Evaluative Report
Department of Financial Institutions Program
Washington State

October 2008

Eric A. Hagedorn, Ph.D.

Hagedorn Evaluation Services
El Paso, TX

MoneySavvy GENERATION
HELPING KIDS GET SMART ABOUT MONEY™

Introduction

The purpose of this study is to evaluate the effectiveness of the Money Savvy U® Intermediate Personal Finance Curriculum on pupils in schools in Washington State that used this curriculum.

To investigate the effectiveness of this program a 10 question multiple choice test, called "The Money Savvy U® Personal Finance Curriculum test" was used. A portion of the questions were drawn from the JumpStart Coalition Personal Finance Literacy test (items 1, and 5 through 10) and a portion were provided by the curriculum developers (items 2 through 4).

Conclusions

These data indicate that the Money Savvy U® Curriculum had a statistically significant impact on the learning of these school children as measured with the test used. The overall effect size for this gain in test score is medium. Changes in the percentages of participants getting individual items correct before and after instruction clearly indicate statistically significant improvement on 9 out of 10 items. Item 7 was the only item where students did not show a significant improvement from pre to post. As the item characteristics for item 7 are not problematic, this suggests a curricular rather than an item wording difficulty. Some possibilities might include: teachers not thoroughly understanding the concepts underlying the item, the concepts not being adequately discussed in the written curriculum, etc.

Despite these positive results, the fact that on average, the participating students get just over half of the questions on the test correct, after instruction, suggest that the test is a bit difficult for these students. The IRT analyses, particularly the test information curves support this assertion. It is noteworthy, that in 2007-2008 the vast majority of the tested students are in the 6th and 7th grades rather than in high school. This could account for some of their difficulty with the test.

Specific item analyses suggest that Item 6 needs to be revisited. The increased evidence of successfully guessing the correct response to this item on the post-test suggests that some of the distracters are too easily eliminated by even less able students.

A general recommendation based on these results would be to revise or replace some of the more difficult items on the Money Savvy U® Curriculum Test. Several easier items would allow more insight into the improvement of students with other than very high ability.

The general positive finding that, on average, students exhibited statistically demonstrable improvement is consistent with findings from other Money Savvy curriculum evaluation (e.g. Money Savvy Kids®).

Methodology

There were 567 completed pre and post-tests in the sample analyzed here. These included the participating students', teachers', and school's names. This allowed for matching individual pre and post-tests. Once matched and recorded, either a paired-samples t-test or the non-parametric

Wilcoxon Signed Ranks test could be performed on the mean raw scores (out of 10) to determine if student responses changed from pre to post in a statistically significant manner.

Any statistically significant change from pre to post will be identified and interpreted. The effect size of any significant change will also be calculated. The effect size is essentially the ratio of the change to the standard deviation of the change score.

In addition to this overall analysis of raw score improvement, the percentages of students choosing the correct responses and incorrect responses to each item are provided. The percentage change in students getting the item correct on the post-test who had gotten it wrong on the pre-test are also provided. Finally, the McNemar test will be used to test the hypothesis that the proportion of students who get an item wrong on the pre-test and then correct on the post-test is the same as the proportion of students who get that item right on the pre-test and then wrong on the post-test. Simply put, if more students get an item correct on the post-test than did on the pre-test or conversely if more students get an item wrong on the post-test than did on the pre-test, this test assesses how likely such a change in proportion is entirely due to random chance. The percentage changes in students correctly answering post-test items who missed that item on the pre-test will determine which situation the McNemar test indicates is statistically significant or not.

While there were additional unmatchable pre-tests and post-tests (unmarked or from differing schools), the sample of matched tests is sufficiently large to use for making inferences.

Results

Grade Demographics

For the matched tests, the participants were from the grades indicated in Table 1. The predominant number of students (97%) are in the 6th and 7th grades.

Table 1

	Frequency	Percent
6	247	43.6
7	303	53.4
8	9	1.6
12	7	1.2
13	1	0.2
Total	567	100.0

Comparing Matched Tests: Total Scores and Item Percentages Correct

Table 2 shows the mean raw score (total number correct) for the entire group of pre-tested students compared to the mean raw scores for the same group of post-tested students.

