square distributions. General equations that would predict the scaling corrections across item score category levels and test length were desired.

V. B. Research Questions

The following research questions were under study:

1. For items modeled by the GRM having 2, 3, 4, and 5 response categories, can scaling corrections obtained from Monte Carlo based empirical sampling distributions be used to rescale simulated fit statistic distributions so that significance tests against a known chi-square distribution can be performed?
2. For items modeled by the GRM having 2, 3, 4, and 5 response categories, can appropriate scaling corrections be predicted from item and sample characteristics and used to rescale simulated fit statistic distributions so that significance tests against a known chi-square distribution can be performed?
3. Can the prediction equations derived from empirical data be generalized to items having different parameters than those in the simulation study, for purposes of significance testing of the observed fit statistics?

V. C. Conclusions

The following conclusions were drawn with respect to each research question.

V. C. 1. Research Question 1

To answer research question 1, data was simulated under the multidimensional GRM (MGRM) for conditions varying test length (12a, 12b, 12c, 24, and 36 items) and number of score categories (2, 3, 4, and 5). For each test configuration, data sets were simulated using realistic item parameters, and χ^{2*} and G^{2*} were computed for each item. The process of generating data and computing the fit statistics was repeated to form sampling distributions of each fit statistic. The empirically generated sampling distributions were examined through analysis of Q-Q plots, and Type I error rates.

Overall, it was found sampling distributions of both the Pearson and likelihood ratio pseudocounts based fit statistics were approximated fairly well by theoretical chi-square distributions, when the scaling corrections were obtained using the method of moments with the means and variances of the empirical sampling distributions. Without this important finding, it would not have been conceivable that scaling corrections could be predicted from item and sample data, rather than found using empirically generated sampling distributions.

In general, across items and test lengths, the Q-Q plots of rescaled likelihood ratio G^{2*} statistics and Pearson χ^{2*} statistics were linear and followed the line $y = x$. Further, the Type I error rates fell in the range expected with 95% confidence, indicating a match between the rescaled empirical and theoretical chi-square distributions.

More specifically, the rescaled likelihood ratio G^{2*} distributions provided a closer match to the theoretical chi-square distributions than the Pearson χ^{2*} distributions. In addition, the sampling distributions of items with more score response categories and on longer tests more closely approximated theoretical chi-square distributions than items with fewer score response categories, and items on shorter tests.

V. C. 2. Research Question 2

To answer research question 2, multilevel-prediction equations that used sample characteristic (e.g., mean posterior variance and root mean posterior variance) and item characteristics (e.g., item discrimination and number of score categories) were estimated from the data. For each item, the predicted scaling corrections were estimated using the final multilevel prediction equations, and used to rescale the empirically generated sampling distributions. Specifically, the statistics in each sampling distribution were rescaled by the

predicted scaling factor, and compared to the theoretical chi-square distribution with predicted degrees of freedom.

Overall, across item score category levels, the multilevel prediction equations yielded scaling corrections that did not adequately rescale the simulated fit statistic distributions so that they approximated theoretical chi-square distributions. For the 2- and 3- category tests, the sampling distributions of Pearson and likelihood ratio statistics that were rescaled by the predicted scaling factor did not approximate the theoretical chi-square distributions with predicted degrees of freedom. The slopes and intercepts of regression lines fitted to Q-Q plots deviated from 1 and 0, respectively, and Type I error rates did not fall in the range expected with 95% confidence intervals.

For tests consisting of items with 4 and 5 score response categories, the performance of the multilevel prediction equations with respect to rescaling the simulated fit statistic distributions depended on the type of chi-square statistic that was utilized. For the Pearson form of the pseudocounts-based statistic, even at the 5-category level, only approximately half of the items on the simulated tests were adequately rescaled by the predicted scaling corrections. The results for the likelihood ratio form of the statistics at the higher score category levels were better. For the likelihood ratio statistic for the 4- and 5-category tests, the slopes and intercepts of regression lines fitted to Q-Q plots were close to 1 and 0, respectively, and Type I error fell in the range expected with 95% confidence intervals, for most items in the simulated distributions.

Overall, based on the simulated sampling distributions, it was concluded that the scaling corrections obtained using the multilevel prediction equations did adequately rescale the sampling distributions of G^{2*} statistics for items with 4 or 5 score response categories. The results based on the Pearson χ^{2*} form of the statistic for 4- and 5-category items were less

promising. Further, the results indicated that the prediction equations did not adequately rescale sampling distributions of Pearson χ^{2*} statistics or likelihood ratio G^{2*} statistics for items with 2 or 3 score response categories.

Because the multilevel prediction equations could not be applied to 2- and 3-category items, a different approach was taken in an attempt to estimate prediction equations. Rather than attempting to estimate a single set of two general equations across tests, as was done with the multilevel prediction equations, this second approach attempted to predict the scaling corrections within subsets of items that were more similar. The scaling corrections obtained using these additional prediction equations based on item subsets also failed to adequately rescale the simulated sampling distributions for the 2- and 3-category items for both the Pearson χ^{2*} and likelihood ratio G^{2*} forms of the statistics. Further, prediction equations based on item subsets also failed to rescale the χ^{2*} and G^{2*} statistics for the tests at the higher score category levels.

V. C. 3. Research Question 3

To answer research question 3, the multilevel prediction equations were applied to sets of real items to assess their utility. Results based on the multilevel prediction equations were compared against results obtained from other methods of assessing item fit. In answering research question 2, it was found that the multilevel prediction equations did not adequately predict the sampling distributions of either the χ^{2*} or G^{2*} statistics for 2- or 3-category items. As a result, the application of these multilevel prediction equations to additional 2-category and 3-category real items was not reliable.

However, the results from the simulation phase of this study did indicate that the multilevel prediction equations could be used to predict the sampling distributions of the G^{2*} statistics, and potentially the χ^{2*} statistics, for items with more score response categories. The fit

of the items on three forms of the 1991-1992 administration of the QCAI assessment were assessed using the multilevel prediction equations. The assessment of item fit was made using both the Pearson and likelihood ratio forms of the statistic.

Overall, the results of this study indicated that for the likelihood ratio form of the pseudocounts-based fit statistic, the multilevel prediction equations resulted in adequate identification of item misfit for the 5-category items on the three forms of the QCAI. At the α = .05 level of significance, all of the times identified as misfitting by the prediction technique were also flagged as misfitting by Stone et al. (1993). In addition, of the three additional items flagged as misfitting by Stone et al. (1993), two showed statistically significant misfit based on the multilevel prediction equations at the α = .10 level of significance.

Results based on the Pearson χ^{2*} form of the statistics indicated that the multilevel prediction equations could not be applied to 5-category items on the QCAI to identify item misfit. None of the QCAI items were identified as misfitting when the prediction equations were used with the Pearson form of the statistic.

Results from the simulation phase of this study indicated that the multilevel prediction equations could not be used to rescale the sampling distributions of pseudocounts-based fit statistics for 2- or 3-category items. For the 2- and 3-category NAEP items, the identification of item misfit was inadequate when the prediction equations were used with both the Pearson χ^{2*} and likelihood-ratio G^{2*} forms of the pseudocounts-based item fit statistics.

V. D. Final Conclusions

Overall, the results of this study indicated that, for the likelihood ratio form of the pseudocounts-based fit statistic, use of the multilevel prediction equations will result in adequate identification of item misfit for real items having 5 score response categories. A higher level of significance of

$\alpha = .10$ should be employed when making decisions of item fit. It is likely that the utility of the prediction equations would extend to real items having 4-score response categories. However, the application to real items with 4 score response categories was limited to only one item. These results do not generalize to the Pearson form of the statistic. That is, results indicated that the prediction equations should not be used to identify item misfit for 4- and 5-category items when χ^{2*} is used.

Further, the results of this study indicate that the prediction equations cannot be utilized to assess the fit of 2- or 3-category items when either the Pearson χ^{2*} or likelihood-ratio G^{2*} forms of the pseudocounts-based statistics are utilized.

V. E. Implications for Further Study

This study contributed to the knowledge of the sampling distribution of the pseudocounts-based item fit statistics, χ^{2*} and G^{2*}. The results indicated that for 2-, 3-, 4- and 5-category items, the sampling distributions of the statistics may follow scaled chi-square distributions.

While the results indicated that general prediction equations for predicting the scaling corrections across test lengths and item score category levels could not be found, the results did show some potential for the use of prediction equations with 4- and 5-category items, in particular when the G^{2*} statistic is utilized.

The prediction equations did not adequately rescale the pseudocounts-based fit statistics for 2- and 3-category items. However, the results of the study remain promising for applications to many performance-based assessments. Many performance-based assessments, like the QCAI assessment utilized in this study, consist of items having 4- or 5-score response categories. The potential for using the prediction equations to assess item fit for such assessments remains high.

However, the utility of the prediction equations for the NAEP assessment, which consists of items having 2 – 4 score category levels, is limited.

Several factors in the current study could have had a negative impact on the quality of the estimated multilevel prediction equations. First, the main application of the pseudocounts-based fit statistic is to shorter, performance-based assessments, which are likely to consist of 4- and 5-category items. In the current study, 2- and 3-category items were included in order to find a set of general prediction equations across a wide variety of items. However, the relationships seen in the data for the 4- and 5-category items were not necessarily seen for 2-category items. In addition, the 2-category tests added a significant amount of variation in the scaling factors and degrees of freedom values, that potentially had a negative impact on the quality of the estimation of the prediction equations for items with more score response categories. Therefore, a study that focused on the prediction of scaling corrections for tests consisting of only 4- and 5-category items could further investigate the utility of the prediction technique for items with more score response categories. In addition, the fit of only one real item not included in the simulation phase of the study was assessed. Including more 4-category items would allow the utility of the prediction equation for more real 4-category items to be assessed.

Second, the results of the current study indicated that the item discrimination parameter was related to the degrees of freedom of the likelihood ratio form of the statistic. In this study, six different values of the item discrimination parameter and two sets of threshold parameters were used in each test configuration. Results indicated that the values of the threshold parameters did not impact the scaling corrections. Therefore, a study that included items with a wider selection of *a* values might serve to better predict the scaling corrections.

Further, inspection of several item fit tables indicated that the item fit statistics for items with higher discrimination parameters were larger. This could be due to the fact that for highly discriminating items, the score distributions across examinees were spread over fewer rows of the item fit tables, and the expectations in the cells at the extremes of the ability scale were small. Including items with a wider selection of a values might allow the impact of the item discrimination on the cell expectations to be further investigated.

Third and related to this, in the current study, the values of the fit statistics and the shapes of the sampling distributions were affected by small expected values in the item fit tables. This was the case for the Pearson form of the statistic in particular. In the current study, the statistics were computed over the range $-2 \leq \theta \leq 2$ due to sparseness in the expected cell counts beyond that range of the ability scale. Further, any cells for which the expected values were less than 0.01, or for which the pseudocounts were equal to 0, were excluded from the computation of the fit statistic. However, the statistics, especially the Pearson form, were still impacted by small expected pseudocounts.

Other researchers have addressed the problem with small expected values in other ways. Orlando and Thissen (2000), for example, handled the problem by starting at each end of the ability scale, and collapsing cells toward the middle of the list until the expected values in each cell were sufficiently large. An approach like this could be taken with the current statistics as well, to assess how that would impact the sampling distributions of the statistics.

It is possible that even by carrying out a new study that includes more 2- and 3-category items with a wide range of a values, a singe set of prediction equations would not adequately predict the scaling corrections for the pseudocounts-based statistics across different category levels. In that case, item level prediction equations may be necessary. For instance, Ankenmann

(1994) attempted to predict the scaling corrections for 5-category items, and found that test level predictions were not possible. He found that item level predictions worked fairly well, when the item level equations were matched to real items based on either item information or item discrimination.

