

Basic Computer Literacy Assessment Tool: Item Response Analysis

*An Appendix to **Basic eSkills:
Foundation or Frustration: A
Research Study of Entering
Community College Students'
Entering Computer Competency***

By: George Rezendes, PhD
Project Director: Diane J. Goldsmith, PhD

Funded by the Alfred P. Sloan Foundation
May 2006

CONNECTICUT DISTANCE LEARNING CONSORTIUM
W W W . C T D L C . O R G

Basic Computer Literacy Assessment Tool— Item Response Analysis

An Appendix to:

Basic eSkills—Foundation or Frustration: A Research Study of Entering Community College Students' Computer Competency

By

George Rezendes, PhD

Director of Research and Assessment, at Three Rivers Community College

**Project Director: Diane J. Goldsmith, PhD
Dean of Planning, Research, and Assessment**

This report on the Item Response Analysis of the Basic Computer Literacy Assessment Tool, which was piloted as part of this research project, is an Appendix to the report entitled *Basic eSkills—Foundation or Frustration: A Research Study of Entering Community College Students' Computer Competency*. That report is available from the Connecticut Distance Learning Consortium.

This research project was conducted by the Connecticut Distance Learning Consortium under the direction of Dr. Diane J. Goldsmith. The project was funded by a Grant from the Alfred P. Sloan Foundation. The simulation software used to develop the test was provided by Course Technology, a division of Thomson Learning.

For more information please contact:

Dr. Diane J. Goldsmith
Connecticut Distance Learning Consortium
85 Alumni Drive
Newington, CT 06111
860.832.3893
dgoldsmith@ctdlc.org

TABLE OF CONTENTS

TABLE OF CONTENTS.....	3
INTRODUCTION	4
METHODOLOGY	4
Item Response Theory.....	4
The Model	4
Model Assumptions.....	6
Model Estimation	6
Evaluating Model Fit.....	7
Item Characteristics.....	9
Student Characteristics.....	13
REFERENCES	16
APPENDICES	17
Appendix A Estimation Methodology	17
Appendix B. Results of the analysis to support the monotonicity assumption.....	22
Appendix C. Tetrachoric Correlation Coefficients by Ability Group.....	23
Appendix D. Standardized Residuals	25

INTRODUCTION

A basic computer literacy assessment consisting of 17 tasks was administered to 2090 students who had applied for admission to one of five Connecticut Community Colleges (Manchester, Middlesex, Naugatuck Valley, Three Rivers and Tunxis) for the Fall 2005 semester. A total of 1459 of those students completing the assessment actually registered for classes at one of the community colleges for the Fall 2005 semester. This analysis will utilize item response theory to estimate the item characteristics of the assessment as well as basic computer literacy ability characteristics of the students. The characteristics of both the assessment and students are then examined relative to student demographic information.

METHODOLOGY

Item Response Theory

Item Response Theory (IRT) is an alternative testing theory that models individual responses at the item level thus overcoming several of the identified shortcomings of the more traditional testing theory known as Classical True Score Theory (CTST). Specifically, IRT provides item characteristics that are not group-dependent, scores reflect individual ability that are not test-dependent, an assessment of score reliability which is not dependent on parallel tests, and a measure of precision for each ability score. A significant advantage of IRT is the independence of item characteristics and student characteristics. In practical terms this allows the researcher to separate the difficulty of the assessment from the ability of the students. When utilizing CTST the assessment characteristics and student characteristics are always dependent on one another so researchers are unable to determine if a difficult assessment is a result of a group of low ability students or a group of low ability students is the result of a difficult test. IRT overcomes this difficulty. In fact, one of the checks to evaluate IRT model fit is to separate the assessment items into two groups by difficulty and to then estimate student ability utilizing both the group of easy items and the group of hard items showing that no significant difference exists between the two ability estimates obtained.

The Model

IRT modeling of examinee responses on the basic computer literacy assessment will be investigated by fitting the two-parameter logistic model to the student response data. The two-parameter logistic model is best described by its item characteristic curve (ICC), a mathematical function that relates the probability of getting a correct answer on an item to the ability measured by the test and the characteristics of the item. The two parameter logistic model has an ICC given by the equation:

$$P(\theta) = \frac{e^{1.7a_j(\theta-b_j)}}{1 + e^{1.7a_j(\theta-b_j)}} \quad j = 1, 2, \dots, n \quad \text{where}$$

- $P(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item j correctly, and is an S-shaped curve with values between 0 and 1 over the ability scale.

- b_j is the difficulty parameter for item j . This parameter represents the point on the ability scale where the probability of a correct response is 0.5. This parameter is a location parameter and indicates the position of the ICC on the ability continuum. The larger the value of b_j the more ability an individual will require to have a 50% chance of getting the item correct. Difficult items are located on the upper end of the ability scale and easy items are located on the lower end of the ability scale.
- a_j is known as the item discrimination parameter. Items with higher values of discrimination are more useful for separating respondents into different ability levels than are items with smaller values since the discrimination parameter is proportional to the slope of the ICC at the point b_j on the ability scale. High values of a_j produce ICC's with steep slopes while low values of a_j lead to item characteristic functions that increase gradually as a function of ability.
- n is the number of items on the assessment
- e is a transcendental number (like π) whose value is 2.718...

Figure 1 provides two examples of ICC's with different parameter values. When looking at Figure 1 note that the .5 probability value occurs at the level of difficulty (b value) for each item (-1.0 for item #1 and 1.0 for item #2) on each of the curves. In looking at the slopes of the two ICC's in Figure 1 it is clear that the ICC for item 2 has a steeper slope at the .5 probability value than the ICC for item 1. This fact is reflected in the discrimination (a) values in that item 2 has a larger value ($a=2.0$) than item 1 ($a=0.5$).

An alternative way to write $P(\theta)$ is obtained by dividing both the numerator and denominator by the expression $e^{1.7a_j(\theta-b_j)}$ to produce: $P(\theta) = \frac{1}{1 + e^{-1.7a_j(\theta-b_j)}}$ $j = 1, 2, \dots, n$, a more convenient way to write the ICC (Hambleton et al., 1991).

Model Assumptions

In fitting the two-parameter logistic model to the basic computer literacy assessment responses, standard IRT assumptions will be considered to be appropriate. Namely the assumptions of monotonicity (the probability of correctly completing a task increases as ability increases), unidimensionality (the fact that the items constituting a test must measure a single ability) and local independence (examinees' responses to individual test items are independent of the responses to other items on the test, i.e. examinees' responses to individual items are dependent only on their ability level relative to the construct being measured).