Table 2. Descriptive Statistics for Raw Scores on Pre and Post-tests

	N	Mean	Std. Deviation
Pre	567	4.12	2.138
Post	567	5.33	2.317

As these data were not normally distributed, a Wilcoxon Signed Ranks test (the non-parametric equivalent of the paired samples t-test) was used to determine whether the mean increase of 1.21 more questions correct was likely to have occurred by chance. The Wilcoxon Z value was -13.087 which indicates that there was less than one chance in 1000 that this increase occurred by random chance ($p = .000$).

The effect size for this improvement is 0.54. This is considered a "medium" effect size.

Table 3 shows the percentage of students getting each item correct or not on the pre and post-tests. It also provides results of the McNemar tests (which uses the chi-square).

Table 3. Percent of Total Wrong "0" and Correct "1" both Pre and Post and Chi Square Results

Item 1		Pretest %	Posttest %	% students improving	Chi square (sig p)
Valid	0	65.3	56.6	8.7	10.15
	1	34.7	43.4		0.001
	Total	100.0	100.0		
Item 2		Pretest %	Posttest %		
Valid	0	55.6	48.3	7.3	7.24
	1	44.4	51.7		0.007
	Total	100.0	100.0		
Item 3		Pretest %	Posttest %		
Valid	0	52.0	42.2	9.8	15.28
	1	48.0	57.8		0.000
	Total	100.0	100.0		
Item 4		Pretest %	Posttest %		
Valid	0	72.5	45.0	27.5	89.65
	1	27.5	55.0		0.000
	Total	100.0	100.0		
Item 5		Pretest %	Posttest %		
Valid	0	63.3	47.1	16.2	41.41
	1	36.7	52.9		0.000
	Total	100.0	100.0		
Item 6		Pretest %	Posttest %		
Valid	0	34.9	21.9	13.0	28.35
	1	65.1	78.1		0.000
	Total	100.0	100.0		
Item 7		Pretest %	Posttest %		
Valid	0	57.0	54.7	2.3	0.71
	1	43.0	45.3		0.400
	Total	100.0	100.0		
Item 8		Pretest %	Posttest %		
Valid	0	53.6	32.5	21.1	17.07
	1	46.4	67.5		0.000
	Total	100.0	100.0		
Item 9		Pretest %	Posttest %		
Valid	0	44.8	32.5	12.3	25.32
	1	55.2	67.5		0.000
	Total	100.0	100.0		
Item 10		Pretest %	Posttest %		
Valid	0	88.7	76.2	12.5	32.45
	1	11.3	23.8		0.000
	Total	100.0	100.0		

Item Analyses

Classical item parameters (see Table 4) can tell you a great deal about how easy or difficult particular items were to the students who answered these items. They can also tell you about the relationship between students' total scores as compared to whether they got a particular item correct.

Facility, the percentage of students who correctly answered an item, is essentially a measure of item "easiness" and "difficulty." Table 5 shows ranked item difficulties for both the pre and post-tests. Thus, the easiest item on "The Money Savvy U® Personal Finance Curriculum test" was item 6: 65% of the students who completed this item got it correct on the pretest and 78% of them got it correct on the post-test. The least easy item was number 10: only 11% of the students chose the correct response to this item on the pre-test and 24% on the post-test.

"P. Bis" refers to the Point Biserial Correlation coefficients. This is a special type of correlation between a dichotomous item score (either right or wrong, indicated as "1" or "0," respectively) and the total score. More or less, a good, difficult item would be gotten right by a student with a high total score (implying a student with higher ability) and gotten wrong by a student with a lower total score (a student with lesser ability), thus each of these students item and total scores would contribute to a higher correlation coefficient (varying as usual from 0 to 1). A problematic, item would be one that if a student got it correct, they were likely to have a low total score, or on the contrary, if a student got this item wrong, they nevertheless had a higher total score. This kind of item would give a lower point biserial correlation. The rule of thumb for point biserial correlation coefficients is that values less than 0.300 are problematic, items between .300 and .400 are acceptable and items better than .400 are good. Table 6 ranks the items from lowest to highest point biserial correlation coefficients.

