

Copyright
by
Michael John Lustina
2004

**The Dissertation Committee for Michael John Lustina Certifies that this is
the approved version of the following dissertation:**

**A Comparison of Andrich's Rating Scale Model and Rost's
Successive Intervals Model**

Committee:

Barbara G. Dodd, Supervisor

Gary D. Borich

William R. Koch

S. Natasha Beretvas

Anne E. Seraphine

**A Comparison of Andrich's Rating Scale Model and Rost's
Successive Intervals Model**

by

Michael John Lustina, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2004

A Comparison of Andrich's Rating Scale Model and Rost's Successive Intervals Model

Publication No. _____

Michael John Lustina, Ph.D.

The University of Texas at Austin, 2004

Supervisor: Barbara G. Dodd

This study compared and contrasted two IRT models for measuring attitudes: Andrich's rating scale model (ARSM) and Rost's successive intervals model (SIM). While these IRT models require that the attitude scale be unidimensional, they make different assumptions in the development of their item parameters. The ARSM and SIM were compared in the context of a computer adaptive test (CAT). Two data sets were used. The Audit of Administrator Communication (ADCOM) data set is archival and allowed for comparison of the models in an actual testing environment. A second data set was simulated using the linear factor analytic approach.

The models were compared using Pearson product-moment correlations, standard errors and number of items administered during a CAT. In addition, RMSE and bias estimates were calculated. Results indicated that each model has

advantages within a CAT context. The ARSM provided a better estimate of theta and the SIM required fewer items to estimate theta. Suggestions are provided as to the choice of model to use in different research settings.

Table of Contents

List of Tables.....	ix
List of Figures	x
CHAPTER I: INTRODUCTION	1
CHAPTER II: LITERATURE REVIEW	6
Overview	6
Item Response Theory.....	6
Dichotomous IRT Models	9
Polytomous IRT Models	10
Difference Models.....	13
Divide-By-Total Models	13
Comparison of Rating Scale Models.....	19
Computerized Adaptive Testing.....	20
Advantages of CAT over paper and pencil testing	20
Guidelines of Operational Procedures for Polytomous CATS.....	21
Item Bank	21
Item Selection Procedure	22
Trait Estimation Method	25
Stopping Rule	27
Polytomous Cat Research with Rating Scale Models	28
Andrich Rating Scale Model	28
Successive Intervals Model.....	30

CHAPTER III: STATEMENT OF PROBLEM	32
CHAPTER IV: METHOD	34
Overview	34
Data Sets.....	34
Audit of Administrator Communication	34
Simulated Data	35
Parameter Estimation	42
ARSM Parameter Estimates.....	42
SIM Parameter Estimates	43
Information.....	43
Summary of CAT Guidelines for Study.....	43
Data Analysis	45
CHAPTER V: RESULTS	46
Overview	46
Parameter Estimation	46
Item Pool Information	55
Descriptive statistics for known theta, ADCOM data and simulated data.....	60
Comparison of descriptive statistics for theta estimates	64
ADCOM Data	64
Simulated Data	64
Comparison of Descriptive Statistics for Standard Errors	68
ADCOM Data	68
Simulated Data	73

Comparison of Descriptive Statistics for Number of Items Administered	78
ADCOM Data	78
Simulated Data	81
RMSE and Bias	84
Difference Plots	84
CHAPTER VI: DISCUSSION	90
REFERENCES	98
VITA	102

List of Tables

Table 1:	Input Factor Loadings and Commonalities for the Generation of Simulated Data	38
Table 2:	Cutting Points for Generation of Simulated Data	40
Table 3:	ARSM Item Parameter Estimates for ADCOM Data	47
Table 4:	SIM Item Parameter Estimates for ADCOM Data.....	49
Table 5:	ARSM Item Parameter Estimates for Simulated Data	51
Table 6:	SIM Item Parameter Estimates for Simulated Data	53
Table 7:	Descriptive Statistics for Known θ	61
Table 8:	Mean and Standard Deviation of estimated θ , Standard Error, and Number of Items Administered for the ADCOM Data Set.	62
Table 9:	Mean and Standard Deviation of estimated θ , Standard Error, and Number of Items Administered for the SIM Data Set.....	63
Table 10:	Pearson Product-Moment Correlations among Theta Estimates for ADCOM data set.	65
Table 11:	Pearson Product-Moment Correlations among Theta Estimates for Simulated data set.	66
Table 12:	RMSE and Bias values for ARSM and SIM for simulated Data and ADCOM Data.....	85

List of Figures

Figure 1:	Hierarchy of Polytomous IRT Models	12
Figure 2:	The total information function for the ADCOM item pool with the ARSM.....	56
Figure 3:	The total information function for the ADCOM item pool with the SIM.....	57
Figure 4:	The total information functions for the simulated item pool with the ARSM.....	58
Figure 5:	The total information functions for the simulated item pool with the SIM.....	59
Figure 6:	Standard errors of ARSM within the Full Scale condition of the ADCOM data set	69
Figure 7:	Standard errors of ARSM within the CAT condition of the ADCOM data set	70
Figure 8:	Standard errors of SIM within the Full Scale condition of the ADCOM data set	71
Figure 9:	Standard errors of SIM within the CAT condition of the ADCOM data set	72
Figure 10:	Standard errors of ARSM within the Full Scale condition of the simulated data set	74
Figure 11:	Standard errors of ARSM within the CAT condition of the simulated data set	75

Figure 12: Standard errors of SIM within the Full Scale condition of the simulated data set	76
Figure 13: Standard errors of SIM within the CAT condition of the simulated data set	77
Figure 14: Number of Items Administered for ARSM within the CAT condition of the ADCOM data set.....	79
Figure 15: Number of Items Administered for SIM within the CAT condition of the ADCOM data set.....	80
Figure 16: Number of Items Administered for ARSM within the CAT condition of the simulated data set.....	82
Figure 17: Number of Items Administered for SIM within the CAT condition of the simulated data set.....	83
Figure 18: Difference between Full Scale theta and CAT theta for the ARSM with the ADCOM data set.....	86
Figure 19: Difference between Full Scale theta and CAT theta for the SIM with the ADCOM data set.....	87
Figure 20: Difference between Full Scale theta and CAT theta for the ARSM with the simulated data set.....	88
Figure 21: Difference between Full Scale theta and CAT theta for the SIM with the simulated data set.....	89

CHAPTER I

INTRODUCTION

An adaptive test is an efficient method of measurement where an optimal test is given to each examinee by asking only items that are pertinent to examinees' ability or attitude level. Each examinee is administered a test individually tailored to most accurately measure their ability level. The two major milestones that ushered individually adaptive testing into the modern age were item response theory (IRT) and computing power.

Item response theory, the theoretical basis for computer adaptive testing (CAT), developed as an alternative to classical test theory (CTT). The basis for IRT was laid out in Lord and Novick's (1968) *Statistical Theories of Mental Test Scores*, where Allan Birnbaum provided the insights that would allow for a test theory that utilized the item, rather than the entire test, as the fundamental unit (Wainer, 1990). One important IRT contribution is the idea of a latent continuum where items are organized from simplest to most difficult. The object of testing involves identifying where along this continuum an examinee exists. This concept makes it unnecessary to ask a respondent every test or scale item. The idea of ranking examinees on the same continuum without administering the same items provides the foundation for individually tailored tests. However, as IRT entered the 1970's, the computer power for this type of test was still lacking.

Beginning in the 1970's and throughout the 1980's, the topic of adaptive testing research increased in importance (Meijer & Nering, 1999). In the early

stages of adaptive testing research, methods focused on mental testing such as the branching strategy, which selects items sequentially from a predetermined logical branching structure (McBride, 1997). Today, a Graduate Record Exam (GRE) test-taker can be administered the GRE via computer using CAT methods. Educational Testing Service (ETS), the organization that produces the GRE, also offers several of their other large-scale tests via computer. In addition to the GRE, ETS has developed CAT versions for the GMAT (Graduate Management Admission Test), The Praxis Series: Professional Assessments for Beginning Teachers and The National Board for Professional Teaching Standards (NBPTS) Assessments.

The majority of CAT research completed to date has occurred within an educational context and using dichotomous models (Singh, Howell & Rhoads, 1990) where responses must be considered either correct or incorrect. Polytomous models are used when items require responses that are scored using more than two options. Three polytomous models were designed specifically for the case where a Likert-type attitude scale is used: Muraki's (Muraki, 1990) rating scale model (MRSM), Andrich's (Andrich, 1978) rating scale model (ARSM) and Rost's (1988) successive intervals model (SIM). These models are called rating scale models.

The rating scale models are designed to measure a unidimensional trait with ordered response categories. These models can be compared and contrasted in their model type classification, the model parameters and the category thresholds specified. The ARSM and the SIM are both divide-by-total models

where the probability of responding in a particular category is determined through dividing the numerator by the sum of all category probabilities so that the probabilities conditional on theta, the estimate of the trait being measured, sum to unity (Thissen & Steinberg, 1986). The SIM is the more general of the two models and simplifies to the ARSM when certain restrictions are placed on it. The MRSM is a difference model that estimates the probability of responding in each response category by subtracting successive categories (Thissen & Steinberg, 1986). The MRSM is not mathematically related to the other two models.

The second area for comparison is the number of parameters each model estimates. The MRSM and the SIM each estimate one additional parameter over the number used under the ARSM for each item in a scale. The third area of comparison is the treatment of category thresholds. The thresholds for the ARSM and the MRSM are held constant across all items. The SIM allows the category thresholds to vary proportionally across items. The current study will focus on two of these three models: the ARSM and the SIM. The MRSM was not included because it is not a commonly used model. In fact, no applications of the MRSM other than when it was originally proposed could be found in the literature to date.

Market research appears to be the ideal arena to extend CAT research beyond the educational setting and beyond mental testing. One key aspect is that the environment and technology is already in place in the form of Computer Aided Telephone Interviewing (CATI) (Kamakura & Balsubramanian, 1989). A CATI assists telephone interviewers as they survey respondents by identifying the

next question that should be asked. Currently, these are used to take respondents through surveys with skip patterns based on their responses. There is no reason that this technology would not be easily transferable to a CAT using one of the rating scale models, especially since the majority of market research surveys use Likert type scales.

There are several possible benefits to using CAT within market research. Singh, Rhoads, & Howell (1992) organize these into time efficiency, scale efficiency and precision efficiency. Time efficiency is defined by the information obtained during the given survey time. Currently, each respondent is asked every question. The benefit to time efficiency can be realized in either saved cost because of shorter survey time or increased information due to opening up survey space because of CAT efficiency. Scale efficiency is defined by the number of scale items needed to obtain the information desired. As alluded to above, the fewer items used to obtain information will have a benefit in either saved cost due to less time or more information due to more survey space. In addition, respondents should have less fatigue if surveys are not as long and they should be more interested if the items are more relevant to them. Precision efficiency is defined by the information obtained at the desired precision level. Currently, market researchers have a measurement error that is constant for all respondents in a given sample. CAT would be able to provide measurement error at each trait level.

However, there are several obstacles in place that continue to keep market researchers from using CAT. The primary challenge is the item bank. Most

market researchers develop a survey for a specific study and use individual items to measure certain concepts. It will be no small task to create a change where market researchers will begin to not only write several items to measure a concept, but take the time to calibrate the items for future use. In addition, a standard client deliverable breaks out responses to each item. Often, each item is in some way represented in a report. This will show itself in an educational barrier as market researchers will need to convince and educate their audience that item level reporting that includes every respondent is unnecessary and that CAT will provide them with more accurate (and more) information while not asking all questions.

A logical question follows the discussion of these exciting and new possibilities for administering attitude scales with CAT procedures. Which rating scale model would you select for use in a testing situation? Specifically, there is a need for further investigation of the IRT models that allow for CATs using Likert-type attitude scales. As mentioned, this study focuses on two of the three rating scale models: ARSM and SIM. Not only have very few research studies been conducted on these models, but no previous research has systematically compared these models. Therefore, the purpose of this research is to compare the rating scale models in the context of a CAT.

CHAPTER II

LITERATURE REVIEW

Overview

The literature review consists of two major sections. The first section introduces item response theory (IRT). This section begins with a brief overview of IRT and dichotomous IRT models. The focus then turns to polytomous IRT models, specifically the rating scale models that are the crux of this study. The second section presents the general guidelines for operational procedures for a polytomous CAT, and then provides specific research for the rating scale models and general findings/results.

Item Response Theory

Item response theory (IRT) has been described as “the theoretical glue that holds a CAT together” (Wainer, 1990 page 13). IRT is a series of theoretical models that mathematically define the interaction of a respondent and an item. More specifically, IRT models define the probability of a particular response to an item given a respondent’s level on a latent trait.

Hambleton and Swaminathan (1985) describe three advantages that IRT has over classical test theory (CTT). First, assuming a large pool of items, estimates of examinee ability are independent of the particular subset of items administered.

Second, assuming a large population of examinees, item parameter estimates are independent of the particular sample of examinees who are administered the items. These first two advantages together are called parameter invariance and allow respondents with traits at varying levels to be administered items of varying difficulty and still be evaluated on a common metric – an essential feature to CAT. The third advantage is that the IRT precision of measurement statistic is allowed to vary across the trait level. A separate standard error of measurement is available for each trait estimate and is often used as a criterion in ending CAT administration.

There are three main assumptions made under most IRT models, including the rating scale models. First, the majority of IRT models assume that a single trait is being measured. This is called the unidimensionality assumption. Following from unidimensionality is the assumption of local independence of item responses within a given trait level. In other words, responses to item 1 are uncorrelated with all other items for a given trait level. Violation of the local independence assumption often results in inflated trait estimates due to overestimation of item information (Wainer & Lewis, 1990). Finally, it is assumed that the mathematical function which represents the relationship between a trait level and the probability of a given response will be an accurate reflection of that relationship for the data. In other words, a model fits the data it is depicting.

Information functions indicate how much information or precision of measurement an item or test provides conditional on the trait level and are useful

for describing, comparing and selecting items. For dichotomous IRT models, item information can be defined as:

$$I_i(\theta) = \frac{P'_i(\theta)^2}{P_i(\theta)Q_i(\theta)}, \quad (1)$$

where:

$I_i(\theta)$ = information for item i , conditional on trait level, (θ) .

$P_i(\theta)$ = probability of responding correctly to item i , conditional on trait level, (θ) .

$Q_i(\theta)$ = probability of responding incorrectly to item i , conditional on trait level, (θ) .

P'_i = first derivative of $P_i(\theta)$.

Since the contribution of each item to test information is independent, test information is simply the sum of all item information functions. In addition, information is inversely related to the standard error of measurement defined as:

$$SE(\theta) = 1 / \sqrt{I(\theta)} \quad (2)$$

While a dichotomous information function is defined at the item level, a polytomous information function can be estimated for each category as well as for the entire item (Dodd, De Ayala, & Koch, 1995). In addition, polytomous items provide more information than dichotomous items (Dodd, et al., 1995) because information estimates can be obtained for each response category in a polytomous item.

The Samejima (1969) information function, although more complex than information functions derived for a specific model, is convenient in that it is general enough to use across many different models, including the ARSM and SIM. Samejima (1969) expressed the information for a given item as:

$$I_{ix}(\theta) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta)]^2}{P_{ix}(\theta)} \quad (3)$$

where P_{ix} is equal to the probability of obtaining a category score of x for a fixed θ and P'_{ix} is the first derivative of P_{ix} , and m_i is the number of categories. The entire scale information function is the sum of all the item information functions.

DICHOTOMOUS IRT MODELS

Dichotomous models score items as either correct or incorrect. Three general models exist for the unidimensional dichotomous case in IRT— the Rasch model (also called the one-parameter logistic model), the 2 parameter logistic model (2PL) and the 3 parameter logistic model (3PL) (Wainer, 1990).

The Rasch model contains one item parameter, b . This parameter represents the difficulty of an item and places the item along the trait continuum. In a Rasch model, the slopes of the ICCs are parallel and the lower asymptote is zero. The three assumptions in a Rasch model are that all items discriminate equally (the reason slopes are equal in the ICCs), guessing is non-existent, and the measured trait is unidimensional. The raw score in a Rasch model is a sufficient statistic for estimating theta and the proportion of examinees getting an item correct is sufficient for estimating item difficulty. Also, given it only has one item

parameter to estimate, it can be estimated with a smaller calibration sample than models with more parameters.

The 2PL model includes item discrimination, a , in the model thereby releasing one assumption—all items no longer have equal discriminating power. Thus, the ICCs are no longer assumed parallel to each other. The item discrimination parameter describes how well the item discriminates for the trait and is proportional to the slope. The 2PL is most appropriate when items vary in their ability to discriminate. When the item discrimination parameter is set to a constant across all items, the 2PL equals the Rasch model.

In addition to the discrimination item parameter, the 3PL model also adds a pseudo-guessing parameter, c , to the model and drops another assumption—a non-zero left asymptote is now modeled. The pseudo-guessing parameter is defined as the lower asymptote of the item characteristic function and represents the probability of respondents with minimum trait levels responding correctly to the item (Hambleton & Swaminathan, 1985). The 3PL is equal to the 2PL when the lower asymptote equals zero.

These Rasch and 2PL models are generalized into polytomous models when there are more than two response options. Both the Andrich rating scale model and the Rost's successive intervals model are extensions of the Rasch model. Polytomous models will be the focus of the remainder of this paper.