In the current study, it was not possible to estimate item level prediction equations, because each item appeared at most 3 times in the study. A study that produced more replications for each item could be better, because then regression analyses could be done at the item level.

Although the prediction equations obtained from this study could not adequately be used to identify item misfit for 2 and 3 category items, there are several methods that could be used to assess the fit of items when the pseudocounts-based fit statistics, χ^{2*} or G^{2*} are utilized. The resampling method employed by Stone (2000) worked well in identifying item misfit, and produced results quickly. Alternatively, one could generate sampling distributions for each item, as was done in the current study, and use the generated sampling distributions for significance testing. This method is more time consuming than the Stone (2000) resampling method, but may become less computer and time intensive with the development of computer resources, and could eliminate the need for prediction equations.

APPENDIX A

SAS Data Generation Program

To Generate Simulated Data Sets Under the Multidimensional Graded Response Model

```
* program to generate item response data under MGRM;
* Dissertation-- Mary Hansen;
options mprint mlogic notes nodate nonumber;
%global abilseed respfile outfile l;
/*  Control information */
        filename wrkdir 'working-directory';
        %let nitems=12;
        %let nsubj=2000;
    %let ndims = 6;
    %let sigmatrix = {1 0 0 0 0 0, 0 1 0 0 0 0, 0 0 1 0 0 0, 0 0 0 1 0 0, 0 0 0 0 1 0, 0 0 0 0 0 1};
        %let nparm=2;
        %let nparmDim = 7;
        %let maxcat=2;
        %let missing=9;
        %let d=1;
        %let nreps=1;
        %let parmfile='parameterfile.txt';
        %let seedmvn=0;
        %let seedresp=0;
        %let seed=0;
        %let sigma=1;
        %let mean=0;
        %let missing=9;
        %let d=1;  /*d parameter for probability functions */
%macro defineFiles;

        %let outfile='outfile.dat';/*resp, abil for all dims*/
    %let respfile='respfile.dat';     /* responses */
    %let abilseed= 'abilfile.dat'; /*abil seed extra dim*/
%mend defineFiles;
%macro genthetas;
data thetas;
        retain seed &seedmvn;
        seed=%eval(&seedmvn);
    array theta(*) theta1-theta&ndims;
/* generate a N(0,1) ability for each dim for each person*/
                do subjCounter=1 to &nsubj;
                do dimCounter=1 to &ndims;
                        call rannor(seed, theta(dimCounter));
                                file wrkdir(&abilseed) mod;
                        put "&l" '  ' seed;
                        end;
                        output;
                end;
run;
%mend genthetas;
```

```
%definefiles;
%genthetas;
/*  Extract item parameter information */
data item_par_full;
  infile wrkdir(&parmfile) missover;
  input model $ ncat x1 x7-x&nparmdim;
run;
data item_par;
   set item_par_full;
   seed=%eval(&seed);
   retain seed &seed;
   array y{*} x1-x&nparmdim;
   keep p;
   array a{*} a2-a6;
/* make the dim params for the 5 minor dims U(0,.4) */
          do until (x2<.4);call ranuni(seed,x2);            end;
          do until (x3<.4);call ranuni(seed,x3); end;
          do until (x4<.4);call ranuni(seed,x4); end;
          do until (x5<.4);call ranuni(seed,x5); end;
          do until (x6<.4);call ranuni(seed,x6); end;
do j=1 to &nparmdim;
     p=y{j};
     output;
   end;
run;
/*  Create a row vector with item parameters as elements */
proc transpose data=item_par out=item_par prefix=p;
          var p;
run;
/* make a vector of only model type */
data model_type;
          set item_par_full;
          keep model;
run;
proc transpose data=model_type out=model_type prefix=model;
          var model;
run;
/* make a vector of only ncat */
data ncat_info;
     set item_par_full;
           keep ncat;
run;
```

```
proc transpose data=ncat_info out=ncat_info prefix=ncat;     var ncat;
run;
****Multidimensional Graded Response Model*********;
%macro mgr;
 do;
    do;
     if resp =  (ncat[j] - 1) then
        xx=1/(1+exp(-&d*(( p{j,1}*theta1+ p{j,2}*theta2+ p{j,3}*theta3 + p{j,4}*theta4+
p{j,5}*theta5+p{j,6}*theta6) -
(p{j,5+resp+1}*p{j,1}+p{j,5+resp+1}*p{j,2}+p{j,5+resp+1}*p{j,3}+p{j,5+resp+1}*p{j,4}+p{j
,5+resp+1}*p{j,5}+p{j,5+resp+1}*p{j,6} )) ));
     else if resp=0 then
        xx=1-1/(1+exp(-&d*(( p{j,1}*theta1+ p{j,2}*theta2+ p{j,3}*theta3 +
p{j,4}*theta4+p{j,5}*theta5+ p{j,6}*theta6)
 -
(p{j,5+2}*p{j,1}+p{j,5+2}*p{j,2}+p{j,5+2}*p{j,3}+p{j,5+2}*p{j,4}+p{j,5+2}*p{j,5}+p{j,5+2
}*p{j,6} ))));
     else
        xx=1/(1+exp(-&d*(( p{j,1}*theta1 +p{j,2}*theta2+ p{j,3}*theta3 +
p{j,4}*theta4+p{j,5}*theta5 + p{j,6}*theta6)  - ( p{j,5+resp+1}*p{j,1} + p{j,5+resp+1}*p{j,2}
+ p{j,5+resp+1}*p{j,3} +p{j,5+resp+1}*p{j,4} +p{j,5+resp+1}*p{j,5}
+p{j,5+resp+1}*p{j,6}))))) -1/(1+exp(-&d*(( p{j,1}*theta1 +p{j,2}*theta2+ p{j,3}*theta3 +
p{j,4}*theta4+p{j,5}*theta5 + p{j,6}*theta6)  - ( p{j,5+resp+2}*p{j,1} + p{j,5+resp+2}*p{j,2}
+ p{j,5+resp+2}*p{j,3} +p{j,5+resp+2}*p{j,4} +p{j,5+resp+2}*p{j,5}
+p{j,5+resp+2}*p{j,6}))));
    end;
  end;
%mend mgr;
/* Merge item model and number of categories information */
data misc;
        set model_type;
        set ncat_info;
run;
data parms;
        one=1;
        /*  set pointer to one */
        merge thetas;
        set item_par point=one;
        set misc point=one;
run;
%macro gendata;
%do l= 1 %to &nreps;
        data seed;
        seed=%eval(&seed);
data gen_data (keep=ix1-ix&nitems theta1-theta&ndims obs) seeds (keep=seed);
        set parms;
```

```
        array theta{&ndims} theta1-theta&ndims;
        array p{&nitems,&nparmdim} p1-p%eval(&nitems * &nparmdim);
        array ix{&nitems} ix1-ix&nitems;
        array ncat{&nitems} ncat1-ncat&nitems;
        array model{&nitems} model1-model&nitems;
        obs=&l;
        array cumprob{*} cumprob1-cumprob%eval(&maxcat);
        do j=1 to &nitems;
        do k=1 to &maxcat;
          cumprob[k]=.;
        end;
        do resp=0 to ncat{j}-1;
                                    if model[j]="mgr" then do;
                                            %mgr;
                                    end;
           if resp=0 then cumprob{1}=xx;
           else cumprob{resp+1}=xx+cumprob{resp};
        end;
        call ranuni(seed,r01);
       output seeds;
        do resp=1 to ncat{j}-1;
          if resp=1 and r01<cumprob{resp} then
             ix{j}=0;
          else if r01>cumprob{resp} and r01<cumprob{resp+1} then
             ix{j}=resp;
        end;
    end;
         file wrkdir(&outfile);
    put (theta1-theta&ndims) (5.2 +1) +1 (ix1-ix&nitems) (1.);
         file wrkdir(&respfile);
         put (ix1-ix&nitems) (1.);
         output seeds;
%end;
%mend gendata;
%gendata
quit;
```