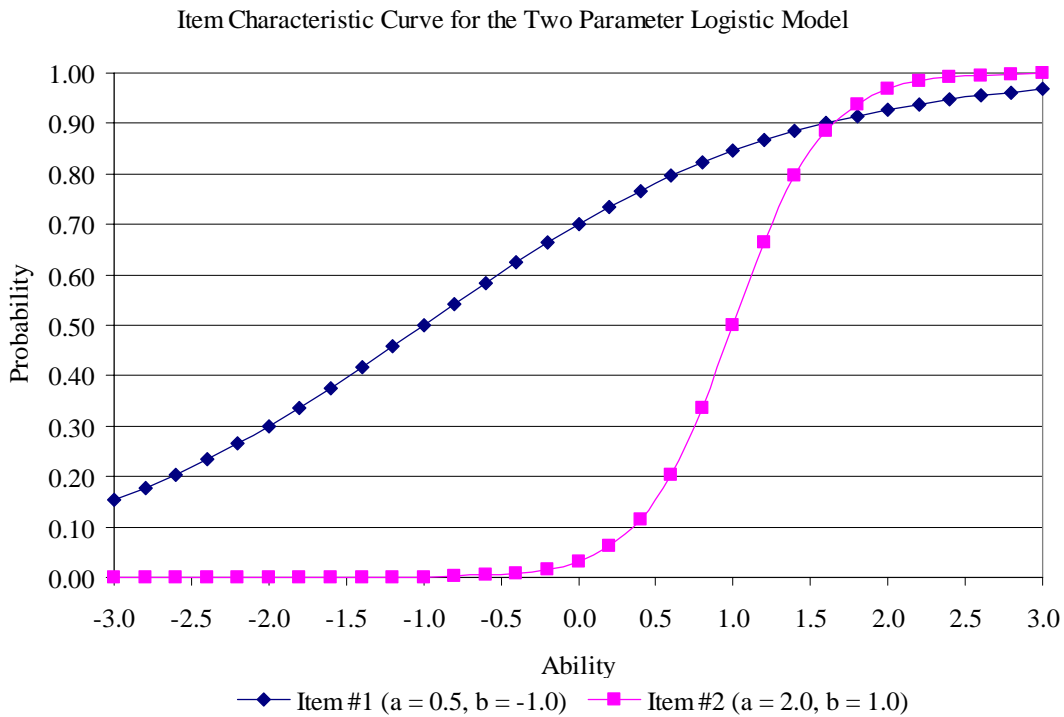


Figure 1.
Example of ICC's for the two parameter logistic IRT model

Model Estimation

Estimates of the IRT model parameters will be calculated by implementing Bayesian estimation techniques with Markov Chain Monte Carlo (MCMC) methods using the *MATHEMATICA* computer algebra system (Wolfram, 1996). MCMC is a general method for the simulation of stochastic processes having probability densities that are known up to a constant of proportionality. In those situations when it is not possible to simulate independent realizations of some complicated stochastic process, dependent realizations X_1, X_2, \dots are used to form an irreducible Markov chain which has a stationary distribution, the same as the distribution of

interest. Simulating a large enough sample of realizations $\dots X_m, X_{m+1}, \dots$ allows the population characteristics of the stationary distribution (i.e. mean, variance, density, etc.) to be calculated to any degree of accuracy using the simulated values (Geyer, 1992).

Patz and Junker (1999a) make the following remark about standard IRT model estimation techniques:

It is difficult to incorporate uncertainty (standard errors) into the item parameter estimates in calculations of uncertainty (standard errors) about inferences for examinees, and there is no way to assess the extent to which standard errors for examinee inferences are overly optimistic because of this.

What is needed is (a) a method of building IRT models that allows for more complete uncertainty calculations, and (b) a method of estimation that remains straightforward as model complexity increases (p. 148).

In response to these remarks, Patz and Junker (1999a) present a MCMC methodology based on Metropolis-Hastings within Gibbs sampling that can easily be implemented to estimate complex IRT models. Details of the estimation methodology are provided in the appendix of this report but in general the estimation methodology creates a realization of the underlying parameter space for the Basic Computer Literacy Assessment IRT model. Multiple replications of the multi-dimensional empirical distribution will then be used to obtain estimates of all item and student parameters along with the associated variance and estimation error.

Evaluating Model Fit

The fit of the IRT model under consideration will be evaluated by presenting three types of evidence. First, the validity of the underlying assumptions will be investigated. Next, evidence establishing the extent to which the expected properties of the IRT model are obtained (e.g., invariance of the ability parameters) will be presented. Finally, evidence establishing the accuracy of the model predictions will be presented. Hambleton and Swaminathan (1985) note that an IRT model that fits the sampled response data well provides evidence of both construct validity and reliability.

Recall that the underlying assumptions for IRT models include monotonicity, unidimensionality, and local independence. Monotonicity assumes that the probability of correctly completing an assessment task increases with ability. Local independence assumes that given any particular ability, responses to individual assessment items are independent of each other. Unidimensionality assumes that the items on the assessment measure a single ability.

The assumption of monotonicity will be investigated by first fitting the IRT model being considered to obtain the individual ability estimates for each respondent. The ability continuum from -3 to 3 will then be divided into 30 equal intervals and the percentage of correct responses for each item within each ability interval will be determined. This item percentage correct by ability level and item provides an estimate of the probability that an individual with the specified ability will respond correctly to the item. Since monotonicity assumes that the probability of correctly completing an assessment task increases with ability, the Spearman rank correlation

coefficient between ability level and percentage correct will be calculated. The Spearman rank correlation coefficient is used because there are no underlying distributional assumptions associated with the methodology. Validity of the monotonicity assumption will be established if a significant correlation close to one is obtained. Conducting the analysis with the Basic Computer Literacy response data fit to with the IRT model the assumption of monotonicity was supported. Spearman rank correlation coefficients for all items were greater than .80 and highly significant, $p < .001$. Details of the analysis are contained in Appendix B.

The assumption of local independence and unidimensionality will be investigated by fitting the IRT model being considered to obtain individual ability estimates. Using the quartiles of the estimated abilities, each person will be placed into one of four like ability groups and then inter-item tetrachoric correlations will be calculated for each ability group. Validity of the local independence and unidimensionality assumptions will be supported if the off diagonal entries of the correlation matrices are found to be close to zero. The tetrachoric correlation coefficient is used for describing the relationship between two dichotomous variables and is conceived as the manifestations of underlying psychological traits which are normally distributed. The tetrachoric correlation coefficient is the product-moment correlation between these two traits (Divgi, 1979). The tetrachoric correlation coefficient will be used in lieu of phi correlations since phi correlations are known to create biased estimates of correlation when used with dichotomous data (Olsson, 1979). The quartiles for student ability were found to be:

Q1: -0.32
(Median) Q2: 0.19
Q3: 0.54

The tetrachoric correlation coefficients were calculated for the four ability groups, high ability (greater than 0.54), mid high ability ($0.19 < \text{ability} \leq 0.54$), mid low ability ($-0.32 < \text{ability} \leq 0.19$) and low ability (less than -0.32). Considering that there are 17 items on the assessment there are a possibility of 136 unique off diagonal correlation coefficients that were calculated for each of the ability groups. In reviewing the tetrachoric correlation coefficients for the high ability group only 35 coefficients (25%) could be calculated due to a lack of variability in the responses. Considering the 35 coefficients that could be calculated only 15 of those coefficients had a magnitude greater than .5 indicating that 90% of the tetrachoric correlation coefficients were less than .5 or could not be calculated. Reviewing the tetrachoric correlation coefficients for the mid high ability group revealed that 49 coefficients (36%) had been calculated, but only 7 coefficients had a magnitude greater than .5. A similar result existed for the mid low ability group, 64 coefficients (47%) had been calculated and only 8 coefficients had a magnitude greater than .5. The tetrachoric correlation coefficients for the low ability group were very different than those for the other three ability groups. One-hundred and thirty-one (96%) of the coefficients had been calculated and 68 (50%) of the coefficients had a magnitude greater than .5. The analysis does suggest that there may be dimensionality concerns for the low ability students. These students may have difficulty reading the questions or understanding the questions and thus the assessment is measuring more than just basic computer literacy. Further study of this issue is warranted. Overall, considering the results from all four ability groups the collective analysis provides sufficient evidence to conclude that the Basic Computer Literacy Assessment is unidimensional since the large majority of off-diagonal correlation coefficients have either not been calculated due to a lack of response variability or have been calculated and have a low

order of magnitude. The tetrachoric correlation coefficients for each ability group are contained in appendix C.