POLYTOMOUS IRT MODELS

The introduction of several polytomous IRT models has occurred over the past 30 years. Thissen & Steinberg (1986) created 5 classifications for

dichotomous and polytomous models placing them into distinct groupings distinguishing models within a category only by assumptions and parameter constraints imposed by the model. Three of the 5 model classifications are comprised solely of polytomous models: difference models, divide-by-total models, and left-side added divide-by-total models. Difference Models permit calculation of the probability of responding in each response category through subtraction and are appropriate for ordered responses (Thissen & Steinberg, 1986). The divide-by-total models calculate the probability of responding in a particular category directly. As the name suggests, the numerator is divided by the sum of all category probability that can appear in the numerator.

Divide-by-total models are appropriate for either nominal or ordered responses (Thissen & Steinberg, 1986). Left-side added divide-by-total models are designed for polytomous models that incorporate multiple choice items with the assumption of guessing. Rating scale models, the focus of this paper, do not comprise any of the left-side added divide-by-total models. For this reason, only difference and divide-by-total classifications will be discussed.

Dodd, De Ayala, and Koch (1995) arranged the difference models and the divide-by-total models so that the most general model would appear at the top and the most simplistic model would appear at the bottom of each classification (Figure 1). A line between two models indicates that the simpler model can be obtained by placing certain constraints on the more general model.

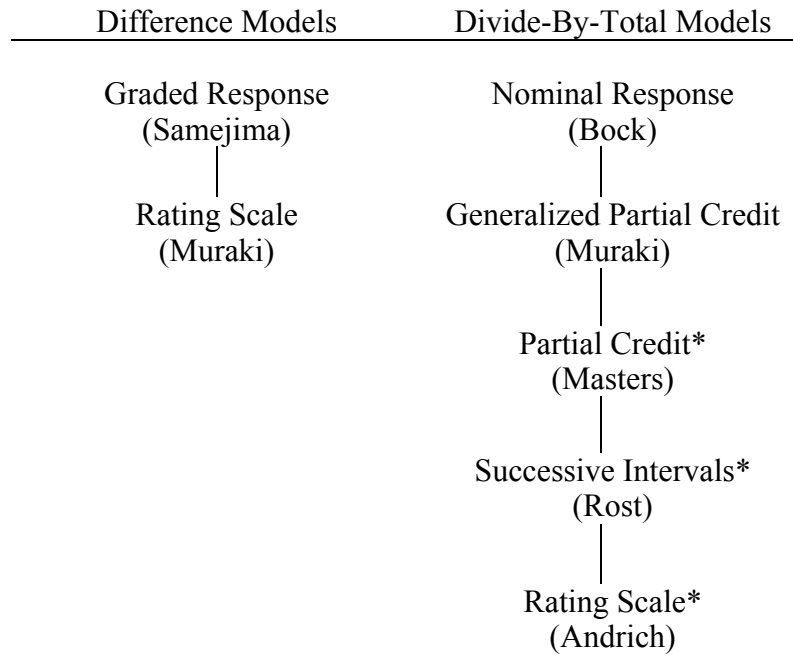


Figure 1: Hierarchy of Polytomous IRT Models

(*Model is a member of the Rasch family of IRT models)

Difference Models

Samejima's (1969) graded response model (GRM) and Muraki's (1990) rating scale model (MRSM) are two examples of difference models. The GRM is appropriate when responses to an item can be classified into three or more ordered categories. The GRM classifies these responses into sequentially ordered categories so that each successive category represents more of the measured trait. The GRM includes a discrimination parameter and set of category boundaries for each item. The MRSM was specifically created from the GRM for use with rating scale data. The MRSM keeps the discrimination parameter of the GRM and splits the category boundaries into an item location parameter for each item and a set of category boundaries for the entire set of items. A review of the literature revealed that to date, the MRSM has not been used. This may be because of the way attitude scales are constructed. Items are usually equal in terms of discrimination power and therefore using a model that allows items to vary in terms of discrimination has not been necessary. The remainder of this paper will focus solely on the other two rating scale models, the successive intervals model and Andrich rating scale model.

Divide-By-Total Models

The divide-by-total model classification includes Bock's (1972) nominal response model (NRM), Muraki's (1992) generalized partial credit model (GPCM), Master's (1982) partial credit model (PCM), Rost's (1988) successive intervals model (SIM) and Andrich's (1978) rating scale model (ARSM). The

NRM can be applied to responses that are ordered or non-ordered in terms of difficulty and is the most general of the divide-by-total models. It is most often used with multiple-choice items where it is difficult to order the responses based on their correctness. The NRM includes a discrimination parameter and an intercept parameter that reflects the interaction between category difficulty and category discrimination. The GPCM and PCM can also be applied to response scales that are ordered or non-ordered in terms of difficulty. Both models include a difficulty parameter for each category threshold in an item. The difference between the two models is that the PCM assumes equal discrimination across all items and the GPCM includes a discrimination parameter for each item. Both Rost's successive intervals model and Andrich's rating scale model were created specifically for attitude measurement. The two models both include a scale value parameter that identifies the location of the item along the attitude continuum and a set of category thresholds for the entire set of items. The SIM differs from the ARSM by including a dispersion parameter allowing category thresholds of each item to vary proportionally from the thresholds created for the entire set of items. The PCM, SIM, and ARSM will be described in more detail since the divide-by-total rating scale models are the focus of this dissertation and they are both specialized cases of the PCM.

Partial Credit Model

The PCM (Masters, 1982) is a generalization of the Rasch model to the polytomous case. Instead of only having one difficulty parameter for each item, as in the dichotomous case, the PCM has one difficulty parameter for every

category threshold that an individual potentially crosses. Thus, if there are 5 categories, there would be 4 difficulty parameters. Each difficulty parameter represents the movement from one category to the next. Masters termed these movements and their difficulty parameters “step difficulties.” The PCM places no restrictions on these threshold parameters (Rost, 1988). Thus, the thresholds of each item are allowed to differ irrespective of their relationship to other items. The model does require the steps within an item be completed in sequential order. The ordered categories represent the number of steps or subtasks involved in completing an item with successive integers (e.g. 0,1,2,3) for category scores with a lower number representing less of a trait or ability. An individual can not receive credit for step 3 before completing both steps 1 and 2. However, no requirement exists concerning the difficulty associated with progressing through the steps. Step 2 may be very difficult to complete, but, once step 2 is completed, step 3 may be completed easily. Finally, the PCM assumes that all items discriminate equally across all levels of θ . The probability that an individual with a given θ will obtain a category score of x on item i is (Dodd, et al., 1995):

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{h=0}^x (\theta - b_{ih})\right]}{\sum_{h=0}^m \exp\left[\sum_{k=0}^h (\theta - b_{ik})\right]} \quad (4)$$

where b_{ik} defines the category thresholds for the item (m_i thresholds for item i). For notational convenience, $\sum (\theta - b_{ik})$ is defined by Masters (1982) as being equal to 0.0 when k is 0.

Successive Intervals Model

The SIM (Rost, 1988) is a special case of the PCM, and thus also a specialized Rasch model. As mentioned, the PCM places no restrictions on its threshold parameters. Rost's proposed model places two requirements on the thresholds in the SIM. First, within an item, thresholds must be equidistant from each other. Second, distances between threshold values must remain proportional to each other throughout the scale. Rost accomplishes this by having two components make up the threshold value parameter (t_j)—a common set of threshold parameters (t_k) for the entire scale and a dispersion parameter (d_i). The dispersion parameter can be thought of as the distance an individual item's threshold values differ from the mean threshold values of the scale (Rost, 1988). This allows item 1 to have threshold values that are more spread out than item 2. However, the distances between the threshold value boundaries in item 2 must remain proportional to the distances between the mean threshold values for the scale. The parameters for the model follow:

b_i = scale value for item i ,

d_i = dispersion parameter for item i (when $d=0$, this model equals the ARSM, described next)

t_x = threshold values between the categories x and $x - 1$ for the set of items, and

θ = attitude level

The probability that a person with a given θ level will respond in a particular category for an item may be expressed as (Dodd, et al., 1995):

$$P_{ix}(\theta) = \frac{\exp \{K_x + x\theta - [xb_i + x(m-x)d_i]\}}{\sum_{h=0}^{mi} \exp \{K_h + h\theta - [hb_i + h(m-h)d_i]\}} \quad (5)$$

where K_x is the negative sum of the threshold value parameters associated with categories 1 to x . For notational convenience, t_0 is defined as equal to 0.0 so that equation 5 can obtain the probability of responding in category 0.

Andrich Rating Scale Model

The ARSM (Andrich, 1978) is a special case of the PCM where the threshold values are held constant across the entire set of items. To develop the model, Andrich extended the Rasch model for dichotomously scored items to the polytomous case for specific use with rating scales. The rating scale model estimates a scale value for each item that reflects the location of the item along the attitude continuum and also estimates a single set of thresholds for the entire set of items in the scale. Because the same response scale is used throughout, response threshold values are assumed to be constant across all items within a scale. Three general differences highlight the extension from one threshold in the dichotomous case to three or more categories in the ARSM (Andrich, 1978). First, rating scale threshold values qualify the scale value of the item (or the other way around). In other words, an agree response to an item with a moderate scale value may be equivalent to a neutral response to an item with a high scale value.

Thus, each threshold value retains the same value with respect to all items on the scale (or the other way around). Secondly, there is a separate response process with respect to each threshold for a given item. For example, in a 5 point strongly disagree to strongly agree scale, it may be easier to distinguish between strongly disagree and disagree than it is to distinguish between agree and strongly agree. Thirdly, the process of ordering multiple response categories and requiring a single response requires the respondent to simultaneously recognize the independence of the decision between each threshold while also recognizing the order. One cannot simultaneously have an opinion that is below the first threshold (i.e., strongly disagree) and above the third threshold (i.e., agree or strongly agree). The model includes the following parameters:

b_i = scale value parameter for item i . This parameter is estimated for each item to reflect the location of the item on the attitude continuum,

t_k = response threshold parameters for the set of items. A single set is estimated for the entire set of items included in the rating scale, and

θ = attitude level

The probability that a person with a given θ level will respond in category x to item i is defined as (Dodd, et al., 1995):

$$P_{xi}(\theta) = \frac{\exp [K_x + x(\theta - b_i)]}{\sum_{h=0}^{mi} \exp [K_h + h(\theta - b_i)]} \quad (6)$$

where K_x is the negative sum of the threshold values passed.

Comparison of Rating Scale Models

The ARSM and SIM are similar in that they were created specifically for Likert response scales, use ordered categories represented by successive integers, measure a unidimensional trait, require m thresholds for $m + 1$ categories and are both divide-by-total models. Because they are in the same classification, one model can be obtained when certain restrictions are placed on the more general model of the two, the SIM. When the dispersion parameter for the SIM is set to equal 0, it simplifies to the ARSM.

The models can be differentiated in the number of item parameters specified and the nature of the threshold restrictions. As with any estimation procedure, error is reduced for a fixed sample size when the number of estimated parameters is decreased because of a greater item parameter to person ratio given a common number of persons. In terms of rating scale models, fewer parameters to estimate equals less possible error—assuming all other factors are equal. The ARSM requires fewer estimated parameters than the SIM. The ARSM estimates a scale value (b) for each item and a set of thresholds (t_x) for the entire set of items. Assuming a 30-item scale with 5 response categories, 34 parameters would be estimated. The same parameters estimated in the ARSM are also estimated for the SIM. In addition, a dispersion parameter (d) is estimated for each item to allow the thresholds to vary proportionally from the mean set of thresholds for all items. 64 parameters would be estimated for a 30-item scale with 5 response categories.

The thresholds for the ARSM are held constant across the entire set of items. The SIM includes in its model the same thresholds as in the ARSM, but allows them to proportionally vary for each item by introducing an item dispersion parameter.

Computerized Adaptive Testing

The primary objective of CAT is to provide an optimal test for each respondent. In the context of attitude scales, the objective becomes to measure a respondent's attitude as precisely as possible while administering as few items as possible. A CAT does this by only administering those items that are pertinent to an individual respondent. Thus, each response in a CAT provides information that is used to determine what the next question will be. The next item administered is the one that will provide the most information at the currently estimated attitude trait level. In contrast to paper and pencil administration where everyone receives the entire scale, CAT respondents receive different items and can have different scale lengths depending on how long it takes to sufficiently estimate their attitude trait level.

ADVANTAGES OF CAT OVER PAPER AND PENCIL TESTING

CAT offers many advantages over the traditional paper and pencil testing format. The purported benefits of CAT include (Wainer, 1990; Meijer and Nering, 1999) increased efficiency in testing, improved test security due to both the increased physical security of the computerized item pool and to the individualized nature of a CAT, reduction in the negative effects of time

constraints for some examinees, reduction in examinee frustration and boredom, elimination of separate answer documents, immediate scoring and feedback to examinees, simple pretesting of items, easy removal of faulty items, and the ability to include new and innovative item types.

GUIDELINES OF OPERATIONAL PROCEDURES FOR POLYTOMOUS CATS

Dodd, et al. (1995) summarized the polytomous CAT operational procedure research to date and provided general guidelines for the four major components of a computer adaptive test: (1) the item bank, (2) the item selection procedure, (3) the trait estimation procedure, and (4) the stopping rule.

Item Bank

In a CAT, each respondent completes an individualized test. Each individualized test can be considered a test “form” created with specific items selected from the larger collection of items that make up the item bank (Wainer, 1990). Thus, a CAT can only be as good as the makeup of its item bank. An important aspect of the item bank is its size, since items need to be available for selection across all levels of theta. If there are not enough items to provide information across all theta levels, the theta estimates near the levels without enough information may not converge or may be poorly estimated. This results in a higher standard error for these theta estimates. Polytomous IRT models tend to have fewer nonconvergence problems because their items provide more information per item than dichotomous items (Dodd, et al., 1995).

Item bank research using polytomous models has demonstrated that an item bank with as few as 30 items may be sufficient for a polytomous CAT.

These studies have investigated the size of an item bank using the GRM (Dodd, et al. 1989), PCM (Dodd, Koch, & De Ayala, 1989; Koch & Dodd, 1989), SIM (Koch and Dodd, 1995) and ARSM (Dodd, 1990; Dodd & De Ayala, 1994). Item banks can be even smaller when using Likert-type attitude measurement. Item banks with as few as 24 items have worked well for the PCM (Koch & Dodd, 1989) and the ARSM (Dodd, 1990; Dodd & DeAyala, 1994). Note that the item bank size studies above did not take into account the impact of other testing issues, such as content validity and test security. When these issues are pertinent in a study, item banks will need to be increased or constraints will need to be placed on the item banks to help control for item exposure. For attitude measurements, test security is less important than in ability testing.

Item Selection Procedure

An item selection procedure is designed to select the next unused item remaining in an item bank that provides the most information at the examinees currently estimated theta level. The most common item selection procedure for polytomous CATs is maximum information (Lord, 1977). Items selected using maximum information are those that provide the most information at the examinee's current trait estimate. Information values are calculated for each remaining item at the current estimated trait level and the item that can provide the most information is administered. Research into rating scale models (ARSM & SIM) has introduced a method that utilizes the scale value parameter of these models (Dodd, 1990; Dodd & De Ayala, 1994; Koch and Dodd, 1995). A scale value estimate is calculated for rating scale models for each item that places the

item on the attitude continuum. Research has demonstrated that using the scale value closest to the current theta estimate for item selection provides similar theta estimation as using the maximum information method in the SIM (Koch & Dodd, 1995) and ARSM (Dodd, 1990; Dodd & De Ayala, 1994). Computational ease is the primary advantage of the scale value selection method. However, since computers have erased the need for calculation ease, this is no longer a sufficient reason to use this method.

Item exposure is an important issue to consider along with item selection. Item exposure problems can arise from the continuous administration of CATs from the same item bank (Meijer & Nering, 1999). The goal of exposure control is to limit item usage by limiting the frequency of administration, often achieved by placing constraints on item selection (Meijer & Nering, 1999). Way (1998) labeled item exposure procedures that control the probability of administering an item with specified criterion (e.g., expected frequency of item usage) “conditional item selection.” An example of a conditional item selection strategy is the Simpson-Hetter. The Simpson-Hetter procedure (Simpson & Hetter, 1985) is one of the most widely known item exposure control method. It attempts to directly control the rate of item exposure by assigning an exposure parameter (between 0 and 1) to each item. This information is used in the selection algorithm by comparing it to a number randomly generated from a uniform distribution. The item is administered if the exposure parameter is greater than the random number. If it is not, a new item is selected.

Recently, Pastor, Dodd, and Chang (2002) examined variations of the Symptom-Hetter and compared them to variations of the a-stratified design in the polytomous case using the GPCM. Stratification procedures group items into strata based on a selected statistical property and then only allow items to be administered from their strata. The a-stratified design (Chang & Ying, 1999) divides the item pool into x different strata based on the value of the item discrimination parameter. The strata are arranged in ascending order of discrimination and the test is divided into x stages to match the strata, beginning with the lowest strata and ending with the most discriminating. Thus, items at all levels of discrimination have equal chances of being used in a test.