Table 4. Classical Test Theory Item Parameters

	Pre		Post	
Item	Facility	P.Bis	Facility	P.Bis
1	0.347	0.389	0.434	0.533
2	0.444	0.516	0.517	0.354
3	0.480	0.569	0.578	0.542
4	0.275	0.336	0.550	0.551
5	0.367	0.459	0.529	0.539
6	0.651	0.539	0.781	0.409
7	0.430	0.485	0.453	0.447
8	0.464	0.495	0.573	0.463
9	0.552	0.488	0.675	0.543
10	0.113	0.159	0.238	0.437

Table 5. Items sorted from easiest to hardest.

	Pre		Post	
Item	Facility	Easiest	Item	Facility
6	0.651		6	0.781
9	0.552		9	0.675
3	0.480		3	0.578
8	0.464		8	0.573
2	0.444		4	0.55
7	0.430		5	0.529
5	0.367		2	0.517
1	0.347		7	0.453
4	0.275		1	0.434
10	0.113	Hardest	10	0.238

Table 6. Items sorted from lowest to highest point biserial correlation.

Pre			Post		
Item	Facility	P.Bis	Item	Facility	P.Bis
10	0.113	0.159	2	0.517	0.354
4	0.275	0.336	6	0.781	0.409
1	0.347	0.389	10	0.238	0.437
5	0.367	0.459	7	0.453	0.447
7	0.43	0.485	8	0.573	0.463
9	0.552	0.488	1	0.434	0.533
8	0.464	0.495	5	0.529	0.539
2	0.444	0.516	3	0.578	0.542
6	0.651	0.539	9	0.675	0.543
3	0.48	0.569	4	0.550	0.551

Item Response Theory Analysis

Item response theory is an improvement on classical test theory. While it has more reasonable underlying assumptions, more useful applications, and a graphically interpretable format, it is far more complicated from a computational point of view.

From data consisting of 0's and 1's for every incorrect and correct item response to the 10 items on the MSU test for 567 students, three item parameters are estimated for each item: 1) an item difficulty parameter "b", 2) a discrimination index "a", and 3) a guessing parameter. These parameters describe an item response curve (see three in Figure 1). Student ability, or theta, is graphed on the horizontal axis, typically from -3 to 3. The vertical axis represents the probability

of getting an item correct. As you can see, the lower you go on theta (student ability), the less probability there is of the student getting the item correct. At high theta, there is a 100% probability (1.0) of getting the item correct. Three curves for items of 3 varying difficulties are shown in Figure 1. The most difficult item has a b of 1; the least difficult a b of -1. An item of moderate difficulty has a b of 0.

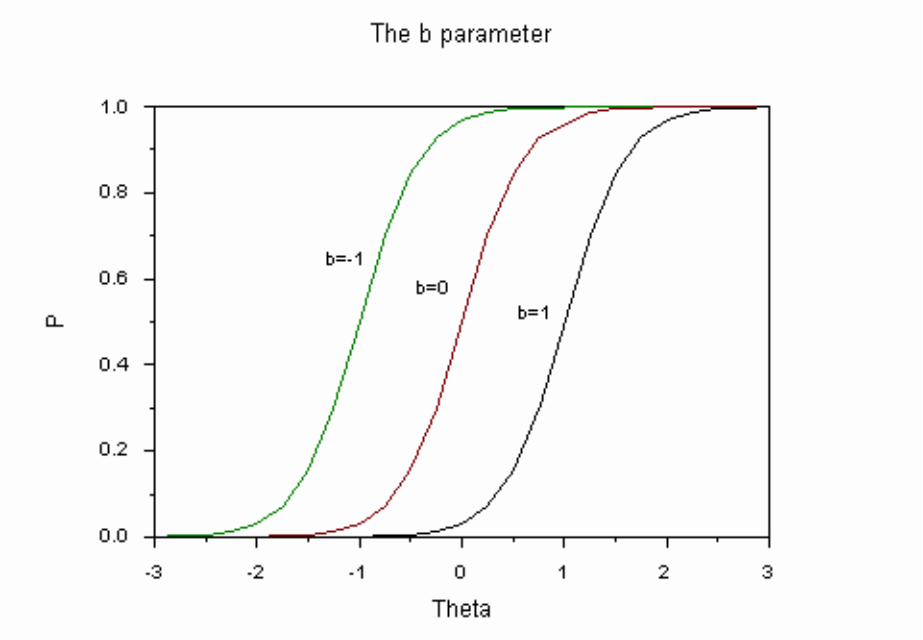

Figure 1. Item difficulty

The "a" parameter determines the steepness of the slope of the item curve. The steeper the slope, the more the item discriminates between students of more and more similar ability. In Figure 2, three equally difficult item curves are shown, but with varying discriminations. The curve with an a=2 clearly differentiates between students with abilities just above 0 on the theta scale and just below. The curve with a=.5 does not differentiate between students of similar abilities.