APPENDIX B

SAS Item Fit Program To Calculate the Pseudocounts-Based Item Fit Statistics

```
/*IRTFIT_RESAMPLE SAS program - This program calculates a goodness-of-fit
        statiistic based on posterior expectations and then uses resampling
        techniques for hypothesis testing.  This program was developed
        under a grant from the U.S. Department of Education, National
        Assessment of Education Progress: Secondary Analysis Program (Grant
        #R902B970008). Relevant references are:

        Stone, C.A. (2000).  Monte-carlo based null distribution for an
          alternative fit statistic.  Journal of Educational Measurement,
          37, 58-75.

        Stone, C.A., & Hansen, M.A.  A computer program for assessing
          goodness-of-fit of item response theory models to NAEP data.
          Final report prepared for the Department of Education, NAEP
          Secondary Analysis Program (PR/Award Number R902B70008).
        Date: 6/1/2000
        Author Contact: Clement A. Stone (cas@pitt.edu)
        */
        ************************************************************;
        * Control Settings for the Program - User Defined;
        ************************************************************;
        * Test control settings;
        %let nitems=12; /* number of items to analyze */
        %let nparm=2;   /* maximum number of item parameters across item set */
                /* # of est parameters per item for adjusting df*/
        %let maxcat=2;  /* max number of response categories across item set */
                /* number of response categories per item */
        %let itemcats=2,2,2,2,2,2,2,2,2,2,2,2;
        %let missing=9; /* missing value code if any */
        * IRT model control settings;
         /* IRT model for each item delimited by commas - if only one listed, model applies to all
items. Programmed models: gr=graded as estimated in multilog, di=dichotomous,
        pc=partial credit, gpc=generalized pc as in parscale */
        %let models=gr,gr,gr,gr,gr,gr,gr,gr,gr,gr,gr,gr;
        %let d=1;     /*d scaling parameter for probability functions */
        * Item Fit Calculation control settings;
        %let nqpt=11;       /* number of discrete levels used for ability dist. */
        %let ub=2;          /* upper boundary for continuous ability distribution */
        %let sigma=1;       /* assummed var of posterior ability distribution */
        %let mean=0;        /* assummed mean of posterior ability distribution */
        %let resample=0;    /* 0=no, 1=yes */
        %let nreps=1;      /* number of MC resamples - 100 or > recommended */
        %let critcut=.02;   /* cutoff for expected prob - cells below cutoff excluded */
        %let abilcut=2.02;  /* upper value for theta used to calculate fit */
        %let addcell=.000001;/* constant added to pseudo-observed cells with 0 count */
        %let seed=0;        /* use computer clock to initialize seed for resampling */
```

```
        * Misc control settings;
        libname dir 'dir';
        %let graphs=0;    /* 0=don't produce response probability graphs, 1=graph */
        %let filename='filename';
        %let varfile='varfile';;
        %let statfile ='statfile';
        %let nsubj=0;     /* input n size to override automatic calculation */
*****************************************************************;
        * READ ITEM Parameters * must be defined by User;
*****************************************************************;
        data item_par;
          infile 'paramfile' missover;
          input   x1 -x&nparm;
          array y{*} x1-x&nparm;
        *if reading in parameters for the 1P or 2P models - uncomment next line;
        *  x3=0;
        if _n_=(&nitems+1) then delete;
           keep p;
           do j=1 to &nparm;
             p=y{j};
             output;
           end;
        proc transpose out=item_par prefix=p;
        var p;
        run;
        ****************************************************************;
        *READ Item Response Data  * must be defined by User ;
        ****************************************************************;
        data item_res;
          infile 'datafile' missover end=lastobs;
          input (ix1-ix&nitems) (1.);
          array rs{*} ix1-ix&nitems;
        * data needs to range from 0 to highest category - below changes range from 1 to
          high category to 0 to high category - uncomment if desired;
        *  do i=1 to &nitems;
        *     rs[i]=rs[i]-1;
        *  end;* check for amount of missing data;
          nmiss=0;
          do i=1 to &nitems;
             if rs[i] = &missing then nmiss=nmiss+1;
          end;
          retain nobs (0);
          if nmiss = &nitems then delete;
          else nobs=nobs+1;
          if (lastobs=1) and (&nsubj = 0) then call symput('nsubj',left(nobs));
        run;
```

```
****Graded Response Model*********;
* As implemented in MULTILOG;
%macro gr(theta);
  do;
    if resp =  (ncat[j] - 1) then
       xx=1/(1+exp(-&d*p{j,1}*(&theta-p{j,resp+1})));
    else if resp=0 then
       xx=1-1/(1+exp(-&d*p{j,1}*(&theta-p{j,2})));
    else
       xx=1/(1+exp(-&d*p{j,1}*(&theta-p{j,resp+1})))
          -1/(1+exp(-&d*p{j,1}*(&theta-p{j,resp+2})));
  end;
%mend gr;
*****************************************************************;
********     Do not make changes below this point      *********;
*****************************************************************;
* These 2 MACROS used to construct calls to probability functions
  for mixed item type tests;
%macro words(string);
  %local count word;
  %let count=1;
  %let word=%qscan(%quote(&string),&count,%str( )%str(,));
  %do %while(&word ne);
    %let count=%eval(&count+1);
    %let word=%qscan(%quote(&string),&count,%str( )%str(,));
  %end;
  %eval(&count-1)
%mend words;
%macro probcall;
 %if %words(%quote(&models))=1 %then
     %unquote(%&models.(theta));
 %else %do;
   %do i = 1 %to %words(%quote(&models));
      %let x&i = %scan(%quote(&models),&i,%str(,));
   %end;
   %let xfinal=;
   %do i = 1 %to %words(%quote(&models));
      %do j = 1 %to %words(%quote(&models));
        %if &i ne &j %then %do;
        %if &&x&i eq &&x&j %then %let x&i=;
      %end;
   %end;
   %let xfinal=&&x&i &xfinal;
   %end;
   %put These are the unique model values --> &xfinal;
   %do i = 1 %to %words(%quote(&xfinal));
```

```
      %let x&i = %scan(%quote(&xfinal),&i,%str( ));
      %let tmp=&&x&i;
      %if &i=1 %then
       if model[j]="&&x&i" then %unquote(%&tmp.(theta));
      %else
       else if model[j]="&&x&i" then %unquote(%&tmp.(theta));
  %end;
 %end;
%mend;
*****************************************************************;
*  Compute quadrature points and weights for quadrature points,
  set up number of categories and model arrays;
*****************************************************************;
data weights;
array qpt{&nqpt};
array prior{&nqpt};
xinc=(2*&ub)/(&nqpt-1);
do j=1 to &nqpt;
      qpt{j}=-&ub+(j-1)*xinc;
end;
var=-2*&sigma**2;
total=0;
do j=1 to &nqpt;
        prior{j}=exp((qpt{j}-&mean)**2/var);
        total=total+prior{j};
end;
do j=1 to &nqpt;
        prior{j}=prior{j}/total;
end;
array ncat{&nitems};
do j=1 to &nitems;
    ncat{j} = scan("&itemcats",j);
end;
length model1-model&nitems $3.;
array model {&nitems};
do j=1 to &nitems;
   model[j] = scan("&models",j);
end;
run;
data seed;
seed=%eval(&seed);
*****************************************************************;
*Merge data sets weights, item_par, and item_res and calculate
 posterior expectations and likelihood based on prior weights;
*****************************************************************;
%global wghts;
```

```
%let wghts=prior;
%let l=1;
%let nloops=1;
data item_res(keep=ix1-ix&nitems lk1-lk%eval(&nqpt+1))
    postexp(keep=postn1-postn&nqpt ptr1-ptr%eval(&nitems*&nqpt*&maxcat))
    weights(keep=qpt1-qpt&nqpt prior1-prior&nqpt awghts1-awghts&nqpt
          ncat1-ncat&nitems model1-model&nitems);
 if _n_=1 then do;
    set item_par;
    set weights;
 end;
 set item_res end=lastobs;
 array qpt{&nqpt} qpt1-qpt&nqpt;
 array prior{&nqpt} prior1-prior&nqpt;
 array p{&nitems,&nparm} p1-p%eval(&nitems * &nparm);
 array postn{&nqpt};
 array ptr{&nitems, &nqpt, %eval(&maxcat)};
 array lk{%eval(&nqpt + 1)};
 array ix{&nitems} ix1-ix&nitems;
 array awghts{&nqpt};
 array ncat{&nitems};
 array model {&nitems};
 retain postn1-postn&nqpt 0;
 retain ptr1-ptr%eval(&nitems*&nqpt*&maxcat) 0;
 retain loglike (0);
* compute likelihood of response pattern;
 lk{  %eval(&nqpt + 1)}=0;
 do i=1 to &nqpt;
      lk{ i}=1;
      do j=1 to &nitems;
        resp=ix{j};
        if resp ~= &missing then
          do;
            theta=qpt[i];
            %probcall
            lk{i}=lk{i}*xx;
          end;
       end;
     lk{%eval(&nqpt+1)}=lk{%eval(&nqpt+1)}+lk{i}*&wghts{i};
 end;
 loglike=loglike+log(lk{ %eval(&nqpt + 1)});
* compute posterior expectations;
 do i=1 to &nqpt;
     xx=lk(i)*&wghts(i)/lk( %eval(&nqpt + 1));
     postn{i}=postn{i}+xx;
     do j=1 to &nitems;
```

```
        do k=1 to ncat[j];
          if ix{j}=k-1 then
            ptr{j,i,k}=ptr{j,i,k}+xx;
        end;
      end;
    end;
* compute posterior or empirical weights;
    if (lastobs=1) then
      do;
        if &l=1 then
          do;
            asum=0;
            do i=1 to &nqpt;
              asum=asum+postn{i};
            end;
            do i=1 to &nqpt;
              awghts{i}=postn{i}/asum;
            end;
          end;
        output weights;
        output postexp;
        output item_res;
      end;
    else output item_res;
    if (&l = &nloops) and (lastobs=1) then
      do;
        file "&filename";
        put ' ';
        put ' Program IRTFIT_RESAMPLE v1.0: ';
        put '   Program to compute goodness-of-fit statistics ';
        put '   based on posterior expectations.  Resampling-based';
        put '   Monte Carlo methods used for hypothesis testing.';
        put ' ';
        put ' Summary of Posterior Ability Distribution and Likelihood Statistics';
        put ' ';
        twolog=-2*loglike;
        put +1 '-2*loglikelihood is: ' twolog 8.2;
      end;
run;
*****************************************************************;
*Compute distribution statistics for posterior ability distribution;
*****************************************************************;
data item_res(keep=ix1-ix&nitems lk1-lk%eval(&nqpt+1) eapmean eapvar)
    results(keep=nitems nqpt nsubj mmean var);
  if _n_=1 then do;
    set item_par;
```

```
    set weights;
    set postexp;
  end;
  set item_res end=lastobs;
  array qpt{&nqpt} qpt1-qpt&nqpt;
  array prior{&nqpt} prior1-prior&nqpt;
  array p{&nitems,&nparm} p1-p%eval(&nitems * &nparm);
  array postn{&nqpt};
  array ptr{&nitems, &nqpt, %eval(&maxcat)};
  array lk{%eval(&nqpt + 1)};
  array ix{&nitems} ix1-ix&nitems;
  array awghts{&nqpt};
  array prob{&nqpt};
  array ncat{&nitems};
* compute EAP - mean and variance;
  retain eapmean (0) eapvar (0);
  zz=0;
  z=0;
  xx=0;
  yy=0;
  ssum=0;
  do i=1 to &nqpt;
      xx=xx+lk{i}*qpt{i}*&wghts{i};
      yy=lk{i}*&wghts{i};
      ssum=ssum+yy;
  end;
  eapmean=eapmean+(xx/ssum);
  do i=1 to &nqpt;
      z=z+lk{i}*&wghts{i}*(qpt{i}-(xx/ssum))**2;
  end;
  eapvar=eapvar+(z/ssum);
  output item_res;
 if lastobs=1 then do;
   file "&filename" mod;
   mmean=eapmean/&nsubj;
   var=eapvar/&nsubj;
   nitems=&nitems;
   nqpt=&nqpt;
   nsubj=&nsubj;
   put ' ';
   put '     Group:              Marginal                Weights';
   put '                       Expected N           Prior    Posterior';
   do i=1 to &nqpt;
     put   'qpt{' i 2. '}=' qpt{i} 6.2 '       ' +5 postn{i} 8.2
          +13 prior{i}  8.4 +3 awghts{i} 8.4;
   end;
```

```
  put ' ';
  put 'Mean of Posterior Means and Variances:  ' mmean 8.6 '   ' var 8.6 ;
  put ' ';
  if &nloops=1 then put / "Item Fit results based on Prior Weights:";
  else put / "Results are based on Posterior Weights:";
  put ' ';
  file "&varfile" mod;
  put mmean 8.6 '   ' var 8.6;
   file "&filename" mod;
  output results;
end;
*****************************************************************;
* Compute Item Fit tables with Pearson and LR chi-squared
  statistics.  Chi-Square statistics are adjusted for cells that
  do not meet criteria by substituting an average deviation
  across valid cells;
*****************************************************************;
data results(keep=nitems nqpt nsubj mmean var vlike1-vlike&nitems
                  vchisq1-vchisq&nitems vdf1-vdf&nitems)
       probs(keep=ptrx1-ptrx%eval(&nitems*&nqpt*&maxcat)
                  exx1-exx%eval(&nitems*&nqpt*&maxcat)
                  qpt1-qpt&nqpt ncat1-ncat&nitems item1-item&nitems);
  set item_par;
  set weights;
  set postexp;
  set results;
  file "&filename" mod;
 array qpt{&nqpt} qpt1-qpt&nqpt;
 array prior{&nqpt} prior1-prior&nqpt;
 array p{&nitems,&nparm} p1-p%eval(&nitems * &nparm);
 array postn{&nqpt};
 array ptr{&nitems, &nqpt, &maxcat};
 array awghts{&nqpt};
 array ncat{&nitems};
 array sumcol{%eval(&maxcat+1)};
 array res{&nqpt,&maxcat};
 array ex{&maxcat};
 array vlike{&nitems};
 array vchisq{&nitems};
 array vdf{&nitems};
 array exx{&nitems, &nqpt, &maxcat};
 array ptrx{&nitems, &nqpt, &maxcat};
 array item{&nitems} item1-item&nitems;
 array model {&nitems};
 do j=1 to &nitems;
    put  'ITEM' j 3. /  '  Parameters are: ' @@;
```

```
    do kkk = 1 to &nparm;
      put p{j,kkk} 7.3 @@;
    end;
    put ' ';
    put ' ';
    put +1 'TABLE:         '@@;
    do kkk=0 to ncat[j]-1;
      put '     'kkk 1.0 '  ' @@;
    end;
    put   '    ROW F';
    put ' ';
* obtain marginals for pseudo-observed distribution;
    qpt_tot=0;
    like=0;
    chisq=0;
    df=0;
    addcell=0;
    do i=1 to &maxcat+1;
      sumcol{i}=0;
    end;
    do i=1 to &nqpt;
      sumobs=0;
      do k=1 to ncat[j];
        sumobs=sumobs+ptr{j,i,k};
        sumcol{k}=sumcol{k}+ptr{j,i,k};
      end;
      sumcol{ncat[j]+1}=sumcol{ncat[j]+1}+sumobs;
    end;
* obtain expected score distribution;
    do i=1 to &nqpt;
      sumobs=0;
      do resp=0 to ncat[j]-1;
         res{i,resp+1}=.;
         theta=qpt[i];
         %probcall
         ex(resp+1)=xx;
         exx(j,i,resp+1)=xx;
         sumobs=sumobs+ptr{j,i,resp+1};
      end;
      do resp=1 to ncat[j];
        if sumobs=0 then ptrx(j,i,resp)=0;
        else ptrx(j,i,resp)=ptr(j,i,resp)/sumobs;
      end;
* check cell count criteria and accumulate deviations;
      use=1;
      if abs(qpt{i}) > &abilcut then use=0;
```

```
      if use=1 then qpt_tot=qpt_tot+1;
      do k=1 to ncat[j];
        ex{k}=ex{k}*sumobs;
      end;
      if use=1 and &nsubj>1 then
        do k=1 to ncat[j];
          if ptr{j,i,k} = 0 then do;
            addcell=addcell+1;
            ptr{j,i,k} = &addcell;
          end;
          if ex{k} >= &critcut and ptr{j,i,k}>0.0 then
            do;
              like=like+ptr{j,i,k}*log(ptr{j,i,k}/ex{k});
              res{i,k}=(ptr{j,i,k}-ex{k}) / sqrt(ex{k});
              chisq=chisq+(ptr{j,i,k}-ex{k})**2/ex{k};
              df=df+1;
            end;
        end;
* write out cell contents;
      put +2 'QPT' I 3. ' PSEUDO# '  @@;
      do kkk=1 to ncat[j];
            put ptr{j,i,kkk} 8.2 @@;
      end;
      put sumobs 8.2;

      put  qpt{i} 6.2  '       EXP'  @@;
      do kkk=1 to ncat[j];
        if kkk=ncat[j] then
          put ex{kkk} 8.2;
        else
          put ex{kkk} 8.2 @@;
      end;
      put +10 'STD RES' @@;
      do kkk=1 to ncat[j];
        if kkk=ncat[j] then
            put res{i,kkk} 8.2;
        else
            put res{i,kkk} 8.2 @@;
      end;
    end;
    put ' ';
    put +12 'COL F' @@;
    do kkk=1 to ncat[j]+1;
            put sumcol{kkk} 8.2 @@;
    end;
    put ' ';
```

```
        * adjust chi-square statistics if necessary;
            like=like*2;
            adjust=qpt_tot * ncat[j] / df;
            vlike[j]=like*adjust;
            vchisq[j]=chisq*adjust;
            vdf[j]=adjust;
            ncells=qpt_tot * ncat[j];
            put / +2 'LIKELIHOOD RATIO CHI SQ=' @25 like 8.2 @35 '   Adjusted L2=' @60
vlike[j] 8.2;
            put +2 'PEARSON CHI SQ =' @25 chisq 8.2 @35 '   Adjusted X2=' @60 vchisq[j]
8.2;
            put '  # of cells used to calculate fit = ' df 4. ' of ' ncells 4. ', # of Zero count cells = '
addcell 4.;
            put ' ';
          end;  output probs;  output results;
        ************************************************************;
        * Save probabilities from item fit tables - to be used for
          graphing item response category functions for observed and
          expected probabilities;
        ************************************************************;
        data itemdat(keep=prob type theta resp item );
        set probs;
        array ptrx{&nitems,&nqpt,&maxcat};
        array exx{&nitems,&nqpt,&maxcat};
        array qpt{*} qpt1-qpt&nqpt;
        array ncat{*} ncat1-ncat&nitems;
        array ptr{&nitems,&nqpt,&maxcat};
        array ex{&nitems,&nqpt,&maxcat};
        length type $3;
        label type='Response Prob';
        label resp='Response Category';
        do i=1 to &nitems;
          do j=1 to ncat[i];
            do k=1 to &nqpt;
              theta=qpt[k];
              resp=j;
              item=i;
              prob=ptrx[i,k,j];
              type='Obs';
              output itemdat;
              prob=exx[i,k,j];
              type='Exp';
              output itemdat; end; end;run;
```