The results from Pastor et al. (2002) indicated that identifying which item exposure control method to utilize should depend on the purpose of the test and its need for security of items. A more simplistic approach to exposure control such as the a-Stratified design would be most appropriate with low to medium stakes testing and any of the four other methods investigated would be appropriate for higher stakes testing.

Item exposure controls are generally incorporated to help maintain test security when the stakes of a test are high, as in the SAT and GRE. If an examinee has prior knowledge of an item, their responses will not be an accurate measure of their true trait level. Stocking & Lewis (2000) note that the effort a testing program should put into exposure control depends on the uses of the test scores and whether the testing program can be classified as high, medium, or low stakes. In terms of attitude measurement, item exposure controls may not be

important. For example, in medical outcome assessment, we are interested in a patient's quality of life during the course of treatment. The fact that the patient has seen the item before should not matter in measuring their quality of life.

Trait Estimation Method

As mentioned earlier, CATs provide each respondent with an individualized test. A CAT administers only those items that are most informative for a specific respondent. As each item is being administered, a trait estimation procedure estimates the current level of the trait and feeds the information back into the CAT. Several methods have been developed to obtain the current theta estimate. Maximum likelihood estimation (MLE), Bayesian methods, and Warm's (1989) weighted likelihood estimation (WLE) will be discussed here.

Maximum likelihood estimation (MLE) estimates the likelihood function given a set of responses to items already administered during the CAT session. MLE assumes that items fit the IRT model and that the item parameters are known. The advantage of MLE is that it is both consistent and efficient (Hambleton & Swaminathan, 1985). The disadvantage of MLE is that it cannot provide a theta estimate until a response occurs in one of the non-extreme categories. It requires a procedure to choose the next item until a respondent selects a response in a non-extreme category. Rating scale model research indicates that the variable stepsize method is an appropriate selection method (Dodd, 1990). The variable stepsize method sets the new theta halfway between

the current theta estimate and one of the two most extreme item parameter estimates in the item bank, depending on the response to the previous item.

Bayesian methods take into consideration information about the population distribution in their estimates. Allowing an informed prior estimate of the ability distribution reduces error in the final ability estimate and avoids unreasonable values of the theta estimate (Meijer & Nering, 1999). Two common Bayesian methods are maximum a posteriori (MAP) and expected a posteriori (EAP). Maximum a posteriori (MAP) uses the mode of the posterior distribution and expected a posteriori (EAP) uses the mean of the posterior distribution as the first estimate. One major advantage of Bayesian methods is their ability to obtain a theta estimate after one item, whereas MLE procedures require responses to fall in at least two different categories before an estimate can be made. A disadvantage of Bayesian estimates is their tendency to regress toward the prior mean when there is a large difference between the estimated likelihood and the mean of the prior distribution (Meijer & Nering, 1999).

Warm (1989) proposed a weighted likelihood estimation (WLE) procedure to reduce the bias of the estimate. The WLE of the current theta estimate is the value of the estimate that maximizes a weighted likelihood function. For the 1PL and 2PL, this weight equals the square root of the test information function (Meijer & Nering, 1999). Simulation studies have found that WLE ability estimates were less biased than either MLE or Bayesian for 2 and 3 PL tailored tests (Warm, 1989). WLE also used fewer items than MLE over the entire range

of theta to obtain the same precision of measurement, resulting in shorter testing times and lower item exposure.

A few recent studies have investigated the properties of trait estimation procedures in polytomous CATs (Chen, Hou, & Dodd, 1998; Chen, Hou, Fitzpatrick, & Dodd, 1997). The majority of these studies have compared MLE and EAP and have demonstrated that MLE and EAP perform equally well under certain conditions in a polytomous CAT system. One condition is when prior estimates of the ability distributions are fairly accurate, such that they match the actual latent trait distribution. Research indicates that Bayesian EAP performs as well as MLE with respect to accuracy of ability estimates for most polytomous models (Chen, Hou, & Dodd, 1998; Chen, Hou, Fitzpatrick, & Dodd, 1997). Wang and Wang (2001) compared WLE to MLE and Bayesian estimation methods in a GPCM CAT and found that under certain conditions WLE reduced bias compared to Bayesian methods across the theta scale. Gorin, Dodd, Fitzpatrick and Shieh, (2000) investigated the WLE with the PCM and found that the WLE performs equally as well as the MLE and EAP when an optimal item bank is available (one that has information across all levels of theta). EAP was found to outperform the other two when non-optimal item banks were used.

Stopping Rule

There are three options for deciding how to end a CAT: fixed length stopping rule, variable length stopping rule, or a combination of the two. The fixed length stopping rule ends a CAT once a pre-specified number of items is administered. The variable length stopping rule ends a CAT once a pre-specified

standard has been met, such as a certain value for the standard error. The combination of these two would terminate a CAT when a respondent reaches one of these criteria. The combination of the two is the most often used with rating scale research (Koch, Dodd, Fitzpatrick, 1990; Dodd & De Ayala, 1994; Chen, et al. 1997; Koch & Dodd, 1995).

Polytomous Cat Research with Rating Scale Models

The operational procedures of CAT systems based on both the ARSM and the SIM have been studied. The following subsections review the research on polytomous CAT with the ARSM and SIM.

ANDRICH RATING SCALE MODEL

Dodd (1990) investigated the operational procedures of a CAT using the ARSM by systematically varying the item selection procedure (maximum information or closest scale value) and the stepsize method (variable or fixed). The variable method adjusted by half the distance to the extreme category. The fixed used stepsizes of either .4 or .7. Additionally, multiple item banks were tested (39 items, 32 items and 24 items). The CAT using maximum information stopped when there was no item left above the average item information for the theta values (.44). The CAT using scale values terminated when a standard error of .3 was reached. The study used real and simulated data sets and identified three major outcomes: 1) Item banks consisting of as few as 25 items may be adequate for CAT, 2) the variable stepsize procedure produced fewer nonconvergent cases than the fixed stepsize procedure, and 3) The scale value

item selection procedure combined with a minimum standard error stopping rule outperformed the maximum information item selection procedure combined with the minimum information stopping rule in three ways: frequency of nonconvergent cases, number of items administered, and correlation of known estimates and estimated CAT θ .

Koch, Dodd, and Fitzpatrick (1990) investigated CAT efficiency and students' attitudes toward CAT in a field test. Participants took a CAT using the ARSM with the following operational procedures: item bank of 40, maximum likelihood estimation with variable stepsize method (half the distance to the extreme), scale value item selection, and a combination stopping rule terminating when either a minimum standard error of .35 or 20 administered items was reached. Results indicated the students had a favorable attitude toward taking the CAT. Operationally, the item bank was peaked rather than spread across theta levels, leaving relatively few items to measure low or high attitudes. However, there was a 60% decrease in test length in the CAT as compared to a full scale paper-and-pencil test.

Dodd and De Ayala (1994) further investigated the two item selection procedures studied by Dodd (1990): maximum information and scale value item selection. The CATs used the ARSM with real (item banks of 39 and 24) and simulated data sets (item bank of 32), and using maximum likelihood estimation with variable stepsize (half the distance to the extreme). CAT sessions under each procedure administered items until a pre-specified standard error was reached or a

maximum of 20 items had been administered. Results indicated that both methods performed equally well.

Chen, Hou, Fitzpatrick, and Dodd (1997) investigated the effect of population distribution on MLE and EAP in a simulated CAT using the ARSM. The Simulated CAT incorporated item banks of 39 items, 32 items, and 24 items and used the scale value item selection method. Maximum likelihood estimation using variable stepsize (half the distance to the extreme) and EAP with 10 quadrature points were each employed with a combination stopping rule of a minimum standard error of .25 or a maximum of 20 items. Comparisons were made between the MLE, EAP with a normal prior distribution, EAP with a uniform prior distribution generated from a normal trait distribution, and EAP with a uniform prior distribution generated from a negatively skewed trait distribution. Results using EAP were similar to MLE, regardless of whether the prior distribution matched the underlying θ distribution.

SUCCESSIVE INTERVALS MODEL

Koch and Dodd (1995) investigated the operational procedures of the SIM using real and simulated data sets. Results indicated that item banks of 30, 39 or 61 items performed well regardless of whether items were selected using the closest scale value method with variable stepsize (half distance to extreme) or the maximum information selection method with variable stepsize (half distance to extreme). Item banks containing items with small dispersion parameters and item banks with large dispersion parameters performed equally well. The simulated CAT recovered the known attitude trait levels of the simulees and the real data set

CAT performed well in terms of number of items administered and accuracy of trait estimation.

CHAPTER III

STATEMENT OF PROBLEM

The advances over the last 20 years in computer technology have created an ideal environment for CAT. One area where CAT has flourished is in the area of ability testing. This is most evident in the mass testing provided by the Educational Testing Service. Recently, they have started to offer computer adaptive testing as an option for the General Test of the GRE. Unfortunately, there has been no breakthrough like this for non-ability CAT. Specifically, the rating scale IRT models are not being utilized in applied settings.

An ideal setting for this breakthrough is market research. Market researches often use rating scales to gauge the perceptions of consumers and/or customers. In these instances, either items are individually analyzed or a scale is created to analyze. One common dilemma in market research is attempting to obtain responses to many questions in a short period of time with a respondent. Transaction market research is a common type of market research that lends itself to CAT. Transactional research occurs when a survey is given to a customer after an experience such as purchasing a car. The purchaser either fills out a survey and sends it in or is called on a telephone and responds to the survey. Current advances have allowed for on-line surveying or Interactive Voice Response Surveying where an automated survey is given over the telephone. In all these cases outside of the paper survey, it would be quite easy to implement a

CAT. The advantage to this would be in the amount of information the researcher could obtain with a similar amount of questions.

Market research is the applied setting where it makes sense to administer an attitude measurement via CAT. The key question still remaining is which rating scale model should be selected to administer a rating scale CAT in an applied setting. The focus of this study is to compare and contrast the two rating scale models that have been individually studied in the context of CAT.

CHAPTER IV

METHOD

Overview

Two polytomous IRT rating scale models (Andrich rating scale model and Rost's successive intervals model) were compared in their performance on two data sets—The Audit of Administrator Communication (ADCOM) data set and a simulated data set. The ADCOM data set is archival and allowed for a comparison of the models in an actual testing context. The simulated data set was computer generated and allowed for a comparison between actual and estimated thetas.

Data Sets

AUDIT OF ADMINISTRATOR COMMUNICATION

Responses from 491 teachers to the ADCOM scale were available for use in this study (see Koch, 1983 for original data collection). The ADCOM scale (Valentine, 1978) measures attitudes about teachers toward the communication skills of their school administrators. It is a 40-item Likert-type attitude scale scored on a five-point scale, with 0 representing an unfavorable response toward the communication skills of the administration and a 4 representing a favorable response. Exploratory factor analysis of the ADCOM scale (Koch, 1983)

indicated that the scale is approximately unidimensional, with the first factor accounting for about 85% of the common variance.

In previous research (Dodd, 1990), one respondent in the ADCOM data set was found to have responded in the highest category for each item. In addition, this research demonstrated that MLE of the lowest step value for item 31 was unobtainable since no person had responded in the lowest category for this item. This had the effect of making item 31 a three-category item rather than a four-category item as the other 39 still were. Thus, the respondent and item described above were both removed from the data set creating a 490 person data set with responses for 39 items of the ADCOM scale. This was the data set that was utilized in the current study.

SIMULATED DATA

The most common approach for generating data—creating data to fit a particular model—would create a problem in this study. The focus of the study is to compare and contrast the two rating scale models in as controlled an environment as possible. Thus, the proposed model comparison requires that data not be biased toward either of the models. The linear factor analytic approach espoused by Wherry, Naylor, and Fallis (1965) solves this problem by generating data that is neutral towards any model. The linear factor analytic approach described below was used to create two data sets. One data set was used for input into PARSCALE for item calibration and the other was used for input into the CAT program. Two data sets were used to ensure that the results were not capitalizing on outcomes of chance that could occur through the use of the same

data set for item calibration and for the CAT. For each simulated data set, 1000 random numbers in z-score form were selected from a normal distribution. These z-scores were considered to be each simulees' true ability, θ_T . An additional random number, again in z-score form, was then selected for each simulee to serve as a random error component. The mathematical definition used to calculate simulee j's response to item i was:

$$z_{ji} = a_i z_j + z_{eji} \sqrt{1 - h_i^2} \quad (7)$$

where:

a_i = item i's factor loading

z_j = examinee j's randomly selected z-score (i.e., θ_T)

z_{eji} = examinee j's randomly selected z-score error component for item i.

h_i^2 = item i's communality

Beginning with simulee one and item one, the item z-score was compared to the z-score cutting point from item one to determine simulee one's item score (from 0 to 4). This step was repeated until item scores were obtained for simulee one for all items and then repeated for each successive simulee until responses were generated for all simulees for all items.

In order to perform the program described above, two inputs were required: (1) a factor loading matrix for test items, and (2) z-score cutting points. The factor loading matrix was obtained from a principal axis factor analysis of the ADCOM data set so that the data reflected real attitude data. As described earlier, previous exploratory factor analysis into the ADCOM scale (Koch, 1983) revealed one dominant factor. For this reason, only the factor loadings for the first factor were input into the data generation program. The factor loadings and

communalities for each item are provided in Table 1. The z-score cutting points were chosen so that the frequencies of the possible item scores approximated the ADCOM data. Since the theta scale is similar to the z-scale, cumulative frequencies of item scores were converted to the z-scores corresponding to proportions under the normal curve represented by these cumulative frequencies. These z-scores served as the cutting points in the data generation program. The cutting points for the generation of the data sets are reported in Table 2.

Table 1: Input Factor Loadings and Commonalities for the Generation of Simulated Data

Item Number	Factor Loading	Communality
1	0.686	0.470
2	0.595	0.354
3	0.652	0.425
4	0.723	0.523
5	0.787	0.619
6	0.407	0.166
7	0.709	0.502
8	0.826	0.683
9	0.536	0.288
10	0.473	0.224
11	0.518	0.268
12	0.557	0.310
13	0.643	0.413
14	0.598	0.358
15	0.649	0.421
16	0.708	0.501
17	0.723	0.522
18	0.699	0.489
19	0.720	0.519
20	0.588	0.345
21	0.671	0.451
22	0.654	0.428
23	0.670	0.449
24	0.533	0.284
25	0.521	0.272
26	0.700	0.491
27	0.629	0.396
28	0.704	0.496

Table 1: Continued

Item Number	Factor Loading	Communality
29	0.458	0.210
30	0.531	0.281
31	0.705	0.497
32	0.783	0.613
33	0.805	0.648
34	0.723	0.523
35	0.817	0.667
36	0.596	0.355
37	0.657	0.431
38	0.683	0.466
39	0.682	0.466

Table 2: Cutting Points for Generation of Simulated Data

Item Number	z_1	z_2	z_3	z_4
1	-1.355	-0.659	0.312	1.580
2	-1.697	-1.023	-0.427	0.427
3	-1.742	-1.259	-0.590	0.578
4	-1.936	-1.204	-0.659	0.062
5	-1.636	-1.050	-0.416	0.495
6	-2.404	-1.817	-1.194	-0.307
7	-1.528	-0.578	0.248	1.248
8	-1.742	-1.114	-0.536	0.416
9	-2.404	-1.512	-0.737	0.461
10	-2.047	-1.451	-0.730	0.590
11	-2.322	-1.697	-1.086	0.806
12	-2.047	-1.528	-0.792	0.211
13	-1.742	-1.077	-0.372	0.778
14	-1.163	-0.410	0.280	0.902
15	-1.306	-0.659	0.185	1.282
16	-1.742	-1.032	-0.275	0.678
17	-2.254	-1.512	-0.941	0.015
18	-1.655	-0.864	-0.206	0.990
19	-2.254	-1.636	-0.886	0.339
20	-2.322	-1.616	-0.973	0.572
21	-2.652	-1.395	-0.886	0.372
22	-1.318	-0.627	-0.021	0.973
23	-1.817	-0.981	-0.280	1.015
24	-0.501	0.312	0.925	1.616
25	-0.835	0.021	0.857	1.817
26	-1.096	-0.339	0.328	1.183
27	-2.008	-1.293	-0.659	0.211

Table 2: Continued

Item Number	z_1	z_2	z_3	z_4
28	-1.451	-0.717	0.170	1.497
29	-1.655	-0.864	-0.206	0.990
30	-2.254	-1.636	-0.886	0.339
31	-2.322	-1.616	-0.973	0.572
32	-2.652	-1.395	-0.886	0.372
33	-1.318	-0.627	-0.021	0.973
34	-1.817	-0.981	-0.280	1.015
35	-0.501	0.312	0.925	1.616
36	-0.835	0.021	0.857	1.817
37	-1.096	-0.339	0.328	1.183
38	-2.008	-1.293	-0.659	0.211
39	-1.451	-0.717	0.170	1.497

Parameter Estimation

The PARSCALE (Muraki & Bock, 1993) software package was used to estimate the item parameters of the ADCOM data set and the simulated data set. Parameter estimates were obtained for the partial credit model first and then transformed into estimates for each of the models. This makes sense since the PCM is a generalized version of both models.