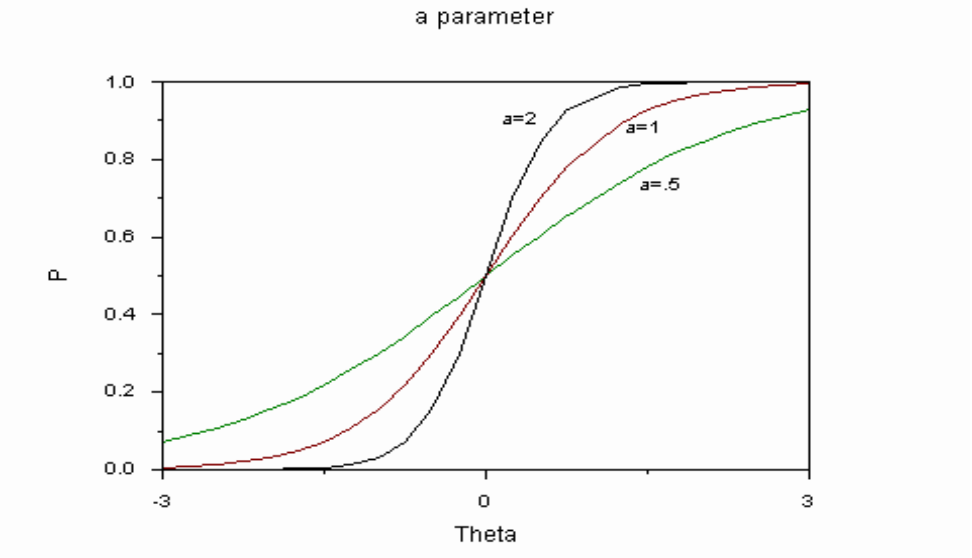

Figure 2. Item discrimination

Finally, the "c" parameter estimates the probability of lower ability students correctly guessing. In a 4 choice, multiple choice test, students simply choosing at random would have a .25 probability of getting the item correct. Three item curves with varying c's are shown in Figure 3.

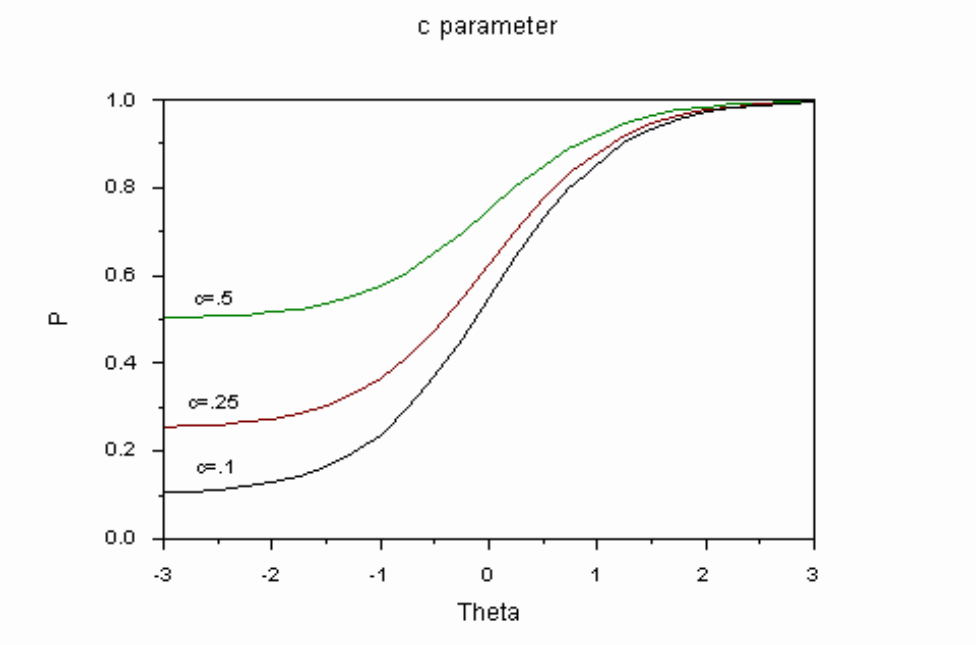

Figure 3. Item guessing

Table 7 shows the estimated parameters for the 10 MSU items as based upon the 567 completed and item scored pretests. Figure 4 shows the curves for each item. Table 7 and Figure 5 present these same results for the post-test. Table 9 compares IRT difficulty parameters for pre and post.