The invariance of the ability parameters will be evaluated by first fitting each of the IRT models under consideration using a randomly selected subset of the data containing 25% of the total response vectors, known as the “sample 1” dataset, to the assessment in order to obtain estimates of the item difficulty and respondent ability parameters. The estimated item difficulties will then be used to divide the assessment responses for another 25% of the observed data, known as the “sample 2” dataset, into two groups, one group containing responses to the difficult items (the upper 50th percentile of items) and the other group containing responses to the easy items (the lower 50th percentile of items). The IRT model will then be fit using both groups of data, (hard items and easy items) to obtain two ability estimates for each respondent, one based on a set of easy assessment items and the other based on a set of hard items. Evidence supporting the invariance of the ability parameter would include: a correlation between the two ability estimates which is close to 1; and the inability to detect any significant difference when using a paired t-test at a .05 significance level. The phi-correlation coefficient between the paired ability estimates was 0.54 and was significant ($p < .01$). The paired t-test was unable to detect a difference in the two ability estimates ($t=1.54, p=0.124$), thus providing evidence of ability parameter invariance.

The accuracy of model predictions will be evaluated through the use of standardized residuals along with the Q_1 chi-square statistic (Yen, 1981) which compares the observed probabilities of getting an item correct to the expected probabilities. Expected probabilities for each item over the ability continuum will be determined by dividing the ability continuum ranging from -3.2 to 2.8 logits into 15 equal intervals of size .4 logits and evaluating the estimated ICC at the midpoint of the each interval. The observed probabilities will be determined by conducting a cross tabulation of correct item responses with individual estimated ability as determined by the fitted model using the same ability intervals determined when calculating the expected probabilities. (Hambleton, et al., 1991). The complete table of standardized residuals for each item across the ability continuum is contained in appendix D. Items 2, 5, 15 and 17 exhibit some areas in the middle of the ability continuum that show considerable lack of model fit as reflect by large residual values. Items 13 and 14 have large residual values at the high end of the ability continuum.

Item Characteristics

Estimates of the item characteristics for the fitted IRT model are summarized in Table 1 with item discrimination estimates in the top section and item difficulty estimates in the bottom section. Table 2 provides a list of basic computer skills (BSC) assessment tasks ordered by item difficulty and shows that the Internet and Email tasks are the more difficulty tasks associated with this assessment. Figure 1 provides the item characteristic curve (ICC) for each of the items contained on the BSC assessment. Recall that the ICC for each item provides the probability that an individual examinee with an ability level of θ will get the item correct. Examining all ICC's for a particular assessment on a single ability continuum provides information about the validity of the assessment in that items should adequately span the ability continuum to satisfactorily distinguish between the ability levels of the respondents. The ICC's will also provide an indication of items that are providing redundant information in that they occupy similar locations

on the ability continuum. An example of two items that appear to be providing redundant information for the BSC assessment are items 1 and 6 which have difficulties of -1.79 and -1.78 and discrimination values of 0.99 and 0.82, respectively. These two tasks, Use the default settings to print the document (item 1) and Apply Bold formatting to the text "50%" in paragraph 1 (item 6) provide redundancy in the assessment and thus one of the tasks may be eliminated to either reduce the time required to complete the assessment or to add an additional task with