APPENDIX C

Data Obtained From the Simulated χ^{2*} and G^{2*} Fit Statistic Distributions

Appendix C. Final Data Set From the Simulated χ^{2*} and G^{2*} Fit Statistic Distributions

								χ^{2*}		G^{2*}	
ID	Test	Item	a	Info	$\bar{p}$	$\sqrt{\bar{p}}$	*ncat*	γ	ν	γ	ν
1	2cat12a	1a-2Cat	0.7	0.5	0.322	0.57	2	0.25	4.89	0.24	5.28
2	2cat12a	2a-2Cat	1.0	0.7	0.322	0.57	2	0.29	5.67	0.23	6.98
3	2cat12a	3a-2Cat	1.4	1.1	0.322	0.57	2	0.28	8.16	0.15	14.45
4	2cat12a	4a-2Cat	1.7	1.4	0.322	0.57	2	0.27	10.59	0.15	17.95
5	2cat12a	5a-2Cat	2.1	1.9	0.322	0.57	2	0.24	14.54	0.16	21.61
6	2cat12a	6a-2Cat	2.4	2.2	0.322	0.57	2	0.22	17.72	0.17	22.97
7	2cat12a	7a-2Cat	0.7	0.5	0.322	0.57	2	0.24	4.70	0.24	4.92
8	2cat12a	8a-2Cat	1.0	0.9	0.322	0.57	2	0.21	6.32	0.19	7.31
9	2cat12a	9a-2Cat	1.4	1.3	0.322	0.57	2	0.19	7.69	0.14	10.32
10	2cat12a	10a-2Cat	1.7	1.6	0.322	0.57	2	0.24	6.26	0.14	10.26
11	2cat12a	11a-2Cat	2.1	2.1	0.322	0.57	2	0.16	8.38	0.10	13.53
12	2cat12a	12a-2Cat	2.4	2.4	0.322	0.57	2	0.13	9.69	0.08	16.29
13	2cat12b	1b-2Cat	0.7	0.5	0.162	0.40	2	0.27	5.11	0.28	5.02
14	2cat12b	2b-2Cat	1.0	0.9	0.162	0.40	2	0.26	5.21	0.26	5.22
15	2cat12b	3b-2Cat	1.4	1.4	0.162	0.40	2	0.25	4.86	0.25	5.03
16	2cat12b	4b-2Cat	1.7	1.7	0.162	0.40	2	0.20	5.12	0.21	5.11
17	2cat12b	5b-2Cat	2.1	2.1	0.162	0.40	2	0.17	5.02	0.18	5.11
18	2cat12b	6b-2Cat	2.4	2.4	0.162	0.40	2	0.17	4.45	0.18	4.44
19	2cat12b	7b-2Cat	0.7	0.5	0.162	0.40	2	0.30	4.51	0.31	4.46
20	2cat12b	8b-2Cat	1.0	0.9	0.162	0.40	2	0.29	4.78	0.29	4.79
21	2cat12b	9b-2Cat	1.4	1.4	0.162	0.40	2	0.21	5.55	0.22	5.51
22	2cat12b	10b-2Cat	1.7	1.7	0.162	0.40	2	0.23	4.54	0.24	4.62
23	2cat12b	11b-2Cat	2.1	2.1	0.162	0.40	2	0.23	4.13	0.22	4.44
24	2cat12b	12b-2Cat	2.4	2.4	0.162	0.40	2	0.16	4.75	0.16	4.94
25	2cat12c	1c-2Cat	0.7	0.5	0.320	0.57	2	0.18	6.18	0.19	6.28
26	2cat12c	2c-2Cat	1.0	0.9	0.320	0.57	2	0.23	5.74	0.21	6.51
27	2cat12c	3c-2Cat	1.4	1.3	0.320	0.57	2	0.20	7.49	0.14	10.27
28	2cat12c	4c-2Cat	1.7	1.6	0.320	0.57	2	0.25	6.05	0.15	10.14
29	2cat12c	5c-2Cat	2.1	2.1	0.320	0.57	2	0.21	6.49	0.11	12.33
30	2cat12c	6c-2Cat	2.4	2.4	0.320	0.57	2	0.11	11.09	0.09	14.45
31	2cat12c	7c-2Cat	0.7	0.5	0.320	0.57	2	0.28	4.52	0.25	4.96

32	2cat12c	8c-2Cat	1.0	0.7	0.320	0.57	2	0.29	5.61	0.23	7.03
33	2cat12c	9c-2Cat	1.4	1.1	0.320	0.57	2	0.23	9.53	0.15	14.23
34	2cat12c	10c-2Cat	1.7	1.4	0.320	0.57	2	0.28	10.05	0.16	17.25
35	2cat12c	11c-2Cat	2.1	1.9	0.320	0.57	2	0.18	18.89	0.13	24.83
36	2cat12c	12c-2Cat	2.4	2.2	0.320	0.57	2	0.20	18.68	0.16	23.85
37	2cat24	1a-2Cat	0.7	0.5	0.119	0.34	2	0.47	5.14	0.45	5.40
38	2cat24	2a-2Cat	1.0	0.7	0.119	0.34	2	0.48	5.76	0.38	7.24
39	2cat24	3a-2Cat	1.4	1.1	0.119	0.34	2	0.53	6.16	0.35	9.11
40	2cat24	4a-2Cat	1.7	1.4	0.119	0.34	2	0.50	7.66	0.27	13.46
41	2cat24	5a-2Cat	2.1	1.9	0.119	0.34	2	0.48	8.88	0.25	16.54
42	2cat24	6a-2Cat	2.4	2.2	0.119	0.34	2	0.41	11.30	0.23	19.35
43	2cat24	7a-2Cat	0.7	0.5	0.119	0.34	2	0.40	5.59	0.40	5.67
44	2cat24	8a-2Cat	1.0	0.9	0.119	0.34	2	0.46	5.40	0.43	5.81
45	2cat24	9a-2Cat	1.4	1.3	0.119	0.34	2	0.44	5.81	0.38	6.88
46	2cat24	10a-2Cat	1.7	1.6	0.119	0.34	2	0.40	6.16	0.35	7.34
47	2cat24	11a-2Cat	2.1	2.1	0.119	0.34	2	0.44	5.21	0.29	7.83
48	2cat24	12a-2Cat	2.4	2.4	0.119	0.34	2	0.49	4.36	0.31	6.88
49	2cat24	1b-2Cat	0.7	0.5	0.119	0.34	2	0.40	5.17	0.43	5.02
50	2cat24	2b-2Cat	1.0	0.9	0.119	0.34	2	0.42	5.00	0.44	4.99
51	2cat24	3b-2Cat	1.4	1.4	0.119	0.34	2	0.43	4.49	0.45	4.61
52	2cat24	4b-2Cat	1.7	1.7	0.119	0.34	2	0.35	4.89	0.36	5.00
53	2cat24	5b-2Cat	2.1	2.1	0.119	0.34	2	0.36	4.02	0.30	5.14
54	2cat24	6b-2Cat	2.4	2.4	0.119	0.34	2	0.39	3.53	0.28	5.11
55	2cat24	7b-2Cat	0.7	0.5	0.119	0.34	2	0.34	6.33	0.36	6.16
56	2cat24	8b-2Cat	1.0	0.9	0.119	0.34	2	0.36	5.77	0.39	5.67
57	2cat24	9b-2Cat	1.4	1.4	0.119	0.34	2	0.40	4.77	0.41	4.91
58	2cat24	10b-2Cat	1.7	1.7	0.119	0.34	2	0.37	4.82	0.38	4.99
59	2cat24	11b-2Cat	2.1	2.1	0.119	0.34	2	0.43	3.52	0.35	4.43
60	2cat24	12b-2Cat	2.4	2.4	0.119	0.34	2	0.43	3.11	0.31	4.44
61	2cat36	1a-2Cat	0.7	0.5	0.094	0.31	2	0.58	5.16	0.52	5.90
62	2cat36	2a-2Cat	1.0	0.7	0.094	0.31	2	0.65	5.13	0.51	6.65
63	2cat36	3a-2Cat	1.4	1.1	0.094	0.31	2	0.87	4.62	0.42	9.13
64	2cat36	4a-2Cat	1.7	1.4	0.094	0.31	2	0.90	5.10	0.43	9.99