Table 7. Item parameters from IRT analysis of Pretest

Item	a steepness of slope: discrimination	b item difficulty	c guessing parameter
1	0.282	1.383	0.000
2	0.692	0.247	0.000
3	0.735	0.090	0.000
4	4.454	1.704	0.241
5	0.473	0.900	0.031
6	0.950	-0.550	0.000
7	0.669	0.743	0.000
8	0.485	0.205	0.000
9	0.614	-0.241	0.000
10	1.987	2.744	0.160

Table 8. Item parameters from IRT analysis of Posttest

Item	a steepness of slope: discrimination	b item difficulty	c guessing parameter
1	0.738	0.550	0.101
2	0.215	-0.191	0.000
3	3.208	0.440	0.363
4	1.523	0.265	0.234
5	1.681	0.498	0.292
6	1.686	0.235	0.623
7	0.379	0.318	0.000
8	0.813	0.490	0.314
9	0.985	-0.595	0.041
10	3.345	1.298	0.146

Table 9. Item Difficulties from IRT analysis of Pre and Post-test compared

Pre			Post	
Item	b item difficulty		Item	b item difficulty
6	-0.550	Easiest	9	-0.595
9	-0.241		2	-0.191
3	0.090		6	0.235
8	0.205		4	0.265
2	0.247		7	0.318
7	0.743		3	0.440
5	0.900		8	0.490
1	1.383		5	0.498
4	1.704		1	0.550
10	2.744	Hardest	10	1.298

The ranking of difficulty for the pre-test items matches the classical ranking. Bear in mind – the rank order is the same, the magnitude of the values is different. The modeling also shows little evidence of guessing – with the exception on Item 4. The ranking of difficulty for the post-test items shows some differences in ranking, particularly with mid-difficulty items. There are, however, on the post-test more estimates of guessing.

It is important to keep in mind that the classical parameters are directly derived from the particular data set that comes from a particular group of students. A weaker group of students could make test items appear more difficult than a stronger group of students. The IRT parameters while estimated from data from a particular group of students, are based on a model that proposes to give the same results regardless of the ability level of the group of students who provide the data. In other words, the IRT model for this test is based upon a distinct set of data, but it generalizes to all students on the ability scale and includes the impact of guessing.