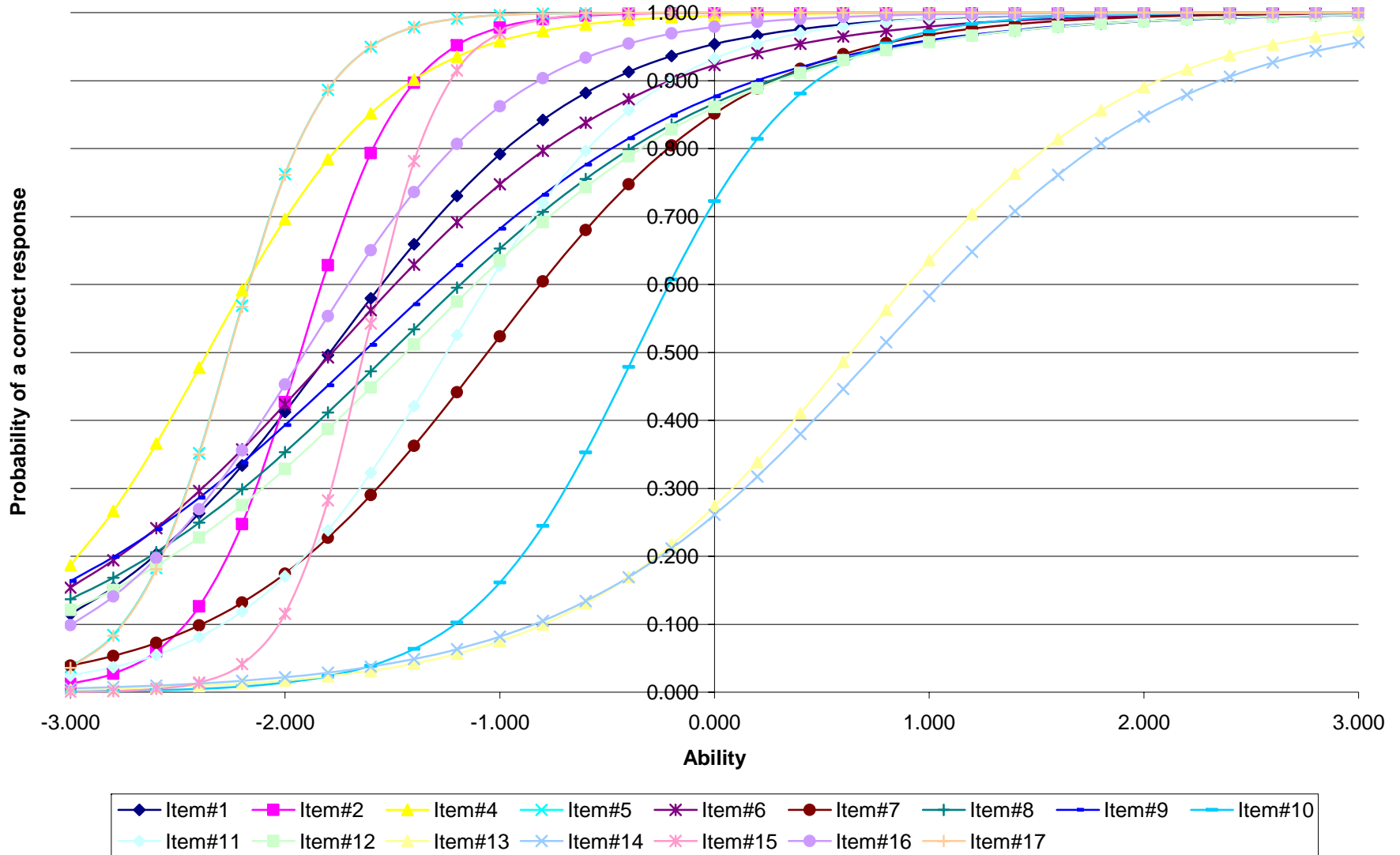
Table 1. Item parameter estimates

Alpha	Mean					Overall	Std Error
(discrimination)	Run#1	Run#2	Run#3	Run#4	Run#5		
item#1	1.004	0.982	0.988	1.002	0.992	0.994	0.0042
item#2	2.421	2.434	2.381	2.409	2.433	2.416	0.0098
item#3	2.371	2.396	2.453	2.390	2.440	2.410	0.0155
item#4	1.349	1.359	1.353	1.355	1.351	1.353	0.0017
item#5	2.582	2.608	2.623	2.601	2.649	2.613	0.0112
item#6	0.824	0.820	0.818	0.823	0.817	0.820	0.0014
item#7	0.969	0.967	0.972	0.975	0.968	0.970	0.0016
item#8	0.727	0.719	0.731	0.734	0.731	0.728	0.0025
item#9	0.710	0.702	0.706	0.705	0.701	0.705	0.0015
item#10	1.547	1.513	1.540	1.527	1.537	1.533	0.0058
item#11	1.249	1.238	1.235	1.244	1.226	1.238	0.0040
item#12	0.754	0.742	0.746	0.750	0.741	0.747	0.0025
item#13	0.887	0.907	0.904	0.903	0.909	0.902	0.0040
item#14	0.801	0.817	0.802	0.816	0.812	0.809	0.0034
item#15	3.136	3.362	3.269	3.230	3.229	3.245	0.0364
item#16	1.188	1.200	1.188	1.189	1.185	1.190	0.0026
item#17	2.610	2.631	2.618	2.598	2.632	2.618	0.0066
Beta	Mean					Overall	Std Error
(difficulty)	Run#1	Run#2	Run#3	Run#4	Run#5		
item#1	-1.766	-1.802	-1.804	-1.781	-1.799	-1.790	0.0073
item#2	-1.739	-1.742	-1.762	-1.754	-1.758	-1.751	0.0044
item#3	-1.914	-1.927	-1.929	-1.934	-1.937	-1.928	0.0039
item#4	-2.346	-2.359	-2.369	-2.352	-2.377	-2.360	0.0057
item#5	-2.254	-2.262	-2.267	-2.262	-2.265	-2.262	0.0022
item#6	-1.752	-1.778	-1.790	-1.782	-1.791	-1.779	0.0071
item#7	-1.037	-1.057	-1.068	-1.058	-1.066	-1.057	0.0055
item#8	-1.506	-1.526	-1.514	-1.497	-1.511	-1.511	0.0049
item#9	-1.616	-1.641	-1.644	-1.635	-1.649	-1.637	0.0056
item#10	-0.356	-0.367	-0.374	-0.365	-0.374	-0.367	0.0034
item#11	-1.229	-1.247	-1.263	-1.242	-1.260	-1.248	0.0061
item#12	-1.419	-1.444	-1.445	-1.427	-1.450	-1.437	0.0059
item#13	0.658	0.639	0.628	0.634	0.624	0.636	0.0060
item#14	0.779	0.754	0.752	0.748	0.749	0.757	0.0056
item#15	-1.621	-1.625	-1.636	-1.634	-1.638	-1.631	0.0033
item#16	-1.902	-1.889	-1.913	-1.910	-1.918	-1.906	0.0051
item#17	-2.250	-2.258	-2.273	-2.264	-2.257	-2.261	0.0039

Table 2. Basic computer literacy assessment tasks ordered by item difficulty

Difficulty	Task Number	Skill set	Activity	Task
-2.36	4	Getting Started with Windows	Open a window by double-clicking	Open the My Computer window by double-clicking. [578]
-2.26	5	Getting Started with Windows	Point to an item	Point to the My Computer icon on the desktop. [571]
-2.26	17	Working with Programs	Close a program	Exit the WordPad program. [576]
-1.93	3	Getting Started on the Internet	Scroll bars	Scroll down to read the description of James Brown's Greatest Hits. [973]
-1.91	16	Working with Files	Open a file from within a program	From within WordPad, open the Sample Text.rtf file. [589]
-1.79	1	Formatting Documents	Print documents	Use the default settings to print the document. (Note: The document will not actually print.) [167]
-1.78	6	Inserting and Modifying Text	Applying character formats	Apply Bold formatting to the text "50%" in paragraph 1. [244]
-1.75	2	Getting Started on the Internet	Go to a Web page by using links	Navigate to the Contact Us page of the Downtown Records Web site.
-1.64	9	Managing Documents	Save a document	Save the current document with the same filename to the "SAM XP" folder. [172]
-1.63	15	Using the Elements of a Web Browser	Finding a previously displayed Web page	Return to the Downtown Records main page by going one page back in the browser. [994]
-1.51	8	Inserting and Modifying Text	Insert text	Insert the word "Northwest" one space after the word "Pacific" in the title text. [101]
-1.44	12	Using e-mail	Open a new e-mail message window	Open a new Untitled Message - Microsoft Word window. [602]
-1.25	11	Searching the Internet	Searching the Web using another search engine	Search for more sites that sell records using Google (www.google.com). Use 'records' as your search term. [1031]
-1.06	7	Inserting and Modifying Text	Copy and paste text	Copy the line containing the text "Paintings and Drawings". Paste it on the line below "Heinrich Strubel". [103]
-0.37	10	Navigating Using the Address bar	Entering a URL in the Address bar	Navigate to the Downtown Records main page by entering the URL in the Address bar. Use the URL www.course.com/downtown_records. [1009]
0.64	13	Using e-mail	Reply to an e-mail message	Create a reply with the text "I completely support this new product line." as a reply to New Products message. Send the reply. [604]
0.76	14	Using e-mail	Send a message with an attachment	Send the Last Year's Product Line document (from the My Documents folder) as an attachment to this e-mail message. [603]

Figure 2. Item Characteristic Curves for Basic Computer Skills Assessment



increased difficulty. An examination of Table 2 shows several sets of items falling in very close proximity relative to difficulty that could be reexamined with an eye for making modifications to the existing BSC assessment to both reduce assessment time and quite possibly add additional difficulty to the assessment.

Student Characteristics

A 50% random sample of respondents (n= 1047) was used to evaluate the student characteristics. A plot of the estimated abilities for these respondents is shown in figure 2. The summary statistics for the ability distribution are displayed in table 3. Of particular note are the quartiles which when used in conjunction with the item characteristic curves in figure 1 can provide a more detailed indication of the skill levels of respondents. For example using Q3 with the ICC for item 13 we can say that 25% of the respondents have at least a .46 chance of responding correctly to item #13, that is they are able to reply to an email message.