65	2cat36	5a-2Cat	2.1	1.9	0.094	0.31	2	0.66	7.72	0.40	12.25
66	2cat36	6a-2Cat	2.4	2.2	0.094	0.31	2	0.62	8.55	0.35	14.66
67	2cat36	7a-2Cat	0.7	0.5	0.094	0.31	2	0.49	5.93	0.49	6.09
68	2cat36	8a-2Cat	1.0	0.9	0.094	0.31	2	0.51	6.11	0.49	6.53
69	2cat36	9a-2Cat	1.4	1.3	0.094	0.31	2	0.53	5.92	0.47	6.94
70	2cat36	10a-2Cat	1.7	1.6	0.094	0.31	2	0.73	4.63	0.51	6.59
71	2cat36	11a-2Cat	2.1	2.1	0.094	0.31	2	0.90	3.33	0.45	6.58
72	2cat36	12a-2Cat	2.4	2.4	0.094	0.31	2	1.09	2.51	0.44	6.08
73	2cat36	1b-2Cat	0.7	0.5	0.094	0.31	2	0.52	5.57	0.56	5.38
74	2cat36	2b-2Cat	1.0	0.9	0.094	0.31	2	0.51	5.46	0.54	5.41
75	2cat36	3b-2Cat	1.4	1.4	0.094	0.31	2	0.41	6.05	0.44	6.15
76	2cat36	4b-2Cat	1.7	1.7	0.094	0.31	2	0.41	5.55	0.40	6.09
77	2cat36	5b-2Cat	2.1	2.1	0.094	0.31	2	0.46	4.49	0.38	5.79
78	2cat36	6b-2Cat	2.4	2.4	0.094	0.31	2	0.53	3.66	0.36	5.57
79	2cat36	7b-2Cat	0.7	0.5	0.094	0.31	2	0.48	5.85	0.52	5.68
80	2cat36	8b-2Cat	1.0	0.9	0.094	0.31	2	0.51	5.45	0.55	5.34
81	2cat36	9b-2Cat	1.4	1.4	0.094	0.31	2	0.42	5.86	0.45	5.93
82	2cat36	10b-2Cat	1.7	1.7	0.094	0.31	2	0.49	4.94	0.44	5.77
83	2cat36	11b-2Cat	2.1	2.1	0.094	0.31	2	0.41	4.93	0.34	6.24
84	2cat36	12b-2Cat	2.4	2.4	0.094	0.31	2	0.52	3.62	0.36	5.47
85	2cat36	1c-2Cat	0.7	0.5	0.094	0.31	2	0.47	6.19	0.48	6.24
86	2cat36	2c-2Cat	1.0	0.9	0.094	0.31	2	0.47	6.20	0.45	6.69
87	2cat36	3c-2Cat	1.4	1.3	0.094	0.31	2	0.63	5.28	0.48	7.10
88	2cat36	4c-2Cat	1.7	1.6	0.094	0.31	2	0.59	5.27	0.43	7.36
89	2cat36	5c-2Cat	2.1	2.1	0.094	0.31	2	1.06	2.88	0.43	6.95
90	2cat36	6c-2Cat	2.4	2.4	0.094	0.31	2	0.86	3.10	0.39	6.79
91	2cat36	7c-2Cat	0.7	0.5	0.094	0.31	2	0.63	4.70	0.57	5.30
92	2cat36	8c-2Cat	1.0	0.7	0.094	0.31	2	0.53	6.22	0.44	7.54
93	2cat36	9c-2Cat	1.4	1.1	0.094	0.31	2	0.99	4.26	0.45	8.83
94	2cat36	10c-2Cat	1.7	1.4	0.094	0.31	2	0.87	5.28	0.42	10.20
95	2cat36	11c-2Cat	2.1	1.9	0.094	0.31	2	1.10	4.69	0.45	10.86
96	2cat36	12c-2Cat	2.4	2.2	0.094	0.31	2	0.69	7.71	0.37	13.95

97	3Cat12a	1a-3Cat	0.7	0.8	0.414	0.64	3	0.65	4.23	0.40	6.54
98	3Cat12a	2a-3Cat	1.0	1.3	0.414	0.64	3	0.57	5.70	0.31	9.54
99	3Cat12a	3a-3Cat	1.4	2.2	0.414	0.64	3	0.50	8.94	0.31	13.12
100	3Cat12a	4a-3Cat	1.7	2.8	0.414	0.64	3	0.49	11.69	0.31	16.68
101	3Cat12a	5a-3Cat	2.1	3.6	0.414	0.64	3	0.76	10.86	0.40	17.41
102	3Cat12a	6a-3Cat	2.4	4.2	0.414	0.64	3	1.20	8.64	0.49	17.22
103	3Cat12a	7a-3Cat	0.7	0.8	0.414	0.64	3	0.50	5.39	0.37	7.09
104	3Cat12a	8a-3Cat	1.0	1.3	0.414	0.64	3	0.62	5.31	0.38	7.93
105	3Cat12a	9a-3Cat	1.4	2.2	0.414	0.64	3	0.63	7.20	0.31	12.91
106	3Cat12a	10a-3Cat	1.7	2.8	0.414	0.64	3	0.58	9.92	0.35	14.79
107	3Cat12a	11a-3Cat	2.1	3.6	0.414	0.64	3	0.65	12.36	0.37	18.47
108	3Cat12a	12a-3Cat	2.4	4.2	0.414	0.64	3	0.95	10.99	0.46	18.66
109	3Cat12b	1b-3Cat	0.7	0.8	0.383	0.62	3	0.40	7.98	0.35	9.05
110	3Cat12b	2b-3Cat	1.0	1.3	0.383	0.62	3	0.41	9.48	0.32	11.80
111	3Cat12b	3b-3Cat	1.4	2.2	0.383	0.62	3	0.35	14.73	0.22	22.22
112	3Cat12b	4b-3Cat	1.7	2.8	0.383	0.62	3	0.51	12.39	0.26	23.35
113	3Cat12b	5b-3Cat	2.1	3.6	0.383	0.62	3	0.31	25.25	0.24	31.56
114	3Cat12b	6b-3Cat	2.4	4.2	0.383	0.62	3	0.29	29.41	0.24	34.73
115	3Cat12b	7b-3Cat	0.7	0.8	0.383	0.62	3	0.60	5.23	0.45	6.73
116	3Cat12b	8b-3Cat	1.0	1.4	0.383	0.62	3	0.66	5.80	0.38	9.40
117	3Cat12b	9b-3Cat	1.4	2.2	0.383	0.62	3	0.70	7.78	0.39	12.33
118	3Cat12b	10b-3Cat	1.7	2.9	0.383	0.62	3	0.76	9.87	0.46	14.01
119	3Cat12b	11b-3Cat	2.1	3.8	0.383	0.62	3	0.84	13.71	0.59	16.51
120	3Cat12b	12b-3Cat	2.4	4.4	0.383	0.62	3	1.18	13.36	0.77	16.99
121	3Cat12c	1c-3Cat	0.7	0.8	0.226	0.48	3	0.58	7.14	0.49	8.43
122	3Cat12c	2c-3Cat	1.0	1.4	0.226	0.48	3	0.55	8.31	0.38	11.61
123	3Cat12c	3c-3Cat	1.4	2.3	0.226	0.48	3	0.57	9.43	0.37	13.74
124	3Cat12c	4c-3Cat	1.7	3.0	0.226	0.48	3	0.59	10.62	0.38	15.10
125	3Cat12c	5c-3Cat	2.1	3.9	0.226	0.48	3	0.58	13.17	0.37	18.70
126	3Cat12c	6c-3Cat	2.4	4.6	0.226	0.48	3	0.69	13.40	0.43	18.86
127	3Cat12c	7c-3Cat	0.7	0.8	0.226	0.48	3	0.44	9.01	0.39	10.33

128	3Cat12c	8c-3Cat	1.0	1.4	0.226	0.48	3	0.60	7.72	0.43	10.45
129	3Cat12c	9c-3Cat	1.4	2.3	0.226	0.48	3	0.60	9.12	0.38	13.32
130	3Cat12c	10c-3Cat	1.7	3.0	0.226	0.48	3	0.66	9.69	0.41	14.47
131	3Cat12c	11c-3Cat	2.1	3.9	0.226	0.48	3	0.67	11.83	0.42	16.85
132	3Cat12c	12c-3Cat	2.4	4.6	0.226	0.48	3	0.73	12.92	0.47	17.60
133	3Cat24	1a-3Cat	0.7	0.8	0.267	0.52	3	0.80	5.51	0.54	7.98
134	3Cat24	2a-3Cat	1.0	1.3	0.267	0.52	3	1.03	4.86	0.55	8.33
135	3Cat24	3a-3Cat	1.4	2.2	0.267	0.52	3	1.14	5.45	0.52	10.57
136	3Cat24	4a-3Cat	1.7	2.8	0.267	0.52	3	0.97	7.43	0.49	12.96
137	3Cat24	5a-3Cat	2.1	3.6	0.267	0.52	3	0.80	12.10	0.48	17.46
138	3Cat24	6a-3Cat	2.4	4.2	0.267	0.52	3	0.99	12.01	0.54	18.31
139	3Cat24	7a-3Cat	0.7	0.8	0.267	0.52	3	0.75	5.74	0.54	7.81
140	3Cat24	8a-3Cat	1.0	1.3	0.267	0.52	3	1.14	4.33	0.52	8.69
141	3Cat24	9a-3Cat	1.4	2.2	0.267	0.52	3	1.14	5.45	0.55	9.95
142	3Cat24	10a-3Cat	1.7	2.8	0.267	0.52	3	0.88	8.41	0.50	13.11
143	3Cat24	11a-3Cat	2.1	3.6	0.267	0.52	3	0.70	13.58	0.44	18.63
144	3Cat24	12a-3Cat	2.4	4.2	0.267	0.52	3	0.86	13.62	0.48	20.59
145	3Cat24	1b-3Cat	0.7	0.8	0.267	0.52	3	0.56	8.02	0.50	9.06
146	3Cat24	2b-3Cat	1.0	1.3	0.267	0.52	3	0.62	8.46	0.44	11.69
147	3Cat24	3b-3Cat	1.4	2.2	0.267	0.52	3	0.69	9.82	0.41	15.49
148	3Cat24	4b-3Cat	1.7	2.8	0.267	0.52	3	0.72	10.63	0.39	18.56
149	3Cat24	5b-3Cat	2.1	3.6	0.267	0.52	3	0.68	13.65	0.39	22.45
150	3Cat24	6b-3Cat	2.4	4.2	0.267	0.52	3	0.47	21.28	0.38	25.75
151	3Cat24	7b-3Cat	0.7	0.8	0.267	0.52	3	0.72	6.02	0.52	8.06
152	3Cat24	8b-3Cat	1.0	1.4	0.267	0.52	3	1.03	5.06	0.57	8.36
153	3Cat24	9b-3Cat	1.4	2.2	0.267	0.52	3	1.05	6.41	0.48	12.11
154	3Cat24	10b-3Cat	1.7	2.9	0.267	0.52	3	0.84	10.42	0.53	14.37
155	3Cat24	11b-3Cat	2.1	3.8	0.267	0.52	3	0.89	14.08	0.65	16.53
156	3Cat24	12b-3Cat	2.4	4.4	0.267	0.52	3	1.16	14.62	0.77	18.57
157	3Cat36	1a-3Cat	0.7	0.8	0.145	0.38	3	0.90	6.94	0.60	10.02
158	3Cat36	2a-3Cat	1.0	1.3	0.145	0.38	3	0.97	6.92	0.54	11.57
159	3Cat36	3a-3Cat	1.4	2.2	0.145	0.38	3	1.29	6.28	0.64	11.24