PARSCALE estimates item parameters through a marginal maximum likelihood EM algorithm (Muraki, 1992). This algorithm consists of two steps: (1) calculate the provisional expected frequency and sample size and (2) estimate the marginal maximum likelihood. Once the item parameters have been estimated, maximum likelihood or EAP is used to estimate person parameters. The PCM item parameter estimates were then transformed into the ARSM and SIM item parameters, respectively. Recall that for each category boundary the PCM estimates a difficulty parameter. In the case of the data used in this study, there were 5 categories and 4 category boundaries. Thus, PARSCALE estimated 4 difficulty parameters.

ARSM PARAMETER ESTIMATES

The ARSM has 1 scale value parameter for each item and one set of thresholds for the entire set of items. The estimate of the item scale value parameter was calculated through the average of the 4 step difficulties per item estimated for the PCM. The estimate of the thresholds for the entire set of items was also obtained from the PCM estimates. First, each of the PCM step value estimates for an item was transformed into a deviation score from the scale value

for the item. Averaging the deviation for each step across the set of items yields the threshold for that step. There were 4 thresholds estimated for each data set.

SIM PARAMETER ESTIMATES

There was one additional step in calculating the parameters for the SIM. The item mean of the differences between adjacent step values was subtracted from the mean of the differences between adjacent step values for the entire set of items. The obtained number was then divided by two and the dispersion value of the item was obtained (Rost, 1988).

Information

The information function was computed for each item in the ADCOM data set and each item in the simulated data set. The information function was based on the item parameter estimates obtained with the process outlined above. The equation specified by Samejima (1969) and described in the literature review was used to calculate item information. Test information was determined by summing the item information functions for each item within a data set. IRTINFO (Fitzpatrick, et al., 1994), a SAS program, was used to calculate the information functions for each model.

Summary of CAT Guidelines for Study

The operational procedures for this study were selected with the intention of using the procedures of CAT consistent with past research and procedures of CAT that would allow for a comparison across the two rating scale models. As

these models have never been compared to each other, the most often used polytomous CAT procedures were selected for use.

The item banks consisted of 39 items each. This is well above the 30 that was recommended by Dodd, De Ayala, and Koch (1995) for a CAT with a rating scale model. For item selection, this study used the maximum information selection procedure. This was a departure from the rating scale literature that has suggested using scale value for item selection. However, the studies comparing these two procedures demonstrated they performed equally and the recommendation was that scale values are a viable alternative to maximum information to save computing time during a CAT (Dodd & De Ayala, 1994; Koch & Dodd, 1995). This is no longer necessary with the power and speed of today's computers. Furthermore, using maximum information is currently the standard for polytomous CAT research.

For trait estimation, this study used maximum likelihood estimation with the variable stepsize method for item selection until an initial trait was estimated. Chen, et al. (1998) demonstrated that MLE performed equally in comparison with various EPA methods, and since MLE is still the most often used trait estimation method for CAT, there is no reason to depart from it in this initial study of the rating scale models.

The stopping rule was a variable length stopping rule that ended when either a standard error of .3 was reached or when a limit of 20 items had been administered. This has proven effective in the studies investigating the operational procedures for CATs using the ARSM and SIM (Koch, Dodd, &

Fitzpatrick, 1990; Dodd & De Ayala, 1994; Chen et al., 1997; Koch & Dodd, 1995).

Data Analysis

Various descriptive statistics were calculated and compared for the CAT conditions of the two models. For each model, Pearson product-moment (PPM) correlation coefficients were calculated between full-scale estimated theta and CAT estimated theta. For the simulated data sets, PPM correlation coefficients were also calculated between z-scores and CAT theta estimates. The accuracy and precision of theta estimation were evaluated with bias and root mean square error (RMSE), respectively. The Bias and RMSE equations follow:

$$\text{Bias} = \frac{\sum_k^n (\hat{\theta}_k - \theta_k)}{n} , \tag{8}$$

$$\text{RMSE} = \left[\frac{\sum_k^n (\hat{\theta}_k - \theta_k)^2}{n} \right]^{1/2} , \tag{9}$$

CHAPTER V

RESULTS

Overview

As outlined in the previous chapter, this study focuses on the comparison between the ARSM and SIM polytomous IRT models. This chapter describes these results through comparisons of the theta estimates, standard errors and number of items administered for each model in a full scale condition and CAT condition for the ADCOM data set and the simulated data set. In addition, for the simulated data set, each of these estimates were compared with the known parameters.

Parameter Estimation

PARSCALE was run on the ADCOM data and the simulated data sets to obtain parameter estimates for the ARSM and SIM with each data set. As described in the previous chapter, parameter estimates were obtained for the partial credit model first and then transformed into estimates for each of the models. The estimated ARSM and SIM parameters for the ADCOM data follow in Table 3 and Table 4, respectively. The estimated ARSM and SIM parameters for the simulated data follow in Table 5 and Table 6, respectively.

Table 3: ARSM Item Parameter Estimates for ADCOM Data

Item Number	<i>b</i>
1	0.058
2	-1.000
3	-1.044
4	-1.401
5	-0.934
6	-2.110
7	-0.232
8	-1.078
9	-1.637
10	-1.311
11	-1.526
12	-1.507
13	-0.863
14	-0.163
15	-0.116
16	-0.879
17	-1.764
18	-0.619
19	-1.638
20	-1.599
21	-1.870
22	-0.331
23	-0.750
24	0.881
25	0.768
26	0.050
27	-1.405

Table 3: Continued

Item Number	<i>b</i>
28	-0.095
29	-1.923
30	-1.593
31	-0.772
32	-0.383
33	-1.251
34	-1.065
35	-1.359
36	0.391
37	-0.735
38	-1.118
39	-1.857

$t_1 = -1.484$ $t_2 = -0.581$ $t_3 = 0.082$ $t_4 = 1.982$

Table 4: SIM Item Parameter Estimates for ADCOM Data

Item Number	<i>b</i>	<i>d</i>
1	0.058	0.162
2	-1.000	-0.130
3	-1.044	-0.109
4	-1.401	-0.173
5	-0.934	-0.147
6	-2.110	-0.197
7	-0.232	0.132
8	-1.078	-0.123
9	-1.637	0.149
10	-1.311	0.018
11	-1.526	0.190
12	-1.507	-0.149
13	-0.863	0.008
14	-0.163	-0.168
15	-0.116	0.029
16	-0.879	-0.032
17	-1.764	-0.077
18	-0.619	0.088
19	-1.638	-0.010
20	-1.599	0.130
21	-1.870	0.296
22	-0.331	-0.051
23	-0.750	0.157
24	0.881	-0.129
25	0.768	0.080
26	0.050	-0.060
27	-1.405	-0.101

Table 4: Continued

	Item Number	<i>b</i>	<i>d</i>
	28	-0.095	0.180
	29	-1.923	0.051
	30	-1.593	0.079
	31	-0.772	-0.049
	32	-0.383	-0.095
	33	-1.251	-0.044
	34	-1.065	-0.101
	35	-1.359	-0.133
	36	0.391	0.228
	37	-0.735	0.215
	38	-1.118	-0.098
	39	-1.857	-0.017
$t_1 = -1.484$	$t_2 = -0.581$	$t_3 = 0.082$	$t_4 = 1.982$

Table 5: ARSM Item Parameter Estimates for Simulated Data

Item Number	<i>b</i>
1	0.080
2	-0.987
3	-0.996
4	-1.363
5	-0.898
6	-2.112
7	-0.221
8	-1.075
9	-1.884
10	-1.340
11	-1.438
12	-1.397
13	-0.724
14	-0.150
15	-0.163
16	-0.777
17	-1.621
18	-0.600
19	-1.396
20	-1.413
21	-1.935
22	-0.317
23	-0.710
24	0.774
25	0.748
26	0.003
27	-1.378

Table 5: Continued

Item Number	<i>b</i>
28	-0.077
29	-1.810
30	-1.428
31	-0.752
32	-0.385
33	-1.171
34	-1.047
35	-1.325
36	0.341
37	-0.684
38	-1.007
39	-1.724

$t_1 = -1.398$ $t_2 = -0.497$ $t_3 = 0.039$ $t_4 = 1.856$

Table 6: SIM Item Parameter Estimates for Simulated Data

Item Number	b	d
1	0.080	0.175
2	-0.987	-0.099
3	-0.996	-0.066
4	-1.363	-0.158
5	-0.898	-0.153
6	-2.112	-0.059
7	-0.221	0.089
8	-1.075	-0.076
9	-1.884	0.358
10	-1.340	0.101
11	-1.438	0.170
12	-1.397	-0.178
13	-0.724	0.048
14	-0.150	-0.137
15	-0.163	0.017
16	-0.777	-0.062
17	-1.621	-0.105
18	-0.600	0.092
19	-1.396	-0.093
20	-1.413	0.054
21	-1.935	0.353
22	-0.317	-0.053
23	-0.710	0.144
24	0.774	-0.159
25	0.748	0.101
26	0.003	-0.059
27	-1.378	-0.104

Table 6: Continued

Item Number	b	d
28	-0.077	0.135
29	-1.810	0.025
30	-1.428	0.004
31	-0.752	-0.059
32	-0.385	-0.109
33	-1.171	-0.084
34	-1.047	-0.094
35	-1.325	-0.149
36	0.341	0.229
37	-0.684	0.189
38	-1.007	-0.138
39	-1.724	-0.089

$t_1 = -1.398$ $t_2 = -0.497$ $t_3 = 0.039$ $t_4 = 1.856$

Item Pool Information

Four total test information functions were computed and plotted. A total test information function was created for both data sets with both models. These functions are presented in Figures 2 through 5. All four total test information functions are peaked and negatively skewed. This is not surprising since the scale values presented in Tables 3 through 6 were predominately negative across all parameter estimates. It should be noted that the ideal total test information function for CAT usage would have a uniform distribution so that information would be spread across all theta levels. In the case of these 4 total information functions, more information is available at lower levels of theta than higher levels of theta. In all four information functions, the most information occurs between theta levels above -2.0 and below -0.5 .

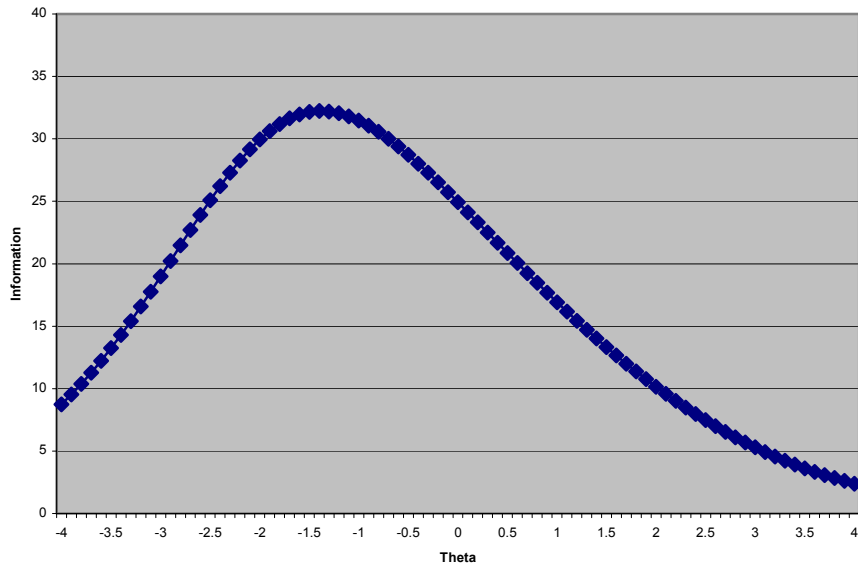


Figure 2: The total information function for the ADCOM item pool with the ARSM.

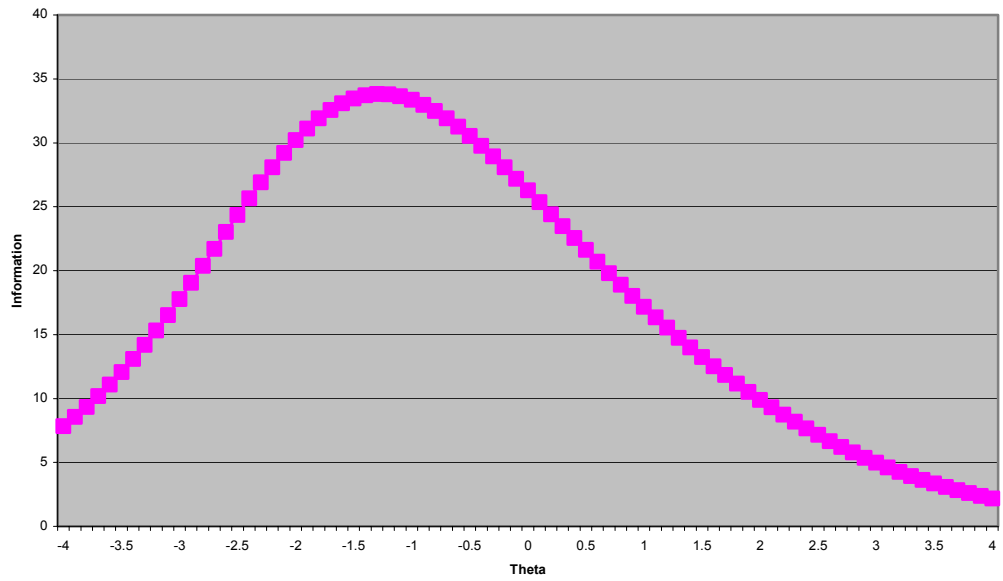


Figure 3: The total information function for the ADCOM item pool with the SIM.

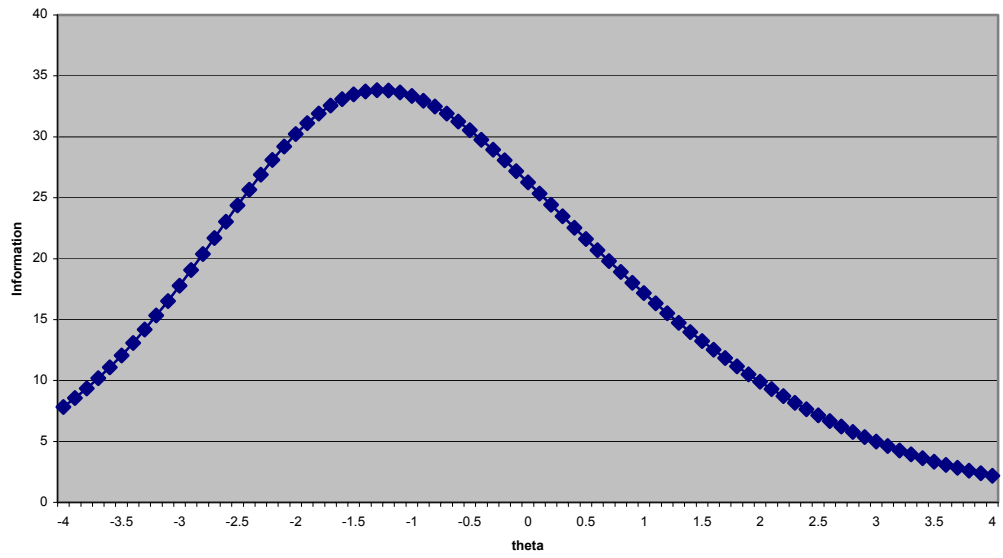


Figure 4: The total information functions for the simulated item pool with the ARSM.

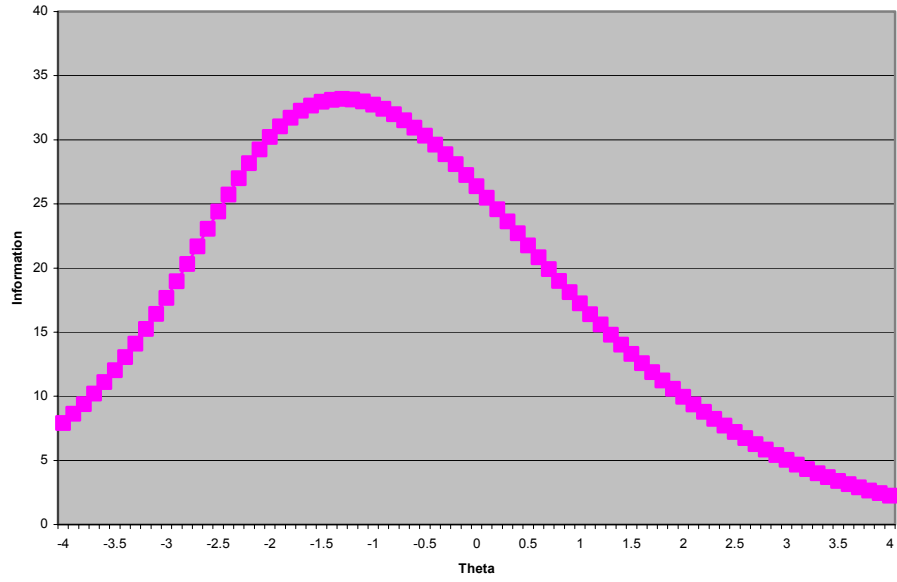


Figure 5: The total information functions for the simulated item pool with the SIM.

Descriptive statistics for known theta, ADCOM data and simulated data

As described in the previous chapter, The simulated data was created by using the factor loading matrix from a principal axis factor analysis of the ADCOM data set. This was done so that the results of the comparisons would more closely represent a real situation. Known theta were compared with thetas estimated from the full scale and CAT conditions based on the ARSM and SIM. The descriptive statistics of the known theta are shown in Table 7. The mean of known theta was just above 0 (.026) and the standard deviation was 1.0. One theta was deleted because it could not be estimated in any of the CAT conditions.