Figure 2. Distribution of respondent basic computer skills abilities

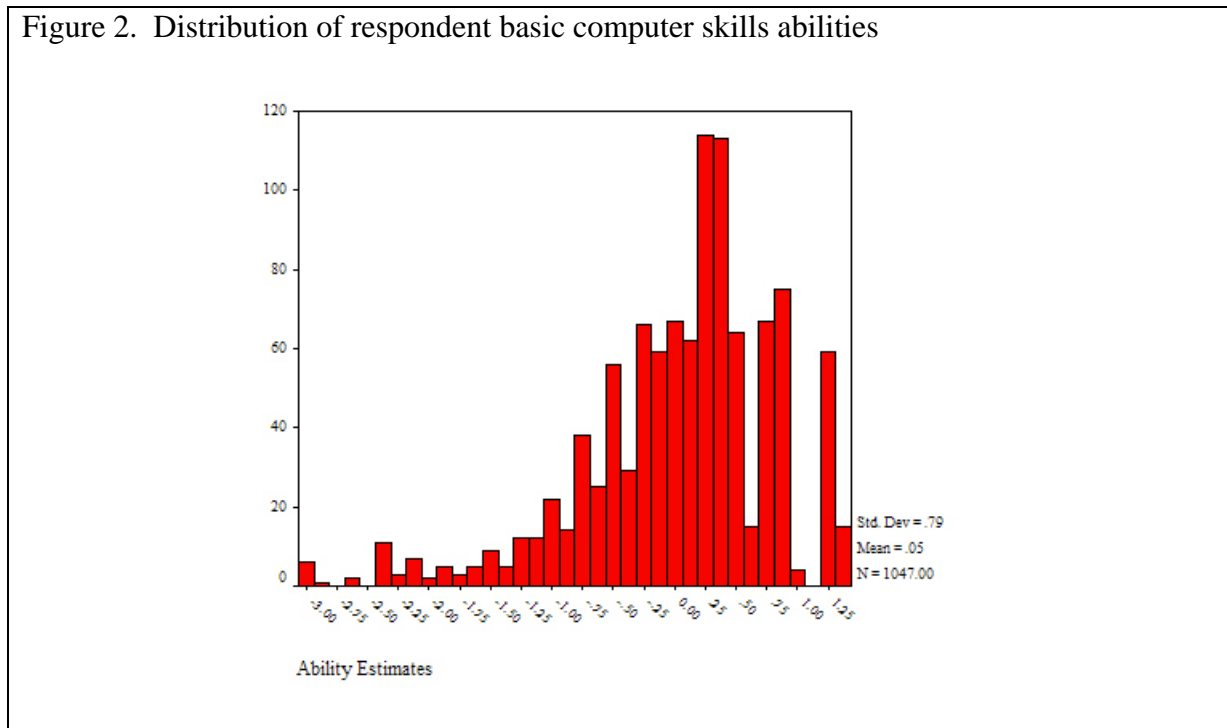


Table 3. Summary statistics for respondent basic computer skill abilities

		Quartiles
n:	1047	Q1: -0.32
Mean:	0.045	Q2: 0.19
Standard Deviation:	0.79	Q3: 0.54

Further analysis of the 1047 respondents contained in selected sample showed that 535 (51.1%) respondents did not register for classes at any of the community colleges during the fall 2005 term while 512 (48.9%) of the selected respondents did have registration records. In an effort to compare the distribution of abilities across the two types of respondents, those that registered and those that did not, the quartiles were used to place the respondents in one of four ability groups. A comparison of the distributions between respondents that registered for classes during fall 2005 and those that did not showed a significant difference in that those students registering for classes during the fall 2005 semester contained more students in the highest ability group and fewer students in the lowest ability group. The mean ability for the group of registered students is also significantly larger than the non registered students ($t=6.19, p <.01$).

Table 4 Comparison of basic computer skills respondents based on registration status

Group	1	2	3	4	Total
Group boundaries	(-3.01,-0.32)	(-0.32,0.19)	(0.19,0.54)	(0.54,1.32)	
Non-registered Students	158 29.5%	138 25.8%	161 30.1%	78 14.6%	535
Registered Students	103 20.1	123 24.0	109 21.3	177 34.6	512

Figure 3. Distribution of basic computer skill abilities for non-registered students

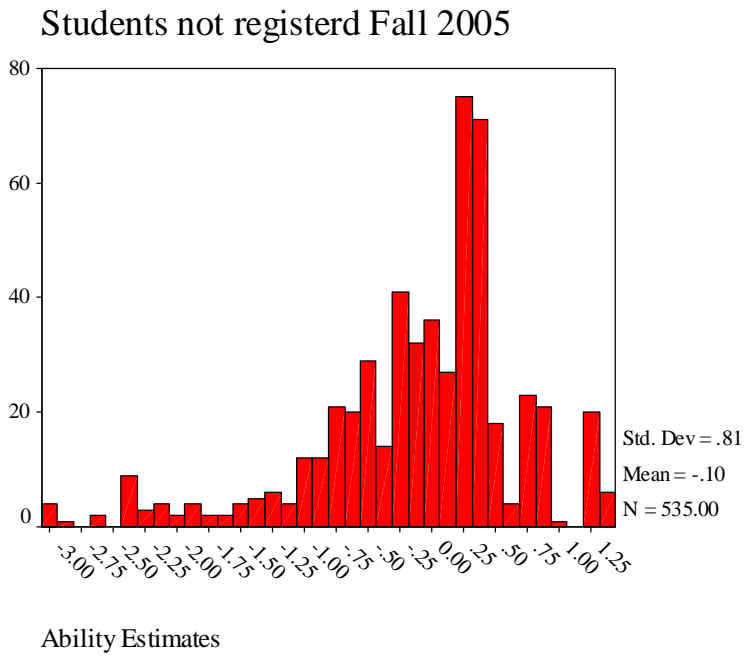
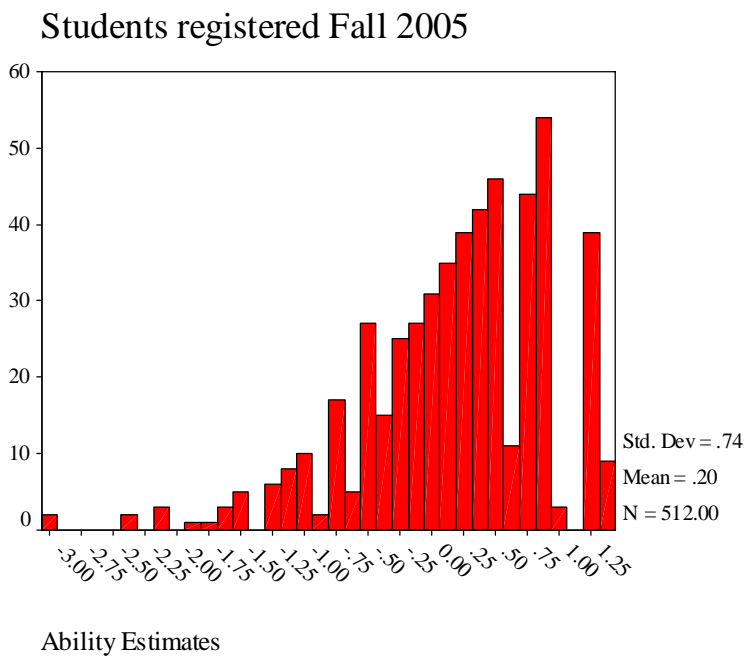