160	3Cat36	4a-3Cat	1.7	2.8	0.145	0.38	3	1.25	7.45	0.69	11.90
161	3Cat36	5a-3Cat	2.1	3.6	0.145	0.38	3	1.03	11.02	0.64	15.29
162	3Cat36	6a-3Cat	2.4	4.2	0.145	0.38	3	1.11	11.90	0.63	17.60
163	3Cat36	7a-3Cat	0.7	0.8	0.145	0.38	3	0.80	7.88	0.55	11.16
164	3Cat36	8a-3Cat	1.0	1.3	0.145	0.38	3	1.15	5.78	0.56	11.15
165	3Cat36	9a-3Cat	1.4	2.2	0.145	0.38	3	1.58	5.09	0.63	11.29
166	3Cat36	10a-3Cat	1.7	2.8	0.145	0.38	3	1.11	8.19	0.62	13.10
167	3Cat36	11a-3Cat	2.1	3.6	0.145	0.38	3	0.93	11.94	0.59	16.40
168	3Cat36	12a-3Cat	2.4	4.2	0.145	0.38	3	1.02	13.06	0.59	19.06
169	3Cat36	1b-3Cat	0.7	0.8	0.145	0.38	3	0.69	9.19	0.59	11.00
170	3Cat36	2b-3Cat	1.0	1.3	0.145	0.38	3	0.75	9.69	0.57	12.73
171	3Cat36	3b-3Cat	1.4	2.2	0.145	0.38	3	0.91	9.56	0.57	14.60
172	3Cat36	4b-3Cat	1.7	2.8	0.145	0.38	3	1.04	9.33	0.53	17.14
173	3Cat36	5b-3Cat	2.1	3.6	0.145	0.38	3	0.83	13.29	0.53	19.53
174	3Cat36	6b-3Cat	2.4	4.2	0.145	0.38	3	0.74	16.20	0.51	22.41
175	3Cat36	7b-3Cat	0.7	0.8	0.145	0.38	3	0.84	7.56	0.62	10.18
176	3Cat36	8b-3Cat	1.0	1.4	0.145	0.38	3	1.02	6.95	0.61	10.80
177	3Cat36	9b-3Cat	1.4	2.2	0.145	0.38	3	1.80	4.79	0.66	11.30
178	3Cat36	10b-3Cat	1.7	2.9	0.145	0.38	3	1.09	9.46	0.61	14.78
179	3Cat36	11b-3Cat	2.1	3.8	0.145	0.38	3	1.02	14.12	0.67	18.46
180	3Cat36	12b-3Cat	2.4	4.4	0.145	0.38	3	1.21	15.24	0.78	19.91
181	3Cat36	1c-3Cat	0.7	0.8	0.145	0.38	3	0.61	10.15	0.53	11.94
182	3Cat36	2c-3Cat	1.0	1.4	0.145	0.38	3	0.82	8.27	0.60	10.99
183	3Cat36	3c-3Cat	1.4	2.3	0.145	0.38	3	1.04	7.39	0.60	11.99
184	3Cat36	4c-3Cat	1.7	3.0	0.145	0.38	3	1.29	6.82	0.76	10.69
185	3Cat36	5c-3Cat	2.1	3.9	0.145	0.38	3	1.27	7.86	0.69	12.94
186	3Cat36	6c-3Cat	2.4	4.6	0.145	0.38	3	1.16	10.06	0.75	13.79
187	3Cat36	7c-3Cat	0.7	0.8	0.145	0.38	3	0.74	8.74	0.62	10.51
188	3Cat36	8c-3Cat	1.0	1.4	0.145	0.38	3	1.09	6.40	0.63	10.69
189	3Cat36	9c-3Cat	1.4	2.3	0.145	0.38	3	1.16	6.83	0.62	11.90
190	3Cat36	10c-3Cat	1.7	3.0	0.145	0.38	3	1.09	7.93	0.63	12.46
191	3Cat36	11c-3Cat	2.1	3.9	0.145	0.38	3	1.14	8.78	0.70	12.94

192	3Cat36	12c-3Cat	2.4	4.6	0.145	0.38	3	1.05	10.87	0.70	14.48
193	4Cat12a	1a-4Cat	0.7	0.8	0.135	0.37	4	0.52	14.95	0.47	16.44
194	4Cat12a	2a-4Cat	1.0	1.6	0.135	0.37	4	0.70	11.70	0.53	15.30
195	4Cat12a	3a-4Cat	1.4	2.9	0.135	0.37	4	0.82	11.18	0.53	16.42
196	4Cat12a	4a-4Cat	1.7	3.9	0.135	0.37	4	0.76	13.45	0.53	17.84
197	4Cat12a	5a-4Cat	2.1	5.3	0.135	0.37	4	0.73	15.92	0.48	21.67
198	4Cat12a	6a-4Cat	2.4	6.3	0.135	0.37	4	0.75	17.62	0.51	22.46
199	4Cat12a	7a-4Cat	0.7	0.8	0.135	0.37	4	0.58	13.53	0.51	15.44
200	4Cat12a	8a-4Cat	1.0	1.6	0.135	0.37	4	0.84	9.96	0.49	16.40
201	4Cat12a	9a-4Cat	1.4	2.9	0.135	0.37	4	0.76	12.06	0.55	15.91
202	4Cat12a	10a-4Cat	1.7	3.9	0.135	0.37	4	0.72	13.90	0.48	19.07
203	4Cat12a	11a-4Cat	2.1	5.3	0.135	0.37	4	0.75	15.62	0.51	20.76
204	4Cat12a	12a-4Cat	2.4	6.3	0.135	0.37	4	0.79	16.73	0.50	22.76
205	4Cat12b	1b-4Cat	0.7	0.8	0.123	0.35	4	0.52	15.76	0.51	16.23
206	4Cat12b	2b-4Cat	1.0	1.6	0.123	0.35	4	0.57	15.08	0.53	16.51
207	4Cat12b	3b-4Cat	1.4	2.8	0.123	0.35	4	0.57	16.76	0.47	20.26
208	4Cat12b	4b-4Cat	1.7	3.8	0.123	0.35	4	0.66	15.51	0.51	19.84
209	4Cat12b	5b-4Cat	2.1	5.3	0.123	0.35	4	0.60	18.95	0.50	22.46
210	4Cat12b	6b-4Cat	2.4	6.4	0.123	0.35	4	0.57	21.01	0.48	24.65
211	4Cat12b	7b-4Cat	0.7	0.8	0.123	0.35	4	0.63	12.81	0.57	14.18
212	4Cat12b	8b-4Cat	1.0	1.6	0.123	0.35	4	0.58	14.74	0.50	16.55
213	4Cat12b	9b-4Cat	1.4	2.9	0.123	0.35	4	0.91	10.90	0.55	16.79
214	4Cat12b	10b-4Cat	1.7	4.0	0.123	0.35	4	0.92	12.68	0.62	17.24
215	4Cat12b	11b-4Cat	2.1	5.4	0.123	0.35	4	0.82	18.18	0.61	21.58
216	4Cat12b	12b-4Cat	2.4	6.5	0.123	0.35	4	0.84	21.46	0.61	25.44
217	4Cat12c	1c-4Cat	0.7	0.8	0.113	0.34	4	0.54	16.00	0.53	16.60
218	4Cat12c	2c-4Cat	1.0	1.5	0.113	0.34	4	0.67	13.53	0.56	16.28
219	4Cat12c	3c-4Cat	1.4	2.8	0.113	0.34	4	0.80	12.06	0.58	16.31
220	4Cat12c	4c-4Cat	1.7	3.8	0.113	0.34	4	0.77	13.16	0.56	17.48
221	4Cat12c	5c-4Cat	2.1	5.2	0.113	0.34	4	0.75	14.74	0.51	20.30
222	4Cat12c	6c-4Cat	2.4	6.3	0.113	0.34	4	0.74	16.48	0.50	22.12
223	4Cat12c	7c-4Cat	0.7	0.8	0.113	0.34	4	0.57	15.36	0.54	16.38

224	4Cat12c	8c-4Cat	1.0	1.5	0.113	0.34	4	0.58	15.67	0.48	19.06
225	4Cat12c	9c-4Cat	1.4	2.8	0.113	0.34	4	0.60	15.51	0.47	19.42
226	4Cat12c	10c-4Cat	1.7	3.8	0.113	0.34	4	0.71	14.54	0.53	18.74
227	4Cat12c	11c-4Cat	2.1	5.2	0.113	0.34	4	0.73	15.05	0.51	20.24
228	4Cat12c	12c-4Cat	2.4	6.3	0.113	0.34	4	0.76	16.11	0.53	20.88
229	4Cat24	1a-4Cat	0.7	0.8	0.069	0.26	4	0.75	15.98	0.63	19.04
230	4Cat24	2a-4Cat	1.0	1.6	0.069	0.26	4	0.92	13.64	0.61	20.06
231	4Cat24	3a-4Cat	1.4	2.9	0.069	0.26	4	1.23	11.05	0.72	17.70
232	4Cat24	4a-4Cat	1.7	3.9	0.069	0.26	4	1.01	14.40	0.63	21.44
233	4Cat24	5a-4Cat	2.1	5.3	0.069	0.26	4	0.99	16.02	0.65	21.99
234	4Cat24	6a-4Cat	2.4	6.3	0.069	0.26	4	1.10	15.54	0.65	23.04
235	4Cat24	7a-4Cat	0.7	0.8	0.069	0.26	4	0.81	14.86	0.63	18.87
236	4Cat24	8a-4Cat	1.0	1.6	0.069	0.26	4	0.91	13.60	0.61	19.86
237	4Cat24	9a-4Cat	1.4	2.9	0.069	0.26	4	0.88	15.36	0.60	21.64
238	4Cat24	10a-4Cat	1.7	3.9	0.069	0.26	4	1.23	11.60	0.70	19.00
239	4Cat24	11a-4Cat	2.1	5.3	0.069	0.26	4	1.03	15.31	0.59	23.88
240	4Cat24	12a-4Cat	2.4	6.3	0.069	0.26	4	1.13	15.20	0.71	21.38
241	4Cat24	1b-4Cat	0.7	0.8	0.069	0.26	4	0.63	18.94	0.62	19.69
242	4Cat24	2b-4Cat	1.0	1.6	0.069	0.26	4	0.63	20.05	0.58	22.12
243	4Cat24	3b-4Cat	1.4	2.8	0.069	0.26	4	0.91	15.12	0.55	24.43
244	4Cat24	4b-4Cat	1.7	3.8	0.069	0.26	4	0.95	15.40	0.62	22.92
245	4Cat24	5b-4Cat	2.1	5.3	0.069	0.26	4	0.94	16.80	0.66	23.29
246	4Cat24	6b-4Cat	2.4	6.4	0.069	0.26	4	0.86	18.92	0.65	24.48
247	4Cat24	7b-4Cat	0.7	0.8	0.069	0.26	4	0.76	15.75	0.67	17.79
248	4Cat24	8b-4Cat	1.0	1.6	0.069	0.26	4	0.81	15.37	0.59	20.63
249	4Cat24	9b-4Cat	1.4	2.9	0.069	0.26	4	1.27	10.98	0.65	19.83
250	4Cat24	10b-4Cat	1.7	4.0	0.069	0.26	4	1.07	14.45	0.69	20.69
251	4Cat24	11b-4Cat	2.1	5.4	0.069	0.26	4	1.12	17.16	0.75	22.86
252	4Cat24	12b-4Cat	2.4	6.5	0.069	0.26	4	1.13	19.50	0.79	24.17
253	4Cat36	1a-4Cat	0.7	0.8	0.045	0.21	4	0.74	20.33	0.62	24.59
254	4Cat36	2a-4Cat	1.0	1.6	0.045	0.21	4	0.91	16.86	0.63	24.07
255	4Cat36	3a-4Cat	1.4	2.9	0.045	0.21	4	1.05	15.98	0.65	24.61