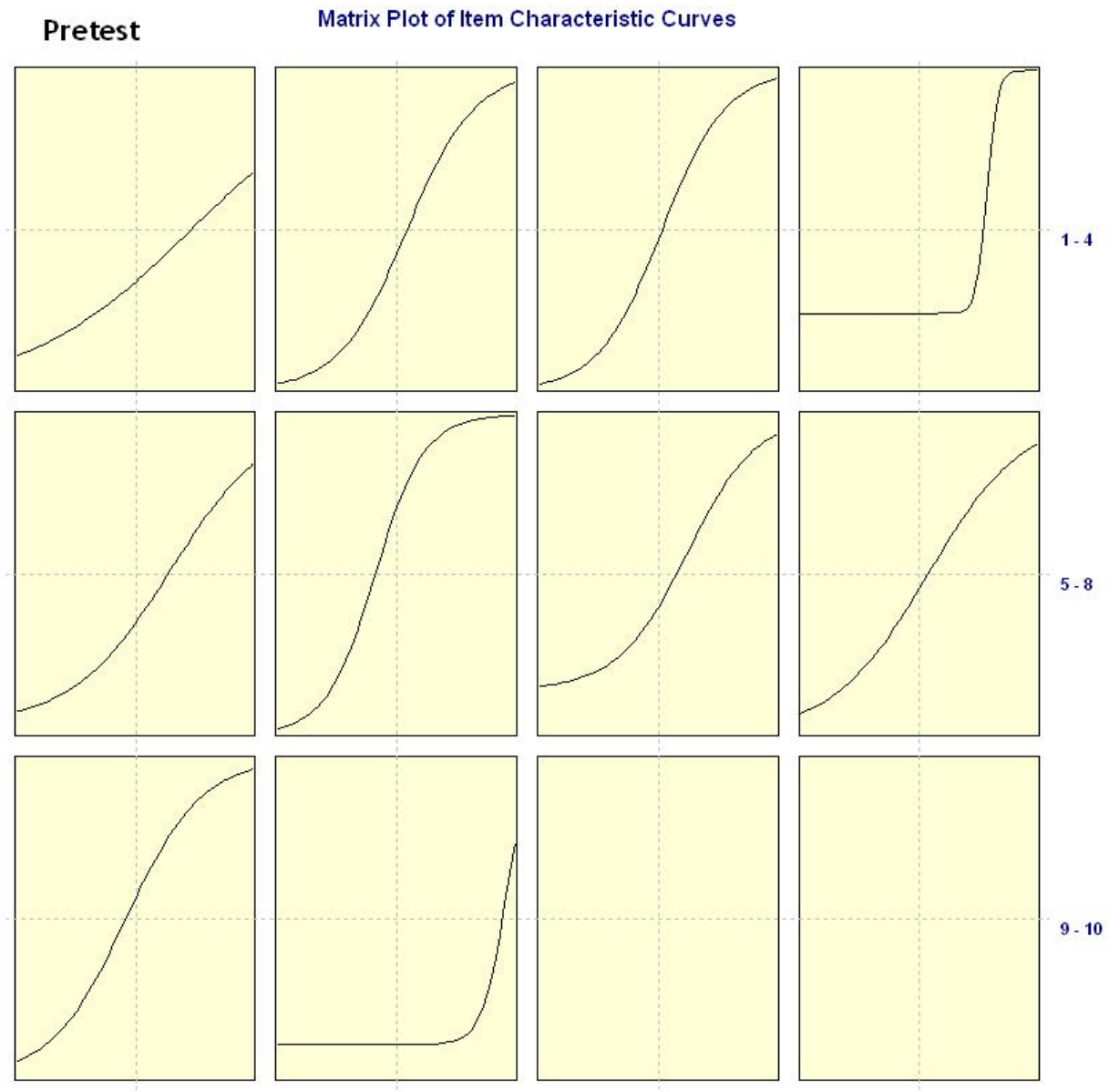

Figure 4. Item Response Curves for all 10 items on pretest.

A visual inspection of Figure 4, suggests that item 1 does not discriminate well. Item 4 is fairly difficult but can be guessed at with roughly a 24% chance of guessing correctly. Item 10 is not only fairly hard, but sharply discriminates between students of slightly varying higher abilities. It is worth noting that this item had a problematic classical measure of discrimination: a point biserial coefficient of 0.159 (Table 7).

A visual inspection of the post-test curves in Figure 5 indicates problems with items 2 and 6. Item 2 does not discriminate between students of varying abilities very well. This corresponds to it having had the lowest point biserial correlation coefficient – although, classically acceptable. Item 6 is unusual in that while fairly easy (classically the easiest and 3rd easiest in the IRT model), the IRT model suggests that the correct answer could be found by guessing roughly 60% of the time.