Figure 4. Distribution of basic computer skill abilities for registered students



REFERENCES

- Gelman A., Roberts, G. O., and Gilks, W. R. (1996). Efficient Metropolis jumping rules. In J.M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, pp. 599-608. New York: Oxford.
- Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), pp. 457-511.
- Geyer, C.J., (1992). Practical markov chain monte carlo. *Statistical Science*, 7(4), pp.473-511.
- Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic Publishers, 101 Philip Drive, Norwell Massachusetts, 02061.
- Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*, Newbury Park, CA: Sage Publications.
- Olsson, U., (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), pp.443-460.
- Patz, R.J., & Junker, B.W., (1999a). A straight forward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), pp.146-178.
- Raftery, A.E., & Lewis, S.M., (1992). How many iterations in the gibbs sampler? In J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, (Eds.), *Bayesian Statistics 4*, (pp.763-773). Oxford University Press.
- Wolfram, S., (1996). *The Mathematica® Book*, 3rd ed., Wolfram Media/Cambridge University Press.

APPENDICES

Appendix A Estimation Methodology

The Gibbs sampler. The Gibbs sampler utilizes the complete conditional distributions to obtain characteristics of the joint distribution. Consider the following example provided by Casella and George (1992) to illustrate the Gibbs Sampler.

Suppose X and Y have conditional distributions that are exponential distributions restricted to the interval (0,B), that is,

$$f(x|y) \propto ye^{-yx}, \quad 0 < x < B < \infty \quad (2.8a)$$

$$f(y|x) \propto xe^{-xy}, \quad 0 < y < B < \infty \quad (2.8b)$$

where B is a known positive constant. The restriction to the interval (0,B) ensures that the marginal distributions $f(x)$ and $f(y)$ exists. Although the form of this marginal is not easily calculable, by applying the Gibbs sampler to the conditionals in (2.8) any characteristic of $f(x)$ can be obtained (p. 168).

The Gibbs sampler in this situation is easy to implement since the complete conditionals are known to be proportional to exponential distributions with the means of $1/x$ and $1/y$ respectively. If $X^{(k)}$ and $Y^{(k)}$ represent the k-th realization of the random values X and Y, then implementation of the Gibbs sampler would proceed as follows:

1. Let $X^{(0)} = K$ such that $0 < K < B$ (note: K is an arbitrary starting value)
2. Let $Y^{(0)}$ be equal to a random draw from an exponential distribution with a mean of

$$1/X^{(0)}. \text{ (i.e. Draw } Y^{(0)} \sim p(y|x^{(0)}) \text{)}$$

3. Let $X^{(1)}$ be equal to a random draw from an exponential distribution with a mean of

$$1/Y^{(0)}. \text{ (i.e. Draw } X^{(1)} \sim p(x|y^{(0)}) \text{)}$$

4. Continue to repeat steps 2 and 3 until the chain of random variables has reached a stationary distribution. Note there are several techniques to determine when the Markov chain has converged that will briefly be discussed later in this section.

Suppose that the above Gibbs Sampler was run for 200 iterations to “burn-in” (i.e. have an opportunity to approach the stationary distribution with the generated data being discarded) and then 500 iterations to sample from the joint distribution (X,Y). The 500 (X,Y) pairs generated would then be used to calculate population characteristics (means, variances, density estimates, etc.) for the joint distribution or either marginal distributions. The usefulness of the Gibbs sampler increases dramatically when dealing with high dimensional problems which make

sampling directly from the desired joint probability distribution difficult if not impossible (Casella & George, 1992).

In general when setting up the Gibbs sampler in a situation involving the random variable X , (responses to an assessment) and two parameters θ and β (note both the random variable and parameters could be vectors) the complete conditionals, $p(\theta|X, \beta)$ and $p(\beta|X, \theta)$ are proportional to the joint distribution $p(X, \theta, \beta) = p(X|\theta, \beta)p(\theta, \beta)$ and that fact can be seen by examining the definition of the complete conditional probability distributions which are given as

$$p(\theta|X, \beta) = \frac{p(X|\theta, \beta)p(\theta, \beta)}{\int p(X|\theta, \beta)p(\theta, \beta)d\theta} \text{ and } p(\beta|X, \theta) = \frac{p(X|\theta, \beta)p(\theta, \beta)}{\int p(X|\theta, \beta)p(\theta, \beta)d\beta}.$$

In each case these expressions point out that the complete conditionals are in fact proportional to the joint distribution (the numerator represents the joint distribution $p(X, \theta, \beta)$ in both equations), but these equations also show that in defining the complete conditional distributions while setting up the Gibbs sampler, the normalizing constants $\int p(X|\theta, \beta)p(\theta, \beta)d\theta$ and $\int p(X|\theta, \beta)p(\theta, \beta)d\beta$ must be calculated (Patz and Junker, 1999a). When calculation of the normalizing constant is difficult, implementation of the pure Gibbs sampling algorithm may not be possible.

The Metropolis-Hastings algorithm. In those situations when the complete conditional distribution can only be identified up to a constant of proportionality (i.e. when the normalizing constants can not be calculated) the Metropolis-Hastings algorithm generates a proposed realization that is then accepted or rejected using the concept of rejection sampling. In rejection sampling a candidate realization (θ^*, β^*) is taken from a convenient proposal distribution and the Markov chain will take the step $(\theta^{(k)}, \beta^{(k)}) = (\theta^*, \beta^*)$ with probability α , which is calculated as a function of the proposal distribution and the target distribution. When calculating the acceptance probability α the proposal distribution will drop out of the calculations if it is a symmetric distribution and the acceptance probability will only depend on the target distribution (Patz & Junker, 1999a). Although the Metropolis-Hastings algorithm resolves the problem associated with the Gibbs sampler of having to calculate the normalizing constant, Patz and Junker (1999a) note that when the parameter space is large (as it will be in most IRT models) it is difficult to maintain reasonable acceptance probabilities while still thoroughly exploring the parameter space with a pure implementation of the Metropolis-Hastings algorithm and so it becomes impractical for estimation of IRT models.

The Metropolis-Hastings algorithm within Gibbs sampling. In an effort to take advantage of the benefits of each technique (Metropolis-Hastings and Gibbs sampling) while minimizing difficulties Patz and Junker (1999a) suggest a mixed algorithm which uses the Gibbs sampler to iteratively sample from the complete conditionals except when the complete conditional distribution is known only up to a constant of proportionality, in which case a single iteration of the Metropolis-Hastings algorithm is executed.

In describing the implementation of the Metropolis-Hastings algorithm within Gibbs sampling, the IRT context of this study the standard two parameter logistic model will be considered.