256	4Cat36	4a-4Cat	1.7	3.9	0.045	0.21	4	1.15	15.42	0.65	25.17
257	4Cat36	5a-4Cat	2.1	5.3	0.045	0.21	4	1.11	17.19	0.67	26.14
258	4Cat36	6a-4Cat	2.4	6.3	0.045	0.21	4	1.24	16.82	0.73	25.30
259	4Cat36	7a-4Cat	0.7	0.8	0.045	0.21	4	0.75	20.07	0.66	23.34
260	4Cat36	8a-4Cat	1.0	1.6	0.045	0.21	4	0.91	17.51	0.58	26.68
261	4Cat36	9a-4Cat	1.4	2.9	0.045	0.21	4	1.22	13.98	0.70	23.08
262	4Cat36	10a-4Cat	1.7	3.9	0.045	0.21	4	1.11	15.72	0.67	24.55
263	4Cat36	11a-4Cat	2.1	5.3	0.045	0.21	4	1.31	15.00	0.70	25.02
264	4Cat36	12a-4Cat	2.4	6.3	0.045	0.21	4	1.24	16.64	0.76	23.94
265	4Cat36	1b-4Cat	0.7	0.8	0.045	0.21	4	0.72	20.97	0.70	22.05
266	4Cat36	2b-4Cat	1.0	1.6	0.045	0.21	4	0.77	20.41	0.68	23.53
267	4Cat36	3b-4Cat	1.4	2.8	0.045	0.21	4	0.88	19.44	0.64	26.14
268	4Cat36	4b-4Cat	1.7	3.8	0.045	0.21	4	1.33	13.71	0.67	25.99
269	4Cat36	5b-4Cat	2.1	5.3	0.045	0.21	4	1.10	17.32	0.75	24.73
270	4Cat36	6b-4Cat	2.4	6.4	0.045	0.21	4	0.98	19.73	0.71	26.82
271	4Cat36	7b-4Cat	0.7	0.8	0.045	0.21	4	0.75	20.12	0.63	24.18
272	4Cat36	8b-4Cat	1.0	1.6	0.045	0.21	4	0.93	17.04	0.66	23.40
273	4Cat36	9b-4Cat	1.4	2.9	0.045	0.21	4	1.30	13.35	0.74	21.96
274	4Cat36	10b-4Cat	1.7	4.0	0.045	0.21	4	1.12	16.70	0.69	24.97
275	4Cat36	11b-4Cat	2.1	5.4	0.045	0.21	4	1.28	17.44	0.77	25.99
276	4Cat36	12b-4Cat	2.4	6.5	0.045	0.21	4	1.34	19.09	0.85	26.24
277	4Cat36	1c-4Cat	0.7	0.8	0.045	0.21	4	0.63	23.83	0.61	25.47
278	4Cat36	2c-4Cat	1.0	1.5	0.045	0.21	4	0.79	19.96	0.64	24.81
279	4Cat36	3c-4Cat	1.4	2.8	0.045	0.21	4	0.93	17.44	0.64	25.14
280	4Cat36	4c-4Cat	1.7	3.8	0.045	0.21	4	1.09	15.45	0.63	25.74
281	4Cat36	5c-4Cat	2.1	5.2	0.045	0.21	4	1.54	11.83	0.66	25.66
282	4Cat36	6c-4Cat	2.4	6.3	0.045	0.21	4	1.28	14.95	0.73	24.12
283	4Cat36	7c-4Cat	0.7	0.8	0.045	0.21	4	0.66	23.09	0.64	24.22
284	4Cat36	8c-4Cat	1.0	1.5	0.045	0.21	4	0.76	20.64	0.64	24.57
285	4Cat36	9c-4Cat	1.4	2.8	0.045	0.21	4	1.07	15.46	0.67	24.35
286	4Cat36	10c-4Cat	1.7	3.8	0.045	0.21	4	0.99	17.07	0.68	24.19
287	4Cat36	11c-4Cat	2.1	5.2	0.045	0.21	4	1.29	14.21	0.77	22.33

288	4Cat36	12c-4Cat	2.4	6.3	0.045	0.21	4	1.29	14.98	0.69	25.34
289	5Cat12a	1a-5Cat	0.7	0.9	0.106	0.33	5	0.70	17.76	0.59	21.27
290	5Cat12a	2a-5Cat	1.0	1.7	0.106	0.33	5	0.77	17.07	0.58	22.14
291	5Cat12a	3a-5Cat	1.4	3.1	0.106	0.33	5	0.79	17.76	0.54	24.95
292	5Cat12a	4a-5Cat	1.7	4.4	0.106	0.33	5	0.87	17.33	0.58	24.58
293	5Cat12a	5a-5Cat	2.1	6.2	0.106	0.33	5	0.82	19.63	0.57	26.09
294	5Cat12a	6a-5Cat	2.4	7.6	0.106	0.33	5	0.80	22.06	0.55	29.00
295	5Cat12a	7a-5Cat	0.7	0.8	0.106	0.33	5	0.66	19.25	0.58	21.95
296	5Cat12a	8a-5Cat	1.0	1.6	0.106	0.33	5	0.87	14.79	0.53	24.09
297	5Cat12a	9a-5Cat	1.4	2.9	0.106	0.33	5	0.73	18.79	0.52	25.44
298	5Cat12a	10a-5Cat	1.7	4.0	0.106	0.33	5	0.88	16.68	0.57	24.36
299	5Cat12a	11a-5Cat	2.1	5.5	0.106	0.33	5	0.75	21.46	0.51	29.32
300	5Cat12a	12a-5Cat	2.4	6.8	0.106	0.33	5	0.85	20.70	0.59	26.83
301	5Cat12b	1b-5Cat	0.7	0.8	0.099	0.31	5	0.65	19.59	0.62	20.97
302	5Cat12b	2b-5Cat	1.0	1.6	0.099	0.31	5	0.61	21.86	0.55	24.42
303	5Cat12b	3b-5Cat	1.4	2.9	0.099	0.31	5	0.67	20.84	0.56	25.26
304	5Cat12b	4b-5Cat	1.7	4.1	0.099	0.31	5	0.87	16.76	0.56	25.42
305	5Cat12b	5b-5Cat	2.1	5.8	0.099	0.31	5	0.71	21.27	0.54	27.73
306	5Cat12b	6b-5Cat	2.4	7.1	0.099	0.31	5	0.66	23.33	0.52	29.22
307	5Cat12b	7b-5Cat	0.7	0.9	0.099	0.31	5	0.61	21.05	0.54	23.82
308	5Cat12b	8b-5Cat	1.0	1.7	0.099	0.31	5	0.74	17.96	0.56	23.46
309	5Cat12b	9b-5Cat	1.4	3.2	0.099	0.31	5	1.27	11.55	0.60	23.27
310	5Cat12b	10b-5Cat	1.7	4.5	0.099	0.31	5	0.71	22.24	0.52	28.84
311	5Cat12b	11b-5Cat	2.1	6.3	0.099	0.31	5	0.90	20.81	0.60	28.07
312	5Cat12b	12b-5Cat	2.4	7.8	0.099	0.31	5	0.89	23.69	0.62	30.09
313	5Cat12c	1c-5Cat	0.7	0.8	0.094	0.31	5	0.56	22.86	0.55	23.63
314	5Cat12c	2c-5Cat	1.0	1.6	0.094	0.31	5	0.61	21.60	0.57	23.60
315	5Cat12c	3c-5Cat	1.4	2.9	0.094	0.31	5	0.76	18.17	0.56	24.52
316	5Cat12c	4c-5Cat	1.7	4.0	0.094	0.31	5	0.81	17.46	0.60	23.05
317	5Cat12c	5c-5Cat	2.1	5.7	0.094	0.31	5	0.78	18.73	0.53	25.96
318	5Cat12c	6c-5Cat	2.4	7.0	0.094	0.31	5	0.77	19.93	0.54	26.55
319	5Cat12c	7c-5Cat	0.7	0.8	0.094	0.31	5	0.54	24.26	0.53	25.31