Each data set has a full scale condition and CAT condition for the ARSM and SIM. For all conditions, there was one case of non-convergence due to inconsistent responding leaving a total of 489 cases for each ADCOM condition and 999 cases for each simulated condition. In each non-convergent case, all item responses were in the highest category for each item but 1. The descriptive statistics for the ADCOM and simulated data set conditions are shown in Table 8 and Table 9, respectively. The mean and standard deviations are displayed for estimated theta, standard error, and number of items administered for the ARSM full scale condition, ARSM CAT condition, SIM full scale condition, and SIM CAT condition. These descriptive statistics are discussed in the following sections.

Table 7: Descriptive Statistics for Known θ

	N	Mean	SD	Min	Max
Known θ	999	.026	1.0	-3.344	3.115

Z-score for deleted case was 3.958

Table 8: Mean and Standard Deviation of estimated θ , Standard Error, and Number of Items Administered for the ADCOM Data Set.

Model	Estimated θ		SE		NIA	
	Mean	SD	Mean	SD	Mean	SD
ARSM Full Scale	.016	1.105	.216	.0485	39	0
ARSM CAT	-.011	1.088	.305	.0341	15.378	2.714
SIM Full Scale	.028	1.106	.216	.0486	39	0
SIM CAT	.010	1.060	.306	.0380	14.335	3.097

Table 9: Mean and Standard Deviation of estimated θ , Standard Error, and Number of Items Administered for the SIM Data Set

Model	Estimated θ		SE		NIA	
	Mean	SD	Mean	SD	Mean	SD
ARSM Full Scale	.017	1.011	.209	.045	39	0
ARSM CAT	-.012	.986	.303	.029	14.546	2.882
SIM Full Scale	.027	1.110	.210	.045	39	0
SIM CAT	.006	.953	.304	.033	13.595	3.201

Comparison of descriptive statistics for theta estimates

Tables 8 and 9 display the mean and standard deviation of estimated theta. The results indicate that the means and standard deviations for all four conditions are close to a mean of 0 and standard deviation of 1. Theta estimates will be compared in the context of how well the full scale condition and CAT condition estimates known theta in the simulated data set and how well the CAT theta estimates the full scale theta in the simulated data set and the ADCOM data set. These results are discussed in the context of each data set.

ADCOM DATA

For the ADCOM data set calibrations, Pearson product moment correlations were calculated between the full scale and CAT condition for the ARSM and the SIM. These results are presented in Table 10. The ARSM Full Scale and CAT thetas (.964) are slightly more correlated than the SIM Full Scale and CAT thetas (.950).

SIMULATED DATA

For the simulated data set calibrations, Pearson product moment correlations were calculated between known theta and all four conditions: ARSM Full Scale theta, ARSM CAT theta, SIM Full Scale Theta and SIM CAT theta. In addition, Pearson product moment correlations between Full Scale and CAT theta are provided for both models. These results are presented in Table 11. In the Full Scale conditions, the Pearson product moment correlation between known theta and ARSM and known theta and SIM are almost identical. However, in the CAT

Table 10: Pearson Product-Moment Correlations among Theta Estimates for ADCOM data set.

	ARSM Full Scale	ARSM CAT	SIM Full Scale	SIM CAT
ARSM Full Scale	1.000	.964	1.000	.950
ARSM CAT		1.000	.964	.978
SIM Full Scale			1.000	.950
SIM CAT				1.000

Table 11: Pearson Product-Moment Correlations among Theta Estimates for Simulated data set.

	Known Theta	ARSM Full Scale	ARSM CAT	SIM Full Scale	SIM CAT
Known Theta	1.000	.975	.945	.975	.927
ARSM Full Scale		1.000	.973	.999	.962
ARSM CAT			1.000	.973	.974
SIM Full Scale				1.000	.961
SIM CAT					1.000

condition, the Pearson product moment correlation of .945 between known theta and the ARSM theta is slightly higher than the Pearson product moment correlation of .927 between known theta and SIM theta. Thus, while there is no difference in the ability to estimate known theta with the Full Scale, the ARSM estimates known theta slightly better than the SIM in a CAT situation. A similar result occurs when the Pearson product moment correlation between Full Scale theta and CAT theta are compared. The ARSM Full Scale theta and CAT theta are slightly more highly correlated (.973) than the SIM Full Scale theta and CAT theta (.961), a finding consistent with the results from the ADCOM data set.

Comparison of Descriptive Statistics for Standard Errors

Tables 8 and 9 also show the mean standard errors for the Full Scale and CAT conditions of each model for each data set. The results indicate that the standard errors are similar for all four Full Scale conditions (ranging from .209 to .216) and for all four CAT conditions (ranging from .303 to .306). The mean standard errors for the CAT conditions are higher due to the fact that the stopping rule imposed for the CAT was a standard error of .30 with a maximum administered item limit of 20. Figures 6 to 13 display the plots of standard errors against estimated theta for each condition within each data set. As the graphs demonstrate, error is highest at the highest theta levels in all conditions of both data sets. This is attributable to the lower amount of information available at the higher levels of theta in each data set. These results are discussed in the context of each data set.

ADCOM DATA

Figures 6 and 7 display the plots of the standard errors against estimated theta for the ARSM ADCOM Full Scale and CAT conditions and Figures 8 and 9 display the same information for the SIM ADCOM Full Scale and CAT conditions. As the graphs show, error is almost identical across the theta level for the ARSM and SIM within each condition. When looking at the means in table 8, the mean standard errors for the Full Scale condition are almost identical for both models. In the CAT condition, the mean standard error for the ARSM is .001 lower than the mean standard error for the SIM.

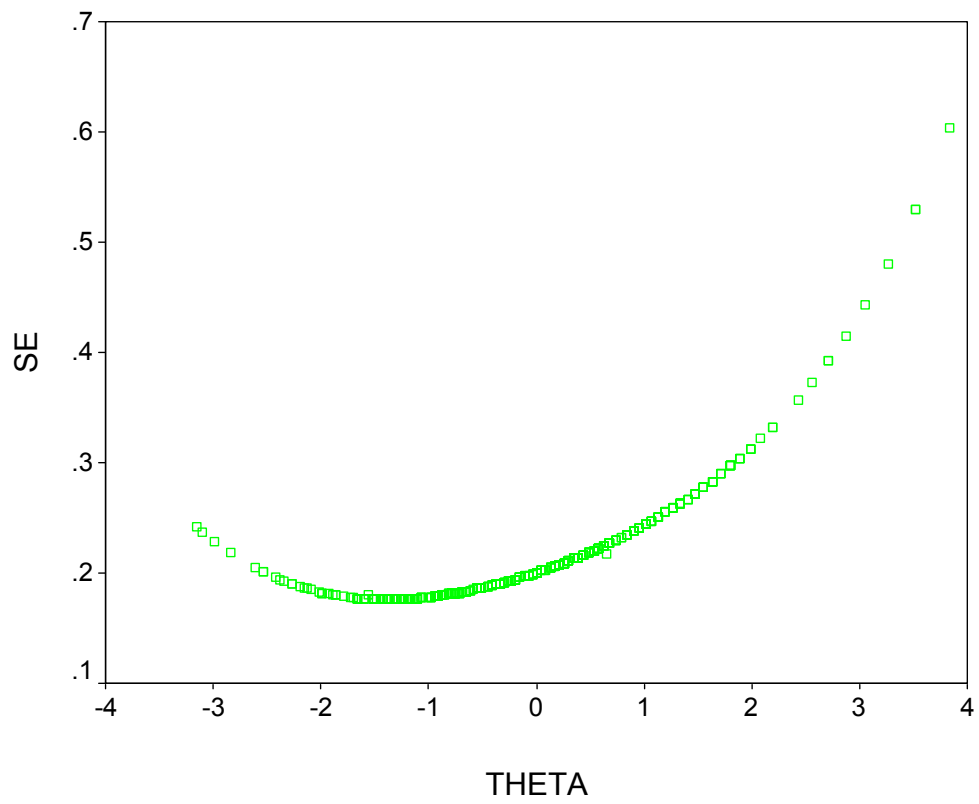


Figure 6: Standard errors of ARSM within the Full Scale condition of the ADCOM data set

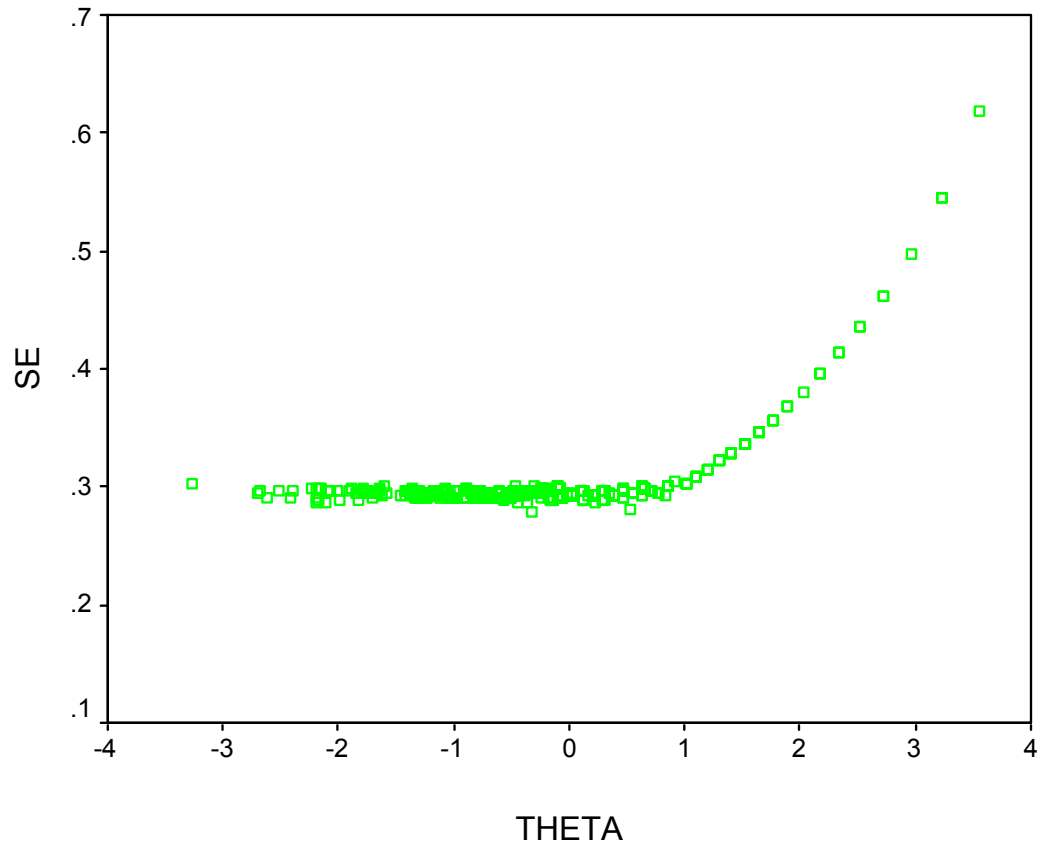


Figure 7: Standard errors of ARSM within the CAT condition of the ADCOM data set

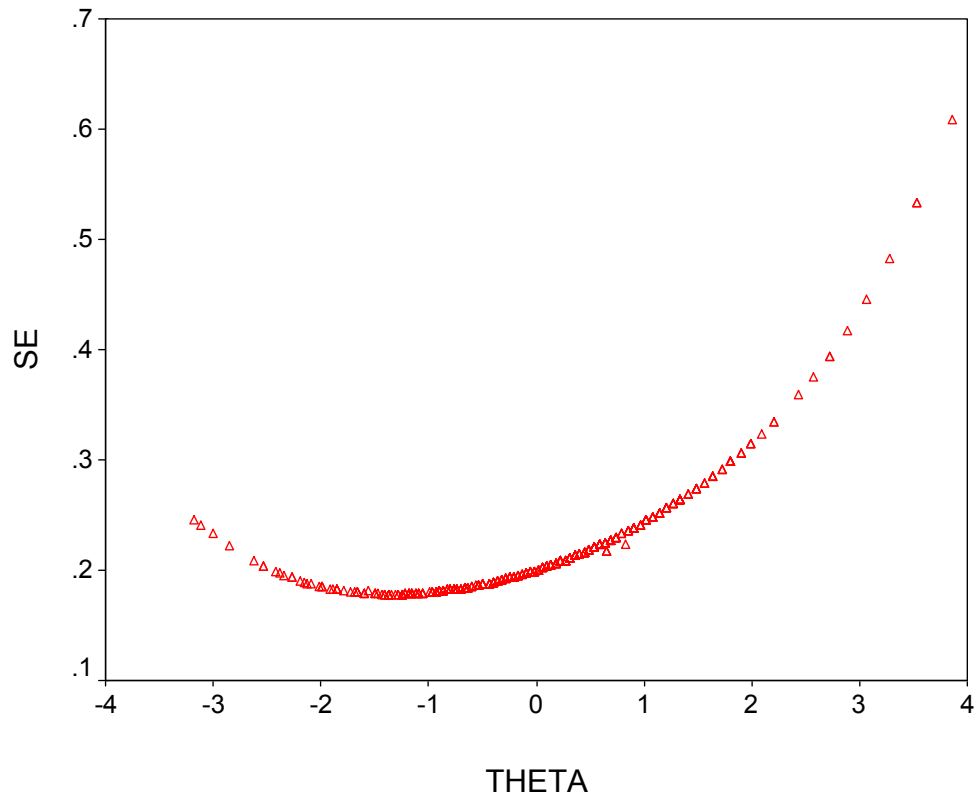


Figure 8: Standard errors of SIM within the Full Scale condition of the ADCOM data set

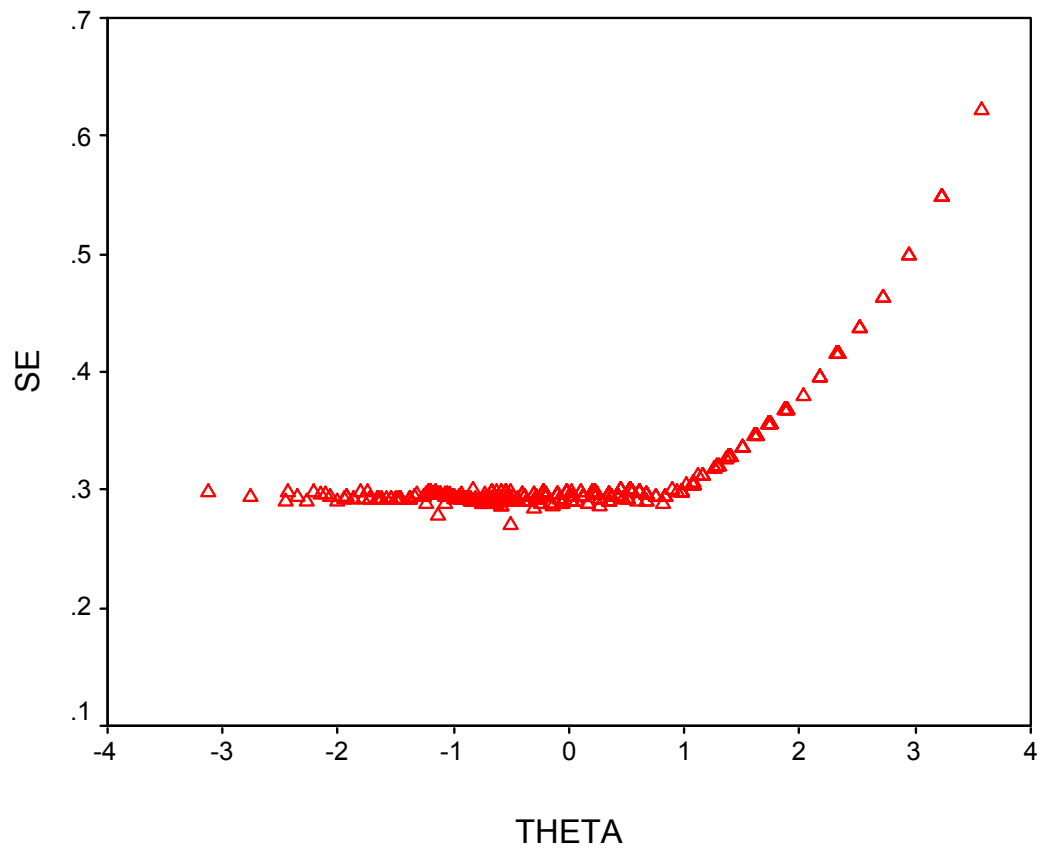


Figure 9: Standard errors of SIM within the CAT condition of the ADCOM data set

SIMULATED DATA

Figures 10 and 11 display the plots of the standard errors against estimated theta for the ARSM simulated data Full Scale and CAT conditions and Figures 12 and 13 display the same information for the SIM simulated data Full Scale and CAT conditions. These results follow the same pattern as in the ADCOM data set, with error almost identical for each model across the theta scale in each condition. When the mean of the standard errors displayed in Table 9 are compared, the ARSM is slightly (.001) lower than the SIM in both the Full Scale and CAT conditions.