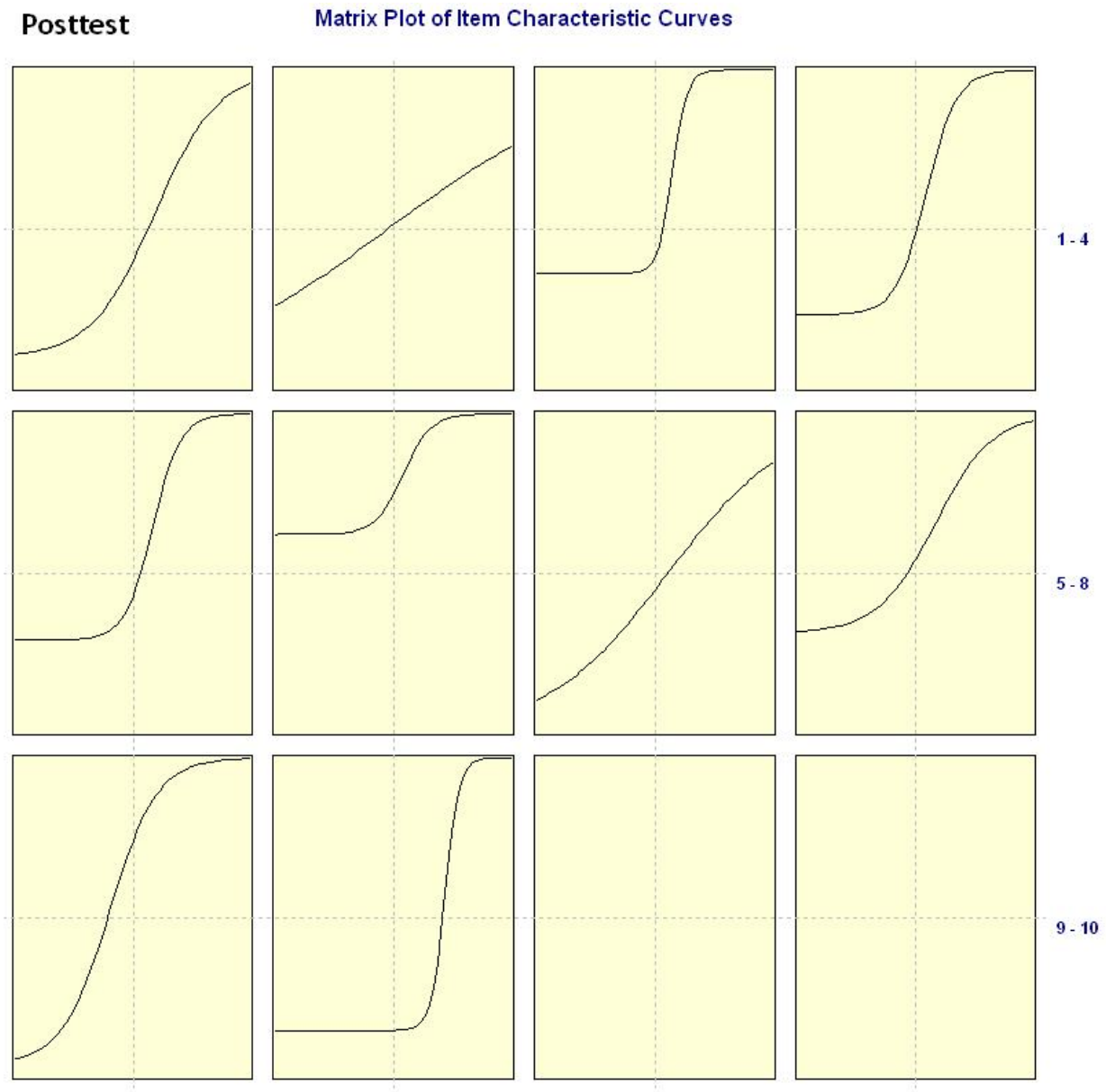

Figure 5. Item Response Curves for all 10 items on pretest.

Finally, based on all the item parameters, IRT models can allow one to describe how well the entire test measures across the ability spectrum (although not in terms of theta, but in terms of a scale score related to theta). This description can be graphically represented in what is referred to as a test information curve. The curves shown in figures 6 and 7. The solid blue line indicates "how much" information we can determine about a student's ability based on their scale score. The dotted line indicates standard error. The more standard error, the less information. (Please note the different Information scales – this was unintentional and cannot be manually set.)

In Figure 6, the large peak to the right of the graph indicates that this test can tell us a great deal about rather high ability/high scoring students, but over a rather narrow range. The fact that the information curve is low and flat for lower scale scores indicates that this test does not tell us very much about lower ability students. Simply put, this graph suggests that this is a rather difficult test.