320	5Cat12c	8c-5Cat	1.0	1.6	0.094	0.31	5	0.57	23.25	0.51	26.39
321	5Cat12c	9c-5Cat	1.4	2.9	0.094	0.31	5	0.78	17.26	0.55	24.42
322	5Cat12c	10c-5Cat	1.7	4.0	0.094	0.31	5	0.74	19.23	0.52	26.82
323	5Cat12c	11c-5Cat	2.1	5.7	0.094	0.31	5	0.77	18.92	0.52	26.78
324	5Cat12c	12c-5Cat	2.4	7.0	0.094	0.31	5	0.86	18.30	0.60	24.23
325	5Cat24	1a-5Cat	0.7	0.9	0.054	0.23	5	0.78	24.10	0.69	27.60
326	5Cat24	2a-5Cat	1.0	1.7	0.054	0.23	5	0.83	22.85	0.61	31.19
327	5Cat24	3a-5Cat	1.4	3.1	0.054	0.23	5	1.09	18.75	0.66	29.89
328	5Cat24	4a-5Cat	1.7	4.4	0.054	0.23	5	1.18	18.13	0.66	30.58
329	5Cat24	5a-5Cat	2.1	6.2	0.054	0.23	5	1.18	18.76	0.69	30.08
330	5Cat24	6a-5Cat	2.4	7.6	0.054	0.23	5	1.16	20.41	0.74	29.28
331	5Cat24	7a-5Cat	0.7	0.8	0.054	0.23	5	0.83	22.54	0.65	29.08
332	5Cat24	8a-5Cat	1.0	1.6	0.054	0.23	5	1.00	19.50	0.63	30.76
333	5Cat24	9a-5Cat	1.4	2.9	0.054	0.23	5	0.98	20.60	0.63	31.24
334	5Cat24	10a-5Cat	1.7	4.0	0.054	0.23	5	1.32	16.44	0.68	29.54
335	5Cat24	11a-5Cat	2.1	5.5	0.054	0.23	5	1.23	18.47	0.66	32.02
336	5Cat24	12a-5Cat	2.4	6.8	0.054	0.23	5	1.14	20.97	0.67	32.75
337	5Cat24	1b-5Cat	0.7	0.8	0.054	0.23	5	0.65	28.37	0.65	29.01
338	5Cat24	2b-5Cat	1.0	1.6	0.054	0.23	5	0.67	27.76	0.60	31.70
339	5Cat24	3b-5Cat	1.4	2.9	0.054	0.23	5	0.83	24.02	0.63	31.65
340	5Cat24	4b-5Cat	1.7	4.1	0.054	0.23	5	1.09	18.77	0.63	31.66
341	5Cat24	5b-5Cat	2.1	5.8	0.054	0.23	5	0.91	23.29	0.64	32.49
342	5Cat24	6b-5Cat	2.4	7.1	0.054	0.23	5	1.04	20.86	0.72	29.42
343	5Cat24	7b-5Cat	0.7	0.9	0.054	0.23	5	0.72	25.85	0.65	29.09
344	5Cat24	8b-5Cat	1.0	1.7	0.054	0.23	5	0.87	22.39	0.67	28.81
345	5Cat24	9b-5Cat	1.4	3.2	0.054	0.23	5	1.23	16.31	0.63	30.61
346	5Cat24	10b-5Cat	1.7	4.5	0.054	0.23	5	1.03	20.76	0.67	30.37
347	5Cat24	11b-5Cat	2.1	6.3	0.054	0.23	5	1.06	22.81	0.72	30.90
348	5Cat24	12b-5Cat	2.4	7.8	0.054	0.23	5	1.22	22.19	0.72	33.31
349	5Cat36	1a-5Cat	0.7	0.9	0.036	0.19	5	0.81	28.67	0.71	33.03
350	5Cat36	2a-5Cat	1.0	1.7	0.036	0.19	5	0.93	25.42	0.73	32.29
351	5Cat36	3a-5Cat	1.4	3.1	0.036	0.19	5	1.16	21.57	0.67	35.99

352	5Cat36	4a-5Cat	1.7	4.4	0.036	0.19	5	1.54	17.21	0.71	34.68
353	5Cat36	5a-5Cat	2.1	6.2	0.036	0.19	5	1.47	19.29	0.77	33.94
354	5Cat36	6a-5Cat	2.4	7.6	0.036	0.19	5	1.25	23.43	0.68	39.33
355	5Cat36	7a-5Cat	0.7	0.8	0.036	0.19	5	0.81	28.67	0.70	33.62
356	5Cat36	8a-5Cat	1.0	1.6	0.036	0.19	5	1.04	23.17	0.70	34.09
357	5Cat36	9a-5Cat	1.4	2.9	0.036	0.19	5	1.09	23.03	0.67	35.98
358	5Cat36	10a-5Cat	1.7	4.0	0.036	0.19	5	1.45	18.18	0.71	34.62
359	5Cat36	11a-5Cat	2.1	5.5	0.036	0.19	5	1.34	20.78	0.76	34.20
360	5Cat36	12a-5Cat	2.4	6.8	0.036	0.19	5	1.31	22.12	0.80	33.34
361	5Cat36	1b-5Cat	0.7	0.8	0.036	0.19	5	0.73	31.56	0.70	33.93
362	5Cat36	2b-5Cat	1.0	1.6	0.036	0.19	5	0.75	31.41	0.69	34.92
363	5Cat36	3b-5Cat	1.4	2.9	0.036	0.19	5	0.93	26.62	0.65	37.96
364	5Cat36	4b-5Cat	1.7	4.1	0.036	0.19	5	1.04	24.24	0.69	36.45
365	5Cat36	5b-5Cat	2.1	5.8	0.036	0.19	5	1.19	22.50	0.75	35.03
366	5Cat36	6b-5Cat	2.4	7.1	0.036	0.19	5	1.32	20.61	0.86	31.05
367	5Cat36	7b-5Cat	0.7	0.9	0.036	0.19	5	0.74	30.59	0.68	34.02
368	5Cat36	8b-5Cat	1.0	1.7	0.036	0.19	5	0.83	28.40	0.65	36.57
369	5Cat36	9b-5Cat	1.4	3.2	0.036	0.19	5	1.24	20.48	0.68	35.77
370	5Cat36	10b-5Cat	1.7	4.5	0.036	0.19	5	1.15	23.15	0.70	36.14
371	5Cat36	11b-5Cat	2.1	6.3	0.036	0.19	5	1.33	22.64	0.76	35.93
372	5Cat36	12b-5Cat	2.4	7.8	0.036	0.19	5	1.60	20.66	0.78	37.26
373	5Cat36	1c-5Cat	0.7	0.8	0.036	0.19	5	0.68	33.05	0.68	34.40
374	5Cat36	2c-5Cat	1.0	1.6	0.036	0.19	5	0.75	31.36	0.64	37.73
375	5Cat36	3c-5Cat	1.4	2.9	0.036	0.19	5	1.14	21.32	0.70	34.60
376	5Cat36	4c-5Cat	1.7	4.0	0.036	0.19	5	1.03	24.22	0.67	36.18
377	5Cat36	5c-5Cat	2.1	5.7	0.036	0.19	5	1.17	21.87	0.64	38.01
378	5Cat36	6c-5Cat	2.4	7.0	0.036	0.19	5	1.47	18.30	0.70	35.72
379	5Cat36	7c-5Cat	0.7	0.8	0.036	0.19	5	0.75	30.04	0.73	31.84
380	5Cat36	8c-5Cat	1.0	1.6	0.036	0.19	5	0.76	30.77	0.64	37.46
381	5Cat36	9c-5Cat	1.4	2.9	0.036	0.19	5	1.08	22.37	0.68	35.07
382	5Cat36	10c-5Cat	1.7	4.0	0.036	0.19	5	1.04	23.68	0.65	36.70
383	5Cat36	11c-5Cat	2.1	5.7	0.036	0.19	5	1.46	17.65	0.72	33.60
384	5Cat36	12c-5Cat	2.4	7.0	0.036	0.19	5	1.19	22.29	0.69	36.19

BIBLIOGRAPHY

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Allen, N. L., Kline, D., & Zelenak, C. (1994). *The NAEP Technical Report*. Washington, DC.: National Center for Educational Statistics.

Ankenmann, R. D. (1994/1995). Goodness of fit and ability estimation in the graded response model (Doctoral dissertation, University of Pittsburgh, 1994). *Dissertation Abstracts International, 55*(10), 3167.

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood estimates for the graded model.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Services No ED. 347 208).

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit test statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*(4), 277-300.

Ansley, T. N., & Bae, H. W. (1989, April). *An empirical investigation of the nature of the distribution of an IRT goodness-of-fit statistic.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Baker, Frank B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker, Inc.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Camilli, G., & Hopkins, K. D. (1978). Applicability of chi-square to 2 x 2 contingency tables with small expected frequencies. *Psychological Bulletin*, *85*(1)*,* 163-167.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A,. (1983). *Graphical methods for data analysis*. Belmont, CA: Wadsworth International Group.

Childs, R., & Oppler, D. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high stakes testing program. *Educational and Psychological Measurement, 60*(6), 939–955.

Cohen, A. S., & Kim, S. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*(4), 116-130.

Cohen, A. S., Kim, S., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335-350.

Collins, L. M, Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-testing for latent class models. *Multivariate Behavioral Research 28* (3), 375-389.

Davey, T., Nering, M., and Thompson T. (1997, July). Realistic simulation of item response data. ACT Research Report Series 97-4.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*(2)*,* 155-170.

DeGroot, M. H. (1986). *Probability and Statistics, 2nd Edition.* Reading, Massachusetts: Addison-Wesley Publishing Company, Inc.

Donoghue, J. R., & Hombo, C. M. (1999, June). *Some asymptotic results on the distribution of an IRT measure of item fit.* Paper presented at the Annual Meeting of The Psychometric Society, Lawrence, KS.

Donoghue, J. R., & Hombo, C. M. (2001a, April). *The distribution of an item fit measure for polytomous items.* Paper resented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Donoghue, J. R., & Hombo, C. M. (2001b, April). *The effect of item parameter estimation of the distribution of an IRT measure of item fit.* Paper resented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Elliot, C. D., Murray, D. J., & Saunders, R. (1977). *Goodness of fit to the Rasch model as a criterion of test unidimensionality*. Manchester: University of Manchester.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement*. (3rd ed.) (pp. 147-200). Phoenix: The Oryx Press.

Hambleton, R. K., & Rogers, J. H. (1986, April). *Promising directions for assessing item response model fit to test data.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principals and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, Jane H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.

Harwell, M., Stone, C. A., Hsu, Tse-Chi, & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

Howell, D. C. (1999). *Fundamental statistics for the behavioral sciences* (4th ed). Pacific Grove, CA: Brooks/Cole Publishing Company.

Khattri, N., Reeve, A. L., & Kane, M.B. (1998). *Principles and practices of performance assessment*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Kim, S., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement 22*(4), 345-355.

Kingston, N., & Dorans, N. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9*(3), 281-288.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment. *Educational Measurement: Issues and Practice, 12*, 16-23.

Lane, S., Liu, M., Ankenmann R. D., & Stone C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.

Lane, S., Stone, C.A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response theory model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8*(4), 313-340.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement 9*(1), 49-57.

Mote, V. L., & Anderson, R. L. (1965). An investigation of the effect of misclassification on the properties of chi-square tests in the analysis of categorical data. *Biometrika,* 52, 95-110.

Muraki, E. & Carlson, J. (1993). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, (19),*1, 73-90.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64.

Reckase, M. D. (1985). The difficulty of tests items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*(2), 127–137.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Iowa City, IA: Psychometric Society.

Singer, Judith A. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*(4), 323-355.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*(1), 58 – 75.

Stone, C.A. (2003). Empirical power and Type I error rates for a goodness-of-fit statistic based on posterior expectations and resampling-based inference. *Educational and Psychological Measurement, 63*, 566-583.

Stone, C.A., Ankenmann, R. D., Lane, S., & Liu, M. (1993, April). *Scaling QUASAR's performance assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Stone, C.A & Hansen, M.A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement*, *60*(6), 974-991.

Stone, C.A., Mislevy, R. J., & Mazzeo, J. (1994, April). *Classification error and goodness-of-fit in IRT models.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Stone, C.A., & Zhang, B. (2003). Assessing goodness-of-fit of IRT Models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*(4), 331 – 352.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0). Mooresville, IN: Scientific Software.

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

U.S. Department of Education. (1999). Office of Educational Research and Improvement. National Center for Education Statistics. *The NAEP 1996 Technical Report,* NCES 1999-452, by Allen, N.L., Carlson, J.E., & Zelenak, C.A.. Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2001). Office of Educational Research and Improvement. National Center for Education Statistics. *The NAEP 1998 Technical Report,* NCES 2001-509, by Allen, N.L., Donoghue, J.R., & Schoeps, T.L. Washington, DC: National Center for Education Statistics.

Write, B. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, *23-48.*

Write, B. D and Mead, R. J. (1977). BICAL: Calibrating items and scales with the Rasch model (Research Memorandum No. 23). Chicago, IL: University of Chicago, Department of Education, Statistical Laboratory.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245-262.

Yen, S. & Rosenberger, K. (1999). *Technical Report: 1998 Maryland School Performance Assessment Program (MSPAP)*. Baltimore: Maryland State Department of Education. [On-line]. Available: http://www.marces.org/mdarch/pdf/mspap98t.pdf.