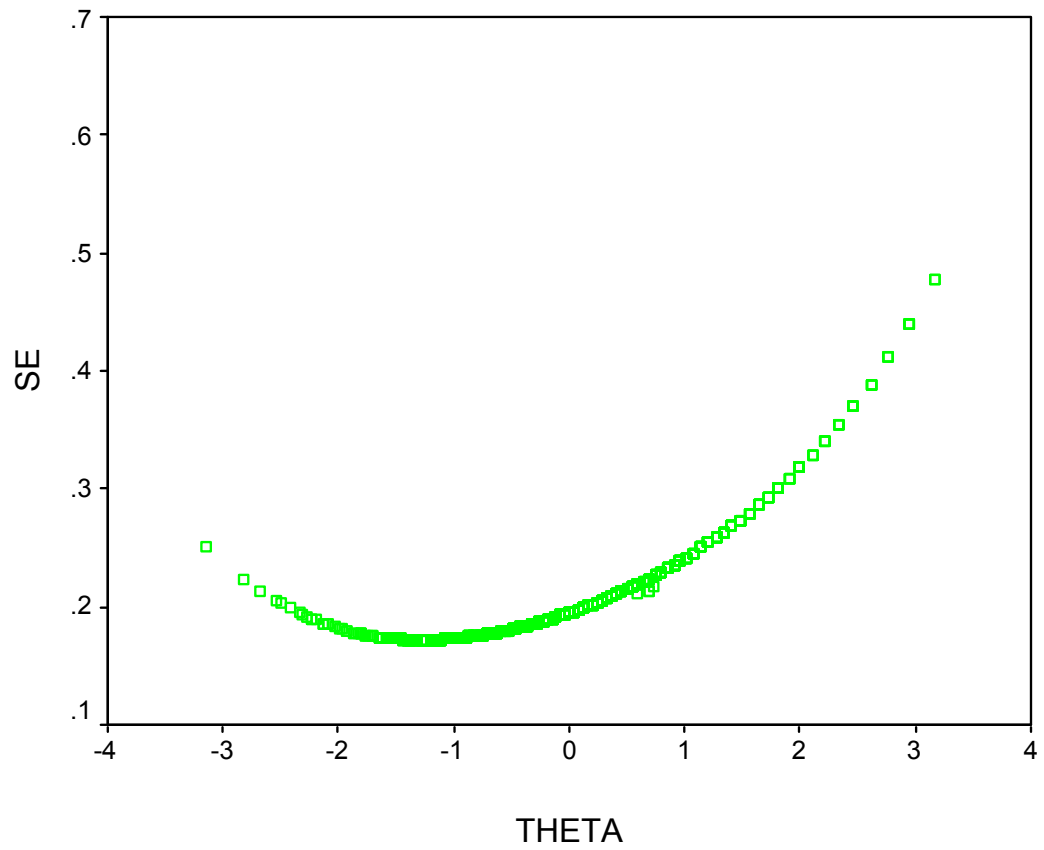


Figure 10: Standard errors of ARSM within the Full Scale condition of the simulated data set

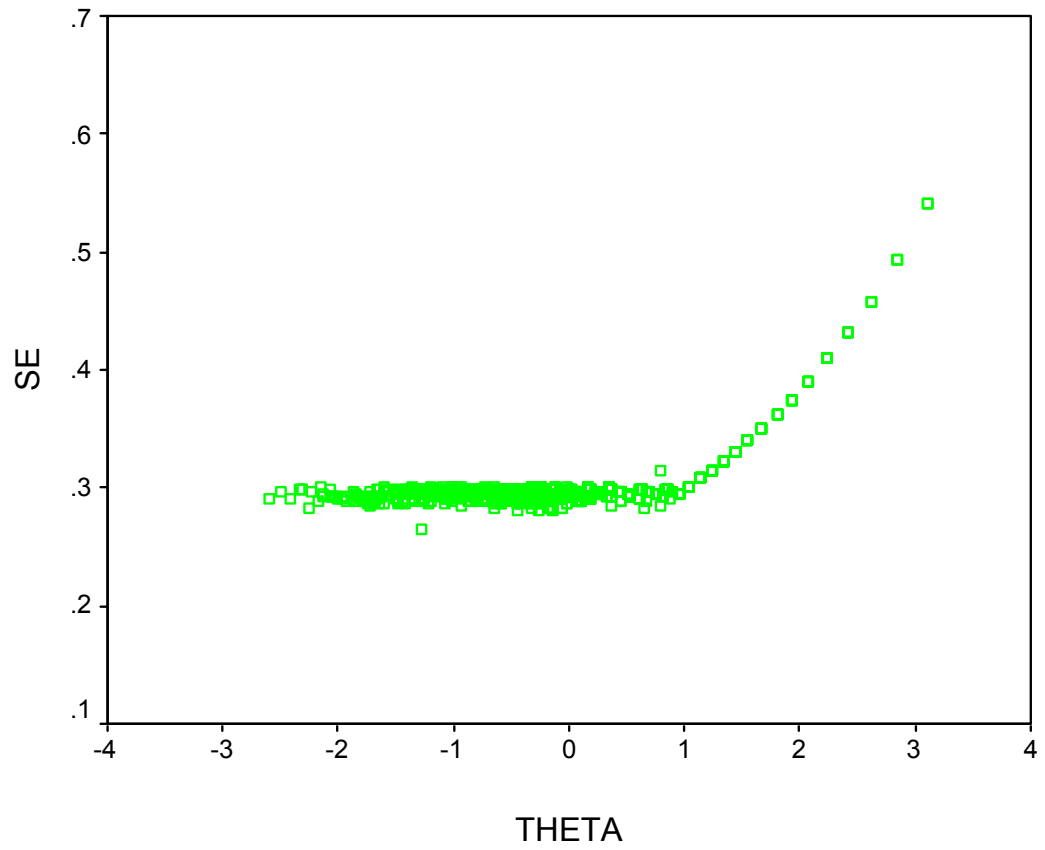


Figure 11: Standard errors of ARSM within the CAT condition of the simulated data set

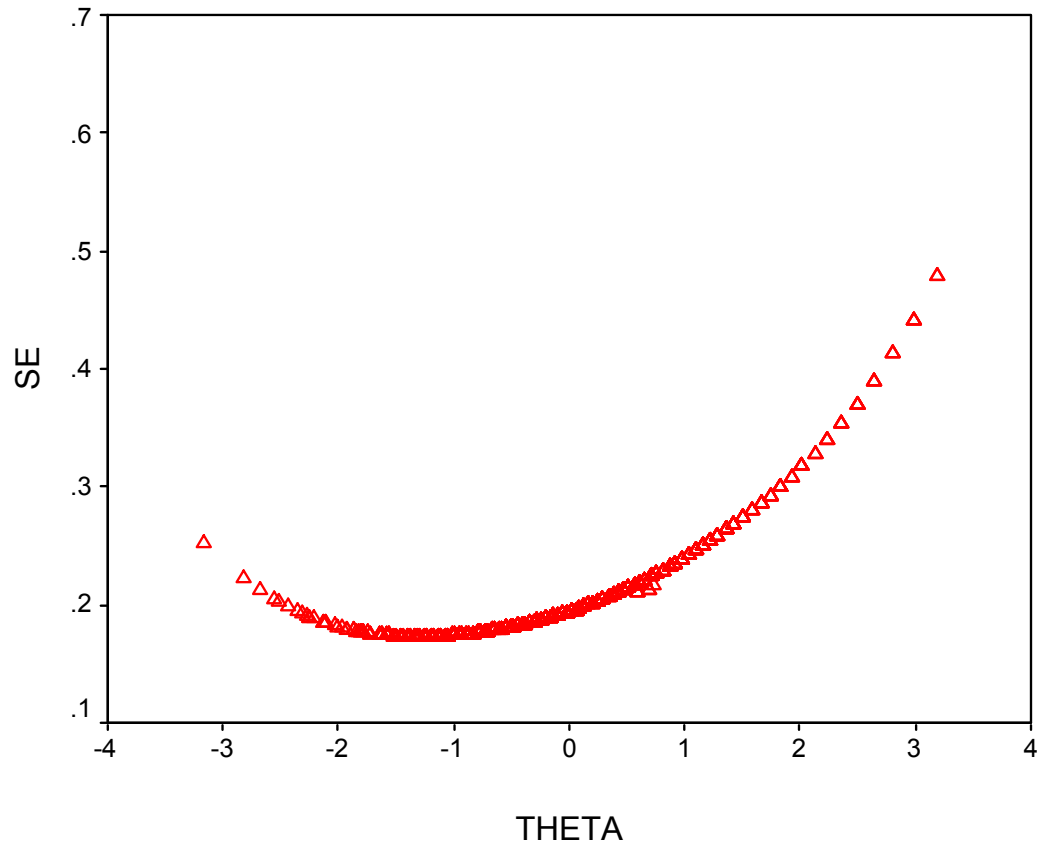


Figure 12: Standard errors of SIM within the Full Scale condition of the simulated data set

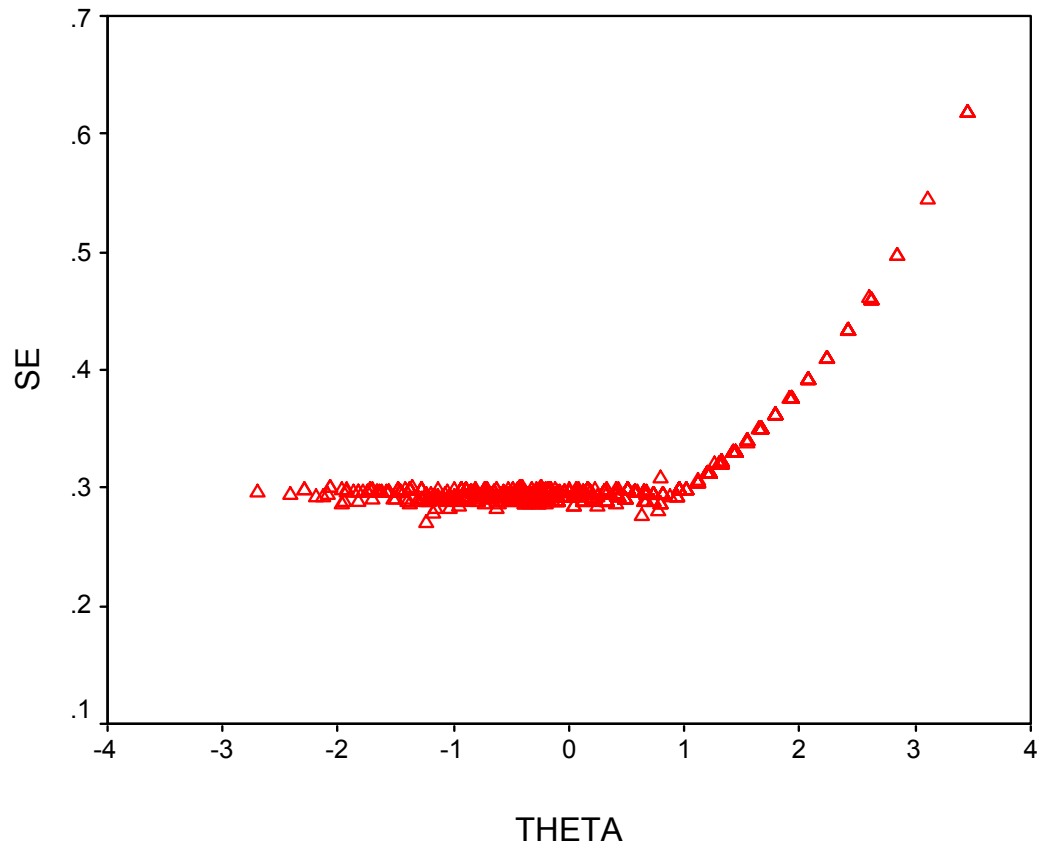


Figure 13: Standard errors of SIM within the CAT condition of the simulated data set

Comparison of Descriptive Statistics for Number of Items Administered

Tables 8 and 9 also show the mean number of items administered for the Full Scale and CAT conditions of each model for each data set. As you might expect, the Full Scale conditions all administered 39 items since this was the item total. The results indicate that the mean number of items administered for all four CAT conditions were less than half the total 39 (ranging 13.595 to 15.378). Figures 14 through 17 display the plots of the mean number of items administered against estimated theta for each CAT condition. Since each Full Scale condition administered exactly 39 items, graphs are not shown for these conditions. As the graphs demonstrate, the most items are administered at the higher levels of theta. As discussed previously, this is attributable to the lower amount of information available at the higher levels of theta in each data set. These results are discussed in the context of each data set.

ADCOM DATA

Figures 14 and 15 display the plots of the mean number of items administered against estimated theta for the ARSM and SIM in the ADCOM data CAT condition, respectively. Although at levels of theta above 1, the ARSM and SIMs both administer the 20 item limit imposed on the CAT program, the graph shows that for levels of theta where there is high information the SIM needs fewer items than the ARSM to estimate theta. When the means of the number of items administered from Table 8 are compared, the SIM (14.335) needs one full item less than the ARSM (15.378) to estimate theta.

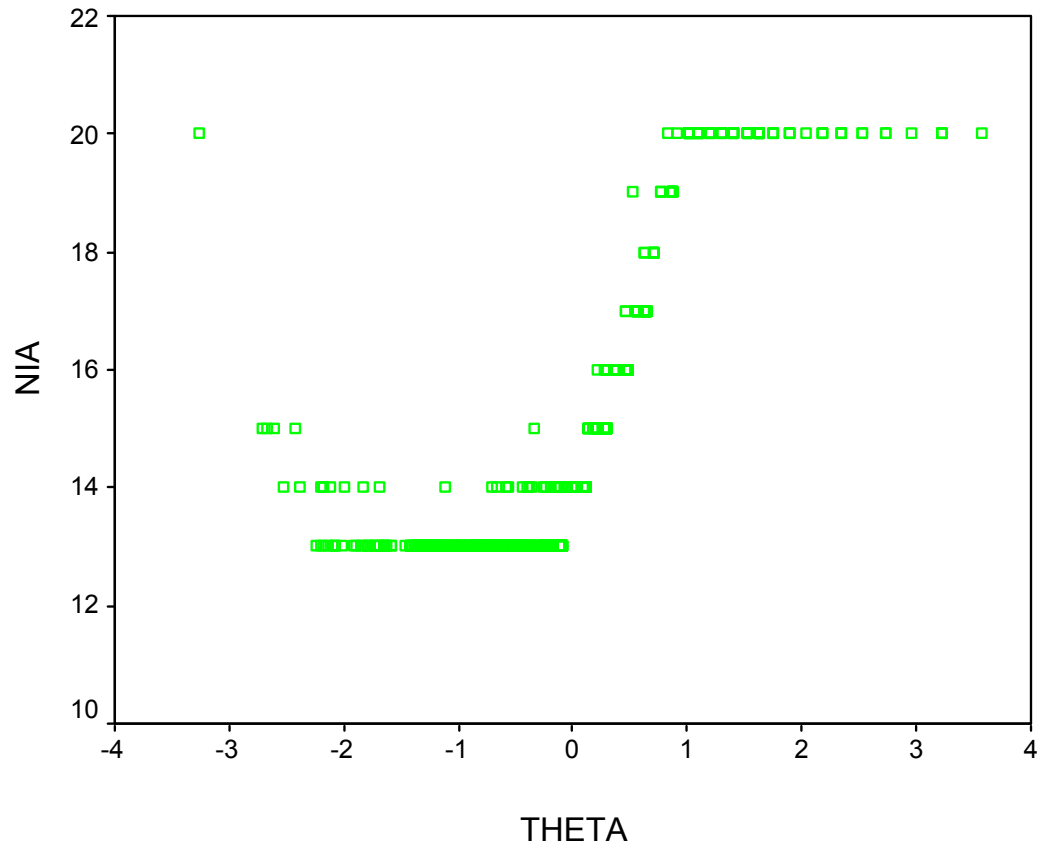


Figure 14: Number of Items Administered for ARSM within the CAT condition of the ADCOM data set

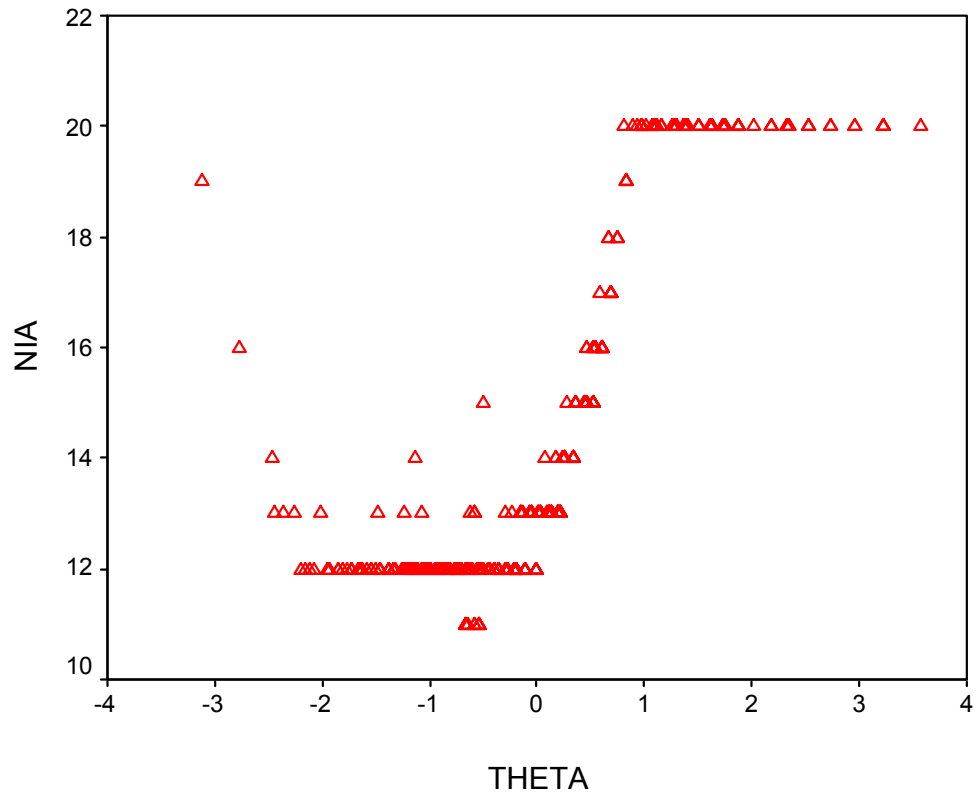


Figure 15: Number of Items Administered for SIM within the CAT condition of the ADCOM data set

SIMULATED DATA

Figures 16 and 17 display the plots of the mean number of items administered against estimated theta for the ARSM and SIM in the Simulated data CAT condition, respectively. These results follow the same pattern as in the ADCOM data set, with the 20 item limit administered at higher levels of theta where less information is available. Also consistent with the ADCOM data, the SIM needs fewer items to estimate theta at the levels of theta where high information is available. When the mean number of items displayed in Table 9 are compared, the SIM (13.595) needs almost one full item less than the ARSM (14.545) to estimate theta.