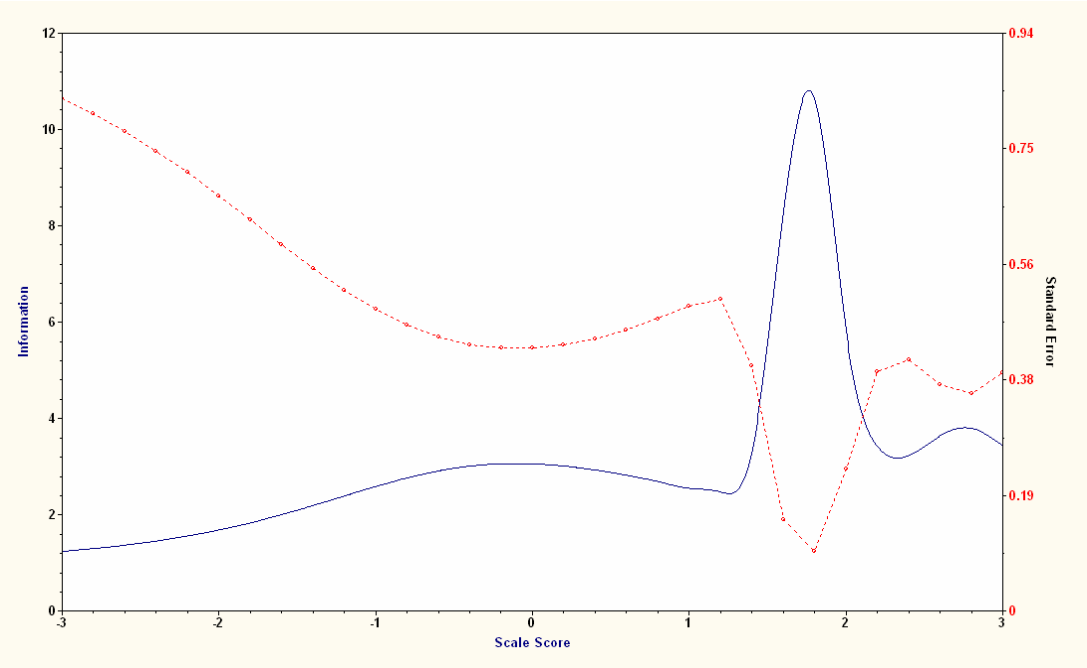

Figure 6. MSU pre-test information curve

The post-test curve in Figure 7 indicates increased information across the range of higher abilities, but still a dearth at lower ability levels.

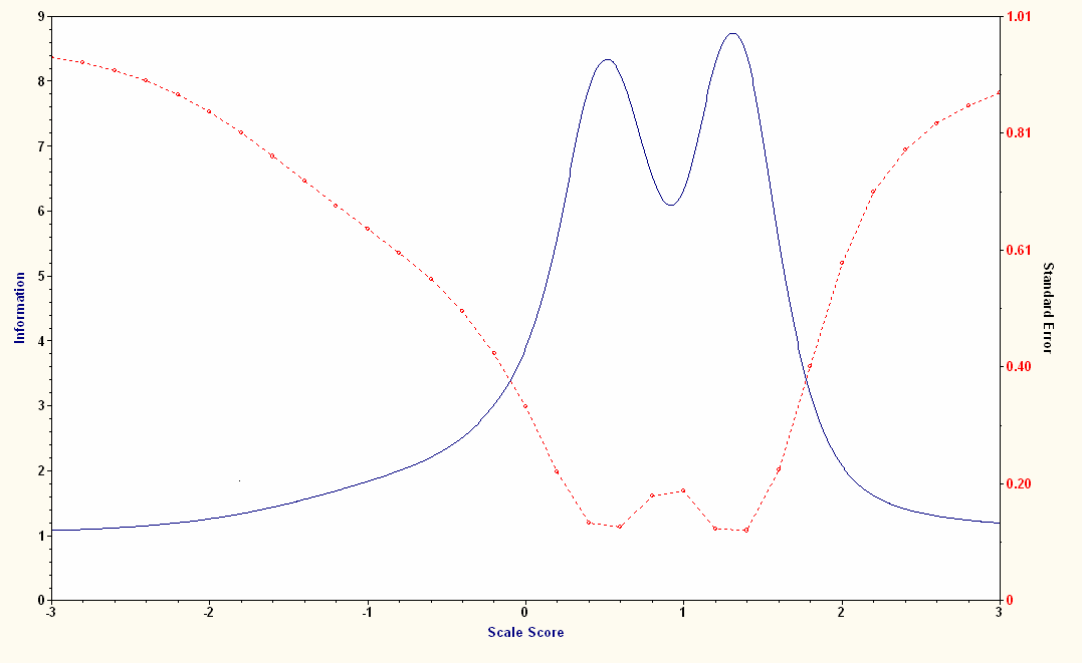

Figure 7. MSU post-test information curve