Recall that the standard two-parameter logistic model can be defined by it's ICC's which yields the probability that person i gets the correct response to item j and is given by:

$$p_{i,j}(\theta_i, a_j, b_j) = P(X_{i,j} = 1 | \theta_i, a_j, b_j) = \frac{1}{1 + \exp[-1.7a_j(\theta_i - b_j)]}$$

with a_j as the discrimination parameter for item j, b_j as the difficulty parameter for item j, and θ_i as the ability parameter for person i. Given I persons and J items the likelihood functions for both models are determined by the product of IJ Bernoulli terms as follows:

$$p(X | \theta, \alpha, \beta) = \prod_i \prod_j p_{i,j}(\theta_i, \alpha_j, \beta_j)^{X_{i,j}} [1 - p_{i,j}(\theta_i, \alpha_j, \beta_j)]^{(1-X_{i,j})}$$

Defining independent prior distributions for each of the parameters,

$$p(\theta_i) \sim Normal(0, \sigma_\theta^2)$$

$$p(b_j) \sim Normal(0, \sigma_b^2)$$

$$p(a_j) \sim LogNormal(0, \sigma_a^2)$$

for $i = 1, 2, \dots, I$ (total persons) and $j = 1, 2, \dots, J$ (total items) completes the model specification. Using the aforementioned model specifications, Patz and Junker (1999a) show that the implementation of the Metropolis-Hastings algorithm within Gibbs sampling for the two parameter logistic model at k-th iteration takes the following form:

1. Attempt to draw $\theta^{(k)}$ from $p(\theta | a^{(k-1)}, b^{(k-1)}, X)$:

(a) Draw $\theta_i^* \sim Normal(\theta_i | \theta_i^{(k-1)}, c_\theta^2)$ independently for each $i = 1, 2, \dots, I$. θ_i^* is the proposed value for $\theta_i^{(k)}$, $\theta_i^{(k-1)}$ represents the (k-1) realization of θ_i and c_θ^2 is the generating variance for θ which is used to calibrate the acceptance probabilities.

(b) Calculate the vector of I acceptance probabilities

$$\alpha_i(\theta_i^{(k-1)}, \theta_i^*) = \min\{1, R_{\theta_i}\},$$

where

$$R_{\theta_i} = \frac{p(X_i | \theta_i^*, a_j^{(k-1)}, b_j^{(k-1)}) p(\theta_i^*)}{p(X_i | \theta_i^{(k-1)}, a_j^{(k-1)}, b_j^{(k-1)}) p(\theta_i^{(k-1)})}$$

$$= \frac{\left[\prod_j p_{i,j}(\theta_i^*, a_j^{(k-1)}, b_j^{(k-1)})^{X_{i,j}} \left(1 - p_{i,j}(\theta_i^*, a_j^{(k-1)}, b_j^{(k-1)})^{1-X_{i,j}}\right) \right] \exp\left(-\frac{(\theta_i^*)^2}{2\sigma_\theta}\right)}{\left[\prod_j p_{i,j}(\theta_i^{(k-1)}, a_j^{(k-1)}, b_j^{(k-1)})^{X_{i,j}} \left(1 - p_{i,j}(\theta_i^{(k-1)}, a_j^{(k-1)}, b_j^{(k-1)})^{1-X_{i,j}}\right) \right] \exp\left(-\frac{(\theta_i^{(k-1)})^2}{2\sigma_\theta}\right)}$$

for $i = 1, 2, \dots, I$.

(c) Accept each $\theta_i^{(k)} = \theta_i^*$ with probability α_i ; otherwise let $\theta_i^{(k)} = \theta_i^{(k-1)}$.

2. Attempt to draw $(a, b)_j^*$ from $p((a, b)_j^* | \theta^{(k)}, X)$:

(a) Draw $a_j^* \sim \text{LogNormal}(a_j | a_j^{(k-1)}, c_a^2)$ and $b_j^* \sim \text{Normal}(b_j | b_j^{(k-1)}, c_b^2)$ independently for each $j = 1, 2, \dots, J$. $(a, b)_j^*$ is the proposed value for $(a, b)_j^{(k)}$, $(a, b)_j^{(k-1)}$ represents the $(k-1)$ realization of $(a, b)_j$ and (c_a^2, c_b^2) are the generating variances for (a, b) which are used to calibrate the acceptance probabilities.

(b) Calculate the vector of J acceptance probabilities

$$\alpha_j((a, b)_j^{(k-1)}, (a, b)_j^{(*)}) = \min\{1, R_{(a,b)_j}\}$$

where

$$\begin{aligned} R_{(a,b)_j} &= \frac{p(X_j | \theta^k, (a, b)_j^*) p((a, b)_j^*) \text{LogNormal}(a_j^* | a_j^{(k-1)}, c_a^2)}{p(X_j | \theta^k, (a, b)_j^{(k-1)}) p((a, b)_j^{(k-1)}) \text{LogNormal}(a_j^{(k-1)} | a_j^*, c_a^2)} \\ &= \frac{\left[\prod_i p_{i,j}(\theta_i^k, (a, b)_j^*)^{X_{i,j}} \left(1 - p_{i,j}(\theta_i^k, (a, b)_j^*)^{1-X_{i,j}}\right) \right]}{\left[\prod_i p_{i,j}(\theta_i^{(k-1)}, (a, b)_j^{(k-1)})^{X_{i,j}} \left(1 - p_{i,j}(\theta_i^{(k-1)}, (a, b)_j^{(k-1)})^{1-X_{i,j}}\right) \right]} \\ &\quad \times \frac{\exp\left(-\frac{(b_j^*)^2}{2\sigma_b}\right) \left(\frac{1}{a_j^*}\right) \exp\left(-\frac{(\log(a_j^*))^2}{2\sigma_a}\right)}{\exp\left(-\frac{(b_j^{(k-1)})^2}{2\sigma_b}\right) \left(\frac{1}{a_j^{(k-1)}}\right) \exp\left(-\frac{(\log(a_j^{(k-1)}))^2}{2\sigma_a}\right)} \times \frac{a_j^{(k-1)}}{a_j^*} \end{aligned}$$

(c) Accept each $(a, b)_j^{(k)} = (a, b)_j^*$ with probability α_j ; otherwise let

$(a, b)_j^{(k)} = (a, b)_j^{(k-1)}$. The factor $\frac{a_j^{(k-1)}}{a_j^*}$ in the acceptance probability ratio $R_{(a,b)_j}$

results from the lack of symmetry in the lognormal proposal density (Patz & Junker, 1999a).

In estimating the two parameter logistic IRT model the acceptance probabilities are controlled by the generating variances c_{θ}^2, c_a^2 , and c_b^2 . The generating variances will have to be determined by running several short MCMC chains with various values while monitoring the acceptance rates. Gelman, Roberts, and Gilkes (1996) as well as Patz and Junker (1999a) note that rates of about 50% for univariate draws, and 25% for higher dimensional blocks of parameters are considered to be reasonably efficient acceptance probabilities when using the Metropolis-Hastings algorithm.