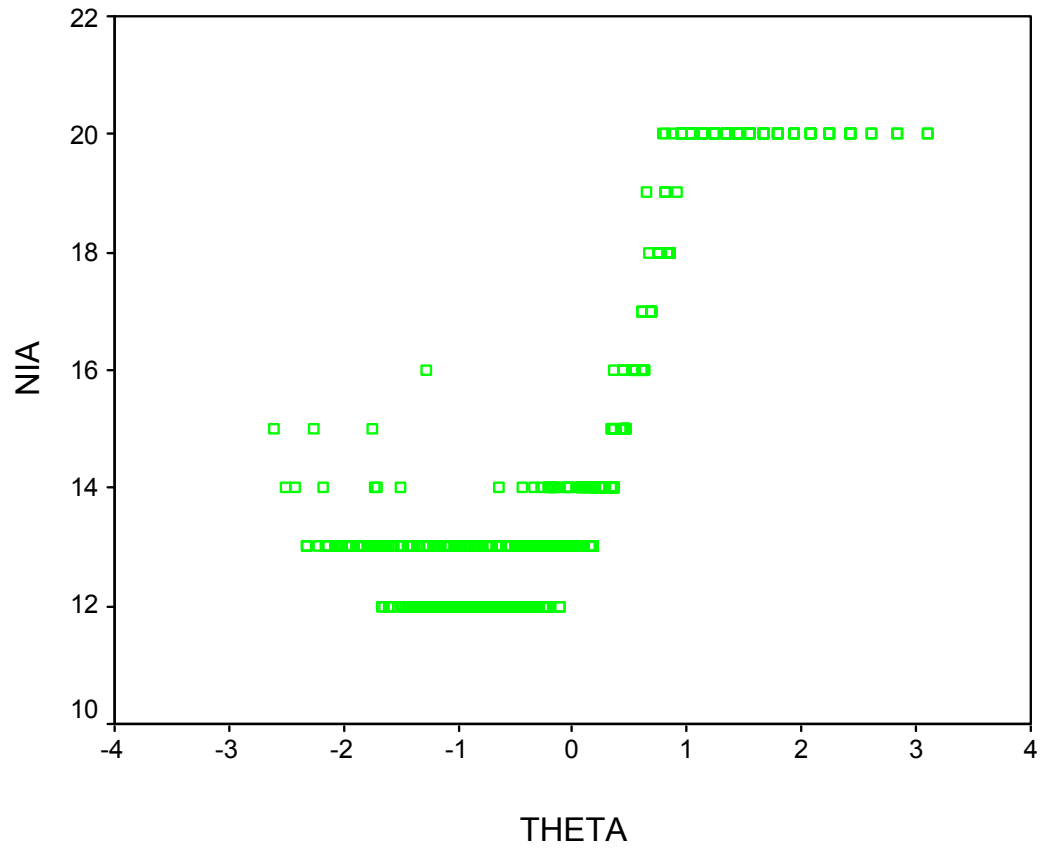


Figure 16: Number of Items Administered for ARSM within the CAT condition of the simulated data set

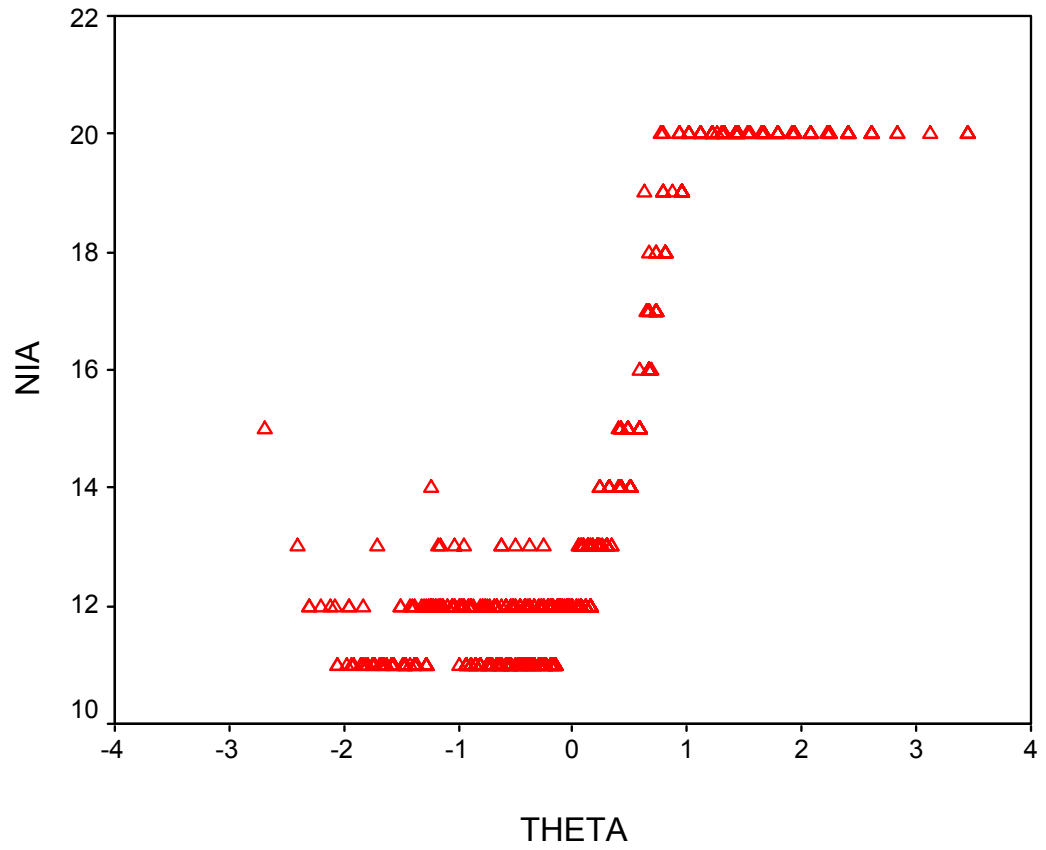


Figure 17: Number of Items Administered for SIM within the CAT condition of the simulated data set

RMSE and Bias

Table 12 shows the RMSE and bias values for simulated and ADCOM Data for the ARSM and SIM. The simulated data includes two RMSE and Bias values for each model: one for estimated theta versus known theta and one for Full Scale theta versus CAT theta. The ADCOM data values are for Full Scale theta versus CAT theta for each model. In all three conditions, the RMSE values are slightly higher for the SIM than for the ARSM. Bias values are slightly better for the SIM in all cases. Theta is slightly underestimated in all cases except the SIM estimated theta versus known theta comparison.

Difference Plots

Difference plots indicate that both the SIM CAT and ARSM CAT approximate the full scale condition in both the ADCOM and simulated data sets. The ADCOM difference plots are represented in Figures 18 and 19 and the simulated difference plots are represented in Figures 20 and 21. The concentration of data appears within .5 of the estimate in either direction with almost all appearing within 1.0 of the estimate. In both the ADCOM and simulated conditions, the ARSM appears to more closely estimate full scale theta than the SIM. This is consistent with Tables 10 and 11 showing that the Pearson product moment correlation between full scale theta and CAT theta for the ARSM within each data set is slightly higher than for the SIM.

Table 12: RMSE and Bias values for ARSM and SIM for simulated Data and ADCOM Data.

	ARSM		SIM	
	RMSE	Bias	RMSE	Bias
ADCOM DATA: Full Scale theta vs. CAT theta	0.295	-0.027	0.347	-0.018
Simulated DATA: known theta vs. CAT theta	0.329	-0.015	0.377	0.003
Simulated DATA: Full Scale theta vs. CAT theta	0.236	-0.029	0.283	-0.025

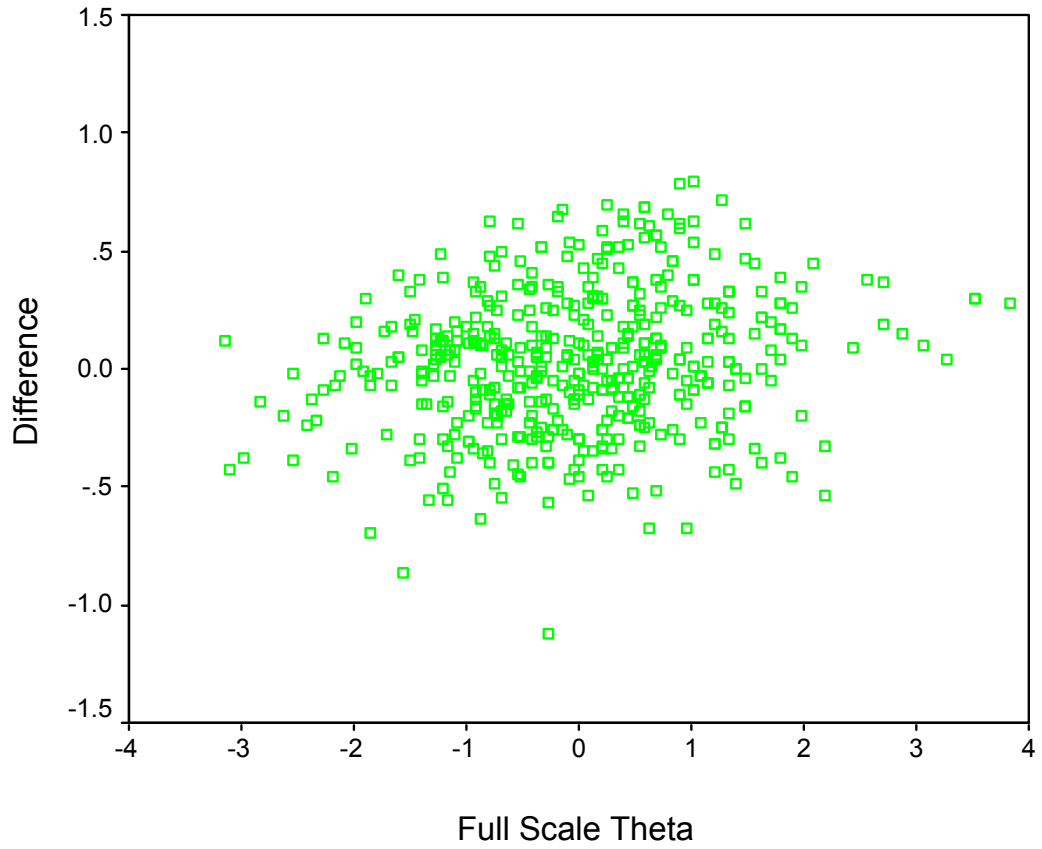


Figure 18: Difference between Full Scale theta and CAT theta for the ARSM with the ADCOM data set.

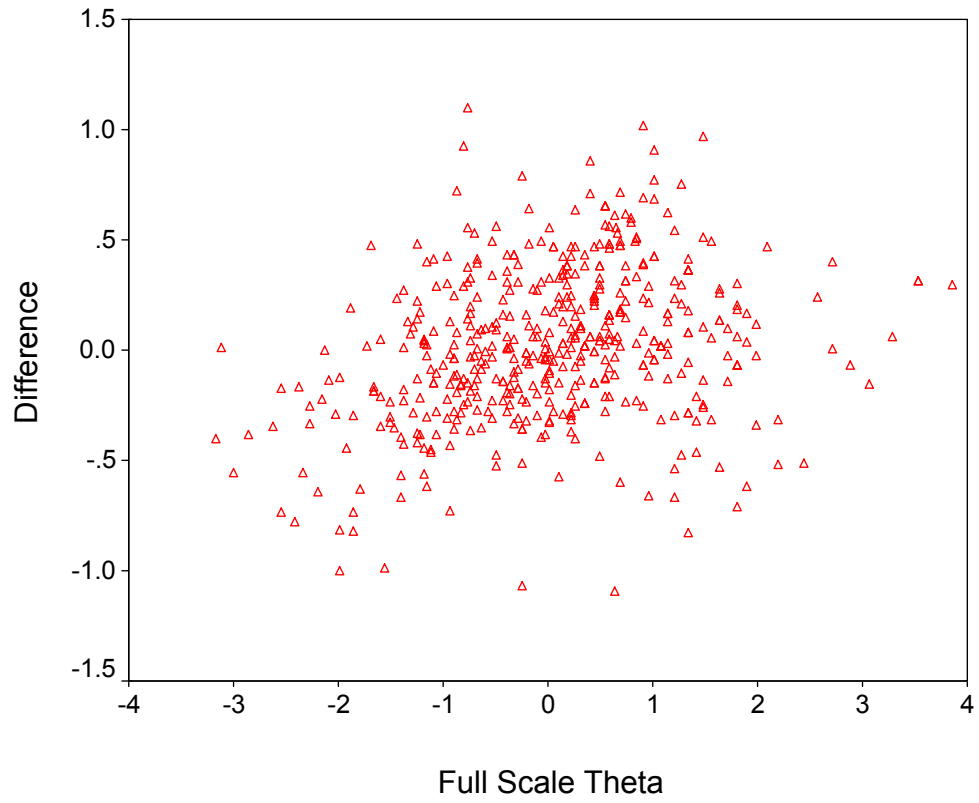


Figure 19: Difference between Full Scale theta and CAT theta for the SIM with the ADCOM data set.

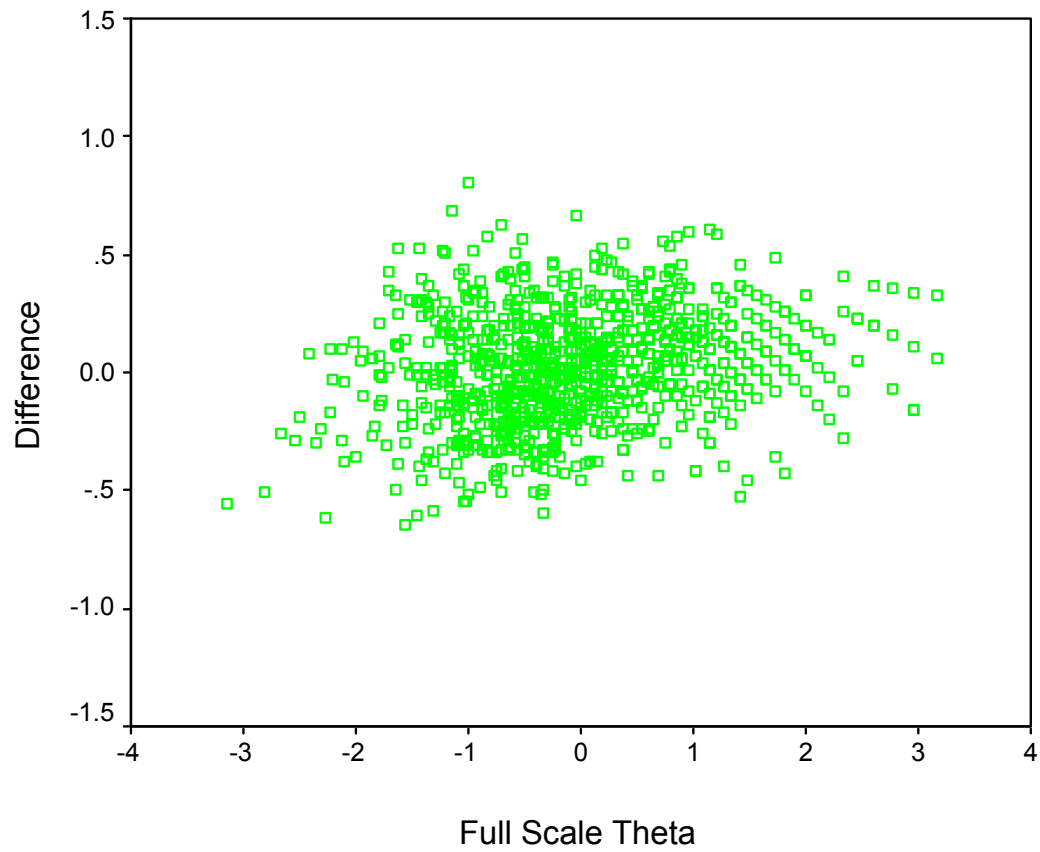


Figure 20: Difference between Full Scale theta and CAT theta for the ARSM with the simulated data set.

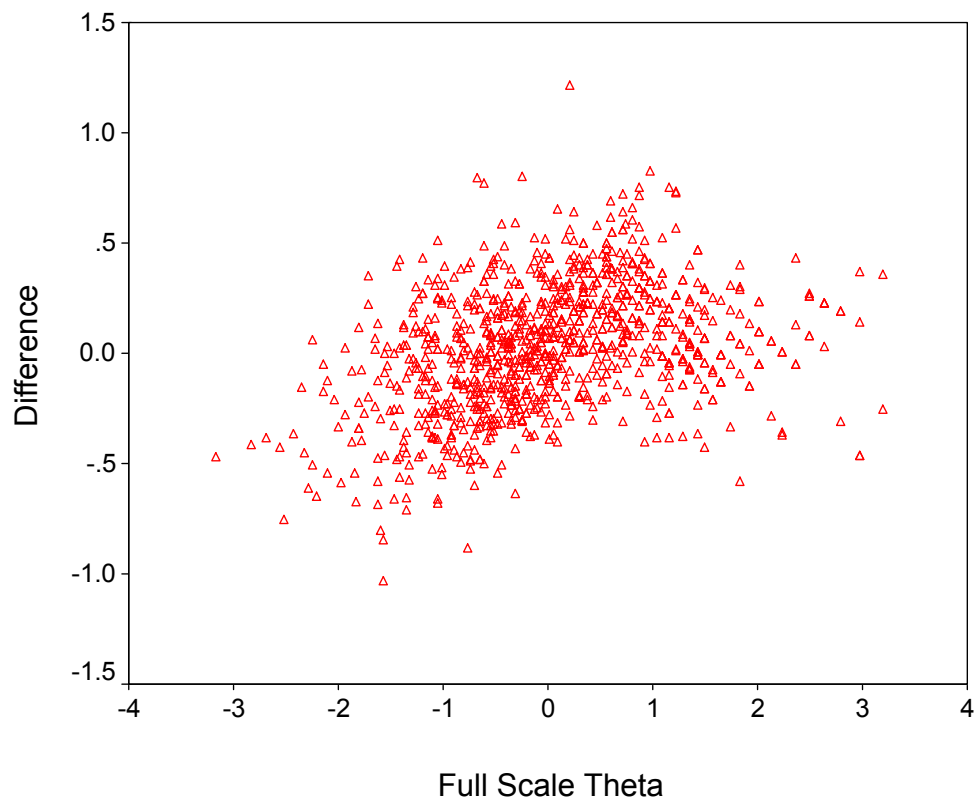


Figure 21: Difference between Full Scale theta and CAT theta for the SIM with the simulated data set.

CHAPTER VI

DISCUSSION

Andrich's rating scale model and Rost's successive intervals model were compared in the context of a computer adaptive test. Results were consistent across the two data sets with the data displaying the same pattern of results for each model in terms of the theta estimate, number of items administered and standard error. This discussion focuses on the results in the context of a CAT as the primary purpose of this study was to identify which model to select when administering an attitude scale via a CAT. Results did not clearly favor one particular model in all comparisons. ARSM and SIM differences were primarily evident in the estimation of theta and in the number of items administered.

The ARSM performed slightly better than the SIM in the performance of the CAT in the context of theta estimation. The ARSM CAT had a slightly higher Pearson product-moment correlation coefficient with Full Scale theta in both the ADCOM and simulated data sets. In the ADCOM data set, the Pearson product-moment correlation between the Full Scale and CAT conditions for the ARSM was .964 compared to .950 in the simulated data set. Similarly, when the simulated data set Pearson product-moment correlations are compared, the ARSM is .973 compared to .961 for the SIM. The Pearson product-moment correlation between CAT and known theta is also available with the simulated data set and the findings are consistent. The correlation between the known theta and CAT theta for the ARSM is .945 compared to .927 for the SIM CAT.

The theta estimate results are supported by the RMSE results. The RMSE was calculated for two conditions in the simulated data set (known theta versus CAT theta, Full-scale theta versus CAT theta) and for one condition in the ADCOM data set (full scale theta versus CAT theta). In all three conditions, the RMSE is lower for the ARSM than the SIM providing supporting evidence that the ARSM theta estimate outperforms the SIM theta estimate.

In contrast, the SIM CAT outperformed the ARSM CAT in terms of number of items administered. In the ADCOM data set, the SIM administered an average of 14.335 items for the CAT and the ARSM administered an average of 15.378 items for the CAT. Similar results were obtained from the simulated data set. The SIM only needed an average of 13.595 items to estimate theta while the ARSM needed an average of 14.545 to estimate theta.

Although they are similar, an interesting result occurred in the standard deviations of the theta estimates. The ARSM and SIM CAT standard deviations of estimated thetas are smaller than the Full Scale standard deviations of estimated thetas. The Full Scale standard deviations are higher than the CAT standard deviations because of the lack of information for higher theta levels. The Full Scale administers all scale items to each respondent, including items that only provide little information for a particular respondent. The CAT selects the items with the most information at the current theta estimate and administers those items until the CAT is completed. However, because of the lack of information at higher levels of theta, a Full Scale respondent at this higher level of theta is administered an entire set of items that provide little information about

their attitude level. In the CAT, these respondents are only administered the 20 item limit.

Outside of the slight differences in the number of items administered and the correlations between the estimated thetas already discussed, the results from the CAT for the two models were quite similar. The item pool, study design, and the actual items provide possible explanations as to why more differences were not evident in the results for each model.

As noted in the results section, the ideal total test information function for a CAT would have a uniform distribution. A uniform distribution allows the CAT to administer appropriate items to respondents at all theta levels as the information is spread across all theta levels. In this study, all total information functions had more information available at lower levels of theta than higher levels of theta. The reason for the similarities between the information functions of the SIM and ARSM is that all data used in this study was essentially a form of the ADCOM data. Recall that this study used a factor loading matrix as an input into the data generation procedure when creating the simulated data set. The factor loading matrix was obtained from a principal axis factor analysis of the ADCOM data set. The ADCOM data factor loadings were used to ensure the data reflected real attitude data. With the ADCOM data, comparisons were between the CAT estimate and the Full Scale estimate of an attitude. The simulated data set was included in this study so that there could be a comparison to known theta.

These peaked information functions restricted the CAT in its ability to estimate theta for respondents outside the area where information was peaked.

This is evident in the number of items administered by each model for each data set displayed in Figures 14 –17. These figures show that the CAT is administering 20 items at these higher levels of theta. This is because the CAT is not reaching the standard error stopping rule of .3. In retrospect, it might have been more interesting to use factor loadings from a data set with a relatively flat test information function. This would have created a data set with items available across the entire theta scale.

Another design tweak that might have created a better opportunity to find differences between the SIM and ARSM would have been to use a different standard error stopping rule. The current study implemented a standard error stopping rule of .3 because that is the most often used standard error in CAT research. It is interesting to consider how the outcomes of this study would have changed had a standard error of .2 or .4 been selected. A lower standard error would likely have increased the mean number of items administered. The restricted test information function suggests that lowering the standard error stopping rule would result in more CATs stopping because of the 20 item limit and less stopping because of obtaining this lower standard error. A higher standard error stopping rule would likely have the opposite effect, with more CATs stopping because of the standard error being obtained and less because of the 20 item limit. However, due to the lack of information at the high theta levels, there would still be a large number of examinees/simulees receiving the 20 item limit.

The makeup of the actual item pool also could influence the effectiveness of each model's performance. The ADCOM survey was designed with a Likert scale. Rost (1988) indicates that the SIM and ARSM provide an interesting contrast in analyzing rating scale data because of their assumptions about the thresholds in a rating scale. The ARSM assumes constant distances between the thresholds for all items. The SIM allows the thresholds for each item to vary proportionally from the mean scale by including a dispersion parameter. It is interesting to consider the impact of the item wording on the effectiveness of these two models. Items can be negative toward the attitude or positive toward the attitude. An example of an item measuring attitudes about college athletics that is negative toward college athletics is "The increased importance of college athletics has lead to a decline in the university educational system". This same thought could be represented in an item that is positive toward college athletics. An example of an item that is positive toward college athletics is "College athletics is one of our societies greatest inventions". An item bank could include all negatively worded items, all positive worded items, or a combination of negative and positively worded items. The results from this study suggest both the SIM and ARSM would perform well as long as the item banks were consistent, regardless of whether they were all negative or all positive. The interesting variation would be in an item pool with a combination of negative and positive worded items. This appears to be a situation where the SIM would be most appropriate because of the dispersion parameter.

The results of this study suggest that future research is warranted with both the ARSM and SIM. The model best suited for use may depend on the actual research being conducted and the importance of the decision being made with the data. Thus, the model to recommend depends on the purpose of the research. For this discussion, research will be divided into high stakes research and low stakes research. Medical research is an example of high stakes research. Medical research is an area where the focus is placed on the absolute accuracy of the information. Attitude measurement is important to medical research studies when considering the effect of certain treatments for patients. For example, a patient's attitude about their quality of life is an important aspect of the overall effectiveness of that specific treatment. An example of low stakes research is market research, where the information is intended to allow the user to make a more informed decision. In market research, the focus is often on the amount of information that can be obtained over the absolute accuracy of the information collected.

The ARSM may be the more appropriate model when the stakes of the research are high because it is a better estimate of Full Scale theta. As discussed, the ARSM CAT theta estimate is more highly correlated with Full Scale theta within each data set. The SIM is the more appropriate model when the stakes of the research are low because it requires fewer items to estimate theta. The SIM outperformed the ARSM in the number of items needed to provide a theta estimate. As mentioned above, research is not always an endeavor in which the outcome is as critical as in the medical field. The SIM may to be the better model

when lower stakes decisions are being made and other factors, including the time it takes to administer a survey, are given a larger role in the consideration of how to conduct the survey.

The next step is to pilot each of these models in a real world setting. As discussed, it would make most sense to pilot the ARSM in the field of medical research and the SIM in the field of market research. There should be ample opportunities to introduce CAT measurement of attitudes in each of these fields. The medical field is continuing to show interest in quality of life issues associated with the treatment of patients. Quality of life is a construct that lends itself to attitude measurement. Similarly, market researchers are always interested in providing more information about the respondents that they measure. The ability to measure attitudes with fewer items would allow market researchers to further define their respondents and possibly provide an attitudinal explanation for differences that exist between measured groups. For example, a financial institution might be interested in differentiating the financial accounts they offer to a particular segment of customers based on the attitudes of these customers. This financial institution could administer a survey identifying the attitudes toward proactive communication of their customers and could offer different types of service to customers with positive attitudes toward proactive communication then they would to customers with negative attitudes toward proactive communication. The service for customers with positive attitudes toward proactive communication could involve various methods of interacting with the customer to serve their account. The service for customers with negative

attitudes toward proactive communication could involve ensuring that these customers are aware of how to contact a person within the financial institution when needed and involve less independent reaching out to them.

In conclusion, both the ARSM and SIM rating scale models are viable models to use in real world research settings. The model to select should depend on the use of the results. As a starting point, the ARSM may be the model to select if the goal is to obtain a CAT theta estimate that will more closely mirror the Full Scale theta estimate. The SIM may be the better model to use if the goal is to obtain a CAT estimate with as few items as possible.

REFERENCES

- Andrich, D. (1978a). A rating formulation for ordered response categories. Psychometrika, *43*, 561-573.
- Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, *2*, 581-594.
- Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, *37*, 29-51.
- Bock, R.D., & Jones, L.V. (1968). The measurement and prediction of judgment and choice. San Francisco: Holden-Day.
- Chang, H. H., & Ying, Z. (1999). A-stratified multi-stage computerized adaptive testing. Applied Measurement in Education, *23*(3), 211-222.
- Chen, S., Hou, L., & Dodd, B.G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation on CAT using the partial credit model. Educational and Psychological Measurement, *53*, 61-77.
- Chen, S., Hou, L., Fitzpatrick, S.J., & Dodd, B.G. (1997). The effect of population distribution and methods of theta estimation on CAT using the rating scale model. Educational and Psychological Measurement, *57*, 61-77.
- Dodd, B.G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. Applied Psychological Measurement, *14*, 355-366.
- Dodd, B.G., & De Ayala, R.J., (1994). Item information as a function of threshold values in the rating scale model. In M. R. Wilson (Ed.) Objective measurement: Theory into practice, Vol 2, 299-315. Norwood NJ: Ablex.
- Dodd, B.G., De Ayala, R.J., & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. Applied Psychological Measurement, *19*, 5-22.

- Dodd, B.G., Koch, W.R., & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. Applied Psychological Measurement, 13, 129-143.
- Edwards, A.L., & Thurstone, L.L. (1952). An interval consistency check for scale values determined by the method of successive intervals and the method of graded dichotomies. Psychometrika, 17, 169-180.
- Fitzpatrick, S.J., Choi, S.W., Chen, S.K., Hou, L., & Dodd, B.G. (1994). IRTINFO: A SAS macro program to compute item and test information. Applied Psychological Measurement, 18, 390.
- Gorin, J.S., Dodd, B.G., Fitzpatrick, S.J., & Shieh, Y.Y. (2000). Computerized adaptive testing with the partial credit model: Estimation Procedures, Population Distributions, and item pool characteristics. Paper presented at the annual meeting of the AERA, New Orleans, LA.
- Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing.
- Koch, W.R. (1983). Likert scaling using the graded response latent trait model. Applied Psychological Measurement, 7, 15-32.
- Koch, W.R., & Dodd, B.G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. Applied Measurement in Education, 2(4), 335-357.
- Koch, W.R., & Dodd, B.G. (1995). An investigation of procedures for computerized adaptive testing using the successive intervals Rasch model. Educational and Psychological Measurement, 55, 976-990.
- Koch, W.R., Dodd, B.G., & Fitzpatrick, S.J. (1990). Computerized adaptive measurement of attitudes. Measurement and Evaluation Counseling and Development, 23, 20-30.
- Lord, F.M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.
- Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and their parallel forms reliability. Psychometrika, 48, 233-245.
- Lord, F.M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

- Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, *47*, 149-174.
- Meijer, R.R., & Nering, M.L. (1999). Computerized Adaptive Testing: Overview and Introduction. Applied Psychological Measurement, *23* (3), 187-194.
- Muraki, E., & Bock, R.D. (1993). The PARSMCALE computer program [Computer program]. Chicago, IL: Scientific Software International.
- Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. Applied Psychological Measurement, *14*, 59-71.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, *16*, 159-176.
- Pastor, D.A., Dodd, B.G., & Chang, H.H. (2002) A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. Applied Psychological Measurement, *26* (2), 147-163.
- Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. Applied Psychological Measurement, *12*, 397-409.
- Samejima, F (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No 17.
- Singh, J., Howell, R. D. & Rhoads, G. K. (1990). Adaptive designs for Likert-type data: An approach for implementing marketing surveys. Journal of Marketing Research, *27*, 304-321.
- Singh, J., Rhoads, G. K. & Howell, R. D. (1992). Adapting marketing surveys to individual respondents. Journal of the Marketing Research Society, *34* (2), 125-147.
- Stocking, M.L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W.J. van der Linden & C.A.W Glass (Eds.), Computerized adaptive testing: Theory and practice (pp. 163-182). Dordrecht. The Netherlands: Kluwer Academic.
- Sympson, J.B., & Hetter, R.D. (1985, October). Controlling item exposure rates in computerized adaptive testing. Paper presented at the annual meeting of

the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Thissen, D.J., & Steinberg, L. (1986). Taxonomy of item response models. Psychometrika, *49*, 501-519.

Valentine, R.J. (1978). Audit of administrators communication. Columbia, MO: Jerry W. Valentine.

Wainer, Howard. (Ed). (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, *54*, 427-450.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computer adaptive testing. Applied Psychological Measurement, *25*(4), 317-331.

Way, W.D. (1998). Protecting the integrity of computerized testing item pools. Educational Measurement: Issues and Practice, *17*(4), 17-27.

Wherry, R.J., Sr., Naylor, J.C., Wherry, R.J., Jr., & Fallis, R.f. (1965). Generating multiple samples of multivariate data with arbitrary parameters. Psychometrika, *30*, 303-313.

VITA

Michael John Lustina was born in Indiana on August 7, 1973, the son of Steve and Angie Lustina. After finishing his work at Andean High School in 1991, he entered Wabash College and majored in Psychology. In May, 1995, he received the Bachelor of Arts degree from Wabash College. In August, 1995, he entered Graduate School of the University of Texas at Austin. In August, 1998, he received the degree of Master of Arts in Educational Psychology from the University of Texas at Austin with a focus on program evaluation.

Permanent address: 5903 Lost Horizon Drive, Austin, Texas 78759

This dissertation was typed by the author.