Assessing convergence. In using iterative techniques such as the Metropolis-Hastings within Gibbs algorithm for estimation there is always a concern as to when the Markov Chain has reached the stationary distribution. That is, how many iterations are required to achieve a desired level of accuracy? There has been a great deal of discussion in the literature about determining the best approach for monitoring the convergence of the Markov Chain. One approach depends on a single long chain, while other approaches utilize multiple chains with random starting values (Gelman & Rubin, 1992; Geyer, 1992; Patz & Junker, 1999a; Raftery & Lewis, 1992). Patz and Junker (1999a) note, in using MCMC iterative techniques it is important to distinguish between what can be called “statistical uncertainty” from what can be called “Monte Carlo uncertainty” in analyzing obtained estimates. Statistical uncertainty is defined to be a measure of the amount of information contained in the data used to estimate the quantity of interest and is determined as soon as the data is collected. On the one hand, Monte Carlo uncertainty is a measure of uncertainty that arises from the estimation technique being used and can always be reduced by increasing the Monte Carlo sample size (i.e. include more iterations in the Monte Carlo chain or increase the number of chains being used to calculate the estimated quantities).

Convergence of the Monte Carlo chains in this study will be monitored by using five Monte Carlo chains, each with randomly selected starting values and run for 15,000 iterations. The first 5,000 iterations of each chain will be discarded as burn-in. Estimates of model parameters along with the associated statistical uncertainty will be obtained by calculating the appropriate summary statistics from the 50,000 values (5 chains times 10,000 values each) remaining after the burn-in iterations are discarded. The Monte Carlo uncertainty will be determined by calculating the variability between the five point estimates of the model parameters that are obtained from each of the single chains.

Appendix C. Tetrachoric Correlation Coefficients by Ability Group

The following results are for:

High Ability Group Tetrachoric correlations

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10	ITEM11	ITEM12	ITEM13	ITEM14	ITEM15	ITEM16	ITEM17
ITEM1	1.000																
ITEM2	.	1.000															
ITEM3	.	.	1.000														
ITEM4	.	.	.	1.000													
ITEM5	1.000												
ITEM6	1.000											
ITEM7	0.407	0.599	1.000										
ITEM8	-0.788	1.000									
ITEM9	0.509	1.000								
ITEM10	0.770	0.702	0.496	.	0.378	1.000							
ITEM11	0.648	0.407	-0.788	0.509	0.906	1.000						
ITEM12	0.725	0.662	0.428	.	0.319	0.840	0.725	1.000					
ITEM13	0.420	0.383	0.276	.	-0.042	0.773	0.952	0.419	1.000				
ITEM14	0.052	0.292	-0.102	-0.871	-0.148	0.420	0.941	0.171	0.056	1.000			
ITEM15	1.000		
ITEM16	0.959	1.000	
ITEM17	1.000

Number of observations: 261

The following results are for:

Mid High Ability Group Tetrachoric correlations

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10	ITEM11	ITEM12	ITEM13	ITEM14	ITEM15	ITEM16	ITEM17
ITEM1	1.000																
ITEM2	.	1.000															
ITEM3	.	.	1.000														
ITEM4	.	.	.	1.000													
ITEM5	1.000												
ITEM6	0.362	.	.	-0.787	.	1.000											
ITEM7	0.218	.	.	-0.782	.	0.335	1.000										
ITEM8	0.177	.	.	-0.784	.	0.075	0.373	1.000									
ITEM9	0.126	0.230	.	0.270	1.000								
ITEM10	0.181	0.313	0.152	-0.096	-0.150	1.000							
ITEM11	0.409	0.388	.	.	.	1.000						
ITEM12	0.177	.	.	-0.784	.	0.075	.	0.328	0.270	.	.	1.000					
ITEM13	-0.313	.	.	-0.247	.	-0.299	-0.369	-0.374	-0.489	-0.637	-0.376	-0.190	1.000				
ITEM14	-0.270	.	.	-0.213	.	-0.559	0.008	-0.239	-0.584	-0.461	-0.138	-0.329	0.360	1.000			
ITEM15	1.000		
ITEM16	0.377	0.443	.	.	0.486	0.604	.	-0.270	0.035	.	1.000	
ITEM17	1.000

Number of observations: 262

The following results are for:

Low Mid Ability Group Tetrachoric correlations

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10	ITEM11	ITEM12	ITEM13	ITEM14	ITEM15	ITEM16	ITEM17
ITEM1	1.000																
ITEM2	.	1.000															
ITEM3	.	.	1.000														
ITEM4	.	.	.	1.000													
ITEM5	1.000												
ITEM6	0.214	-0.805	.	.	.	1.000											
ITEM7	-0.152	0.145	1.000										
ITEM8	0.008	0.189	0.033	1.000									
ITEM9	-0.018	0.329	.	.	.	0.186	-0.088	-0.170	1.000								
ITEM10	-0.080	0.047	.	.	.	-0.511	-0.061	-0.171	-0.240	1.000							
ITEM11	-0.204	-0.056	-0.427	-0.232	0.101	1.000						
ITEM12	-0.284	0.289	.	.	.	-0.202	-0.162	-0.188	-0.170	-0.204	-0.284	1.000					
ITEM13	-0.275	-0.098	0.004	0.306	-0.144	-0.160	-0.275	-0.017	1.000				
ITEM14	-0.300	-0.325	0.079	-0.142	-0.471	0.077	0.010	0.051	0.155	1.000			
ITEM15	0.493	-0.805	0.303	0.289	.	0.889	.	0.289	.	.	1.000		
ITEM16	0.062	0.662	.	.	.	0.041	-0.871	-0.183	0.064	-0.275	.	-0.183	0.873	0.131	.	1.000	
ITEM17	1.000

Number of observations: 263

The following results are for:

Low Ability Group Tetrachoric correlations

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10	ITEM11	ITEM12	ITEM13	ITEM14	ITEM15	ITEM16	ITEM17
ITEM1	1.000																
ITEM2	0.574	1.000															
ITEM3	0.652	0.923	1.000														
ITEM4	0.701	0.787	0.869	1.000													
ITEM5	0.819	0.925	0.926	.	1.000												
ITEM6	0.527	0.553	0.517	0.639	0.692	1.000											
ITEM7	0.313	0.376	0.536	0.482	0.636	0.550	1.000										
ITEM8	0.432	0.458	0.530	0.555	0.691	0.583	0.382	1.000									
ITEM9	0.395	0.564	0.543	0.632	0.720	0.401	0.257	0.338	1.000								
ITEM10	0.348	0.507	0.521	0.325	0.894	0.252	0.232	0.161	0.249	1.000							
ITEM11	0.446	0.575	0.861	0.481	0.772	0.299	0.388	0.305	0.189	0.334	1.000						
ITEM12	0.422	0.512	0.620	0.684	0.791	0.334	0.201	0.189	0.353	0.473	0.371	1.000					
ITEM13	0.103	.	0.882	0.248	0.185	0.372	0.199	0.196	0.158	0.509	0.035	0.317	1.000				
ITEM14	0.168	0.453	0.887	0.125	0.872	-0.073	0.222	0.055	0.083	0.370	0.339	0.381	0.485	1.000			
ITEM15	0.604	0.879	0.952	0.813	.	0.565	0.548	0.454	0.534	0.591	0.726	0.570	0.388	0.171	1.000		
ITEM16	0.441	0.669	0.722	0.687	0.866	0.462	0.480	0.497	0.381	0.330	0.467	0.465	0.448	0.245	0.674	1.000	
ITEM17	0.741	0.873	0.909	0.944	.	0.684	0.759	0.698	0.769	.	0.689	0.667	0.872	0.276	0.844	0.872	1.000

Number of observations: 261

