Evaluative Report
Department of Financial Institutions Program
Washington State

September 2009

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 5, 9, and 10) while the remaining 7 items were provided by the curriculum developers. This study also evaluates the effectiveness of this test in terms of classical and IRT item analysis.

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 large, close to one full standard deviation of improvement. Changes in the percentages of participants getting individual items correct before and after instruction clearly indicate statistically significant improvement on all 10 items.

While there is clearly room for student improvement (as indicated by getting roughly only 6 out of 10 answers correct, on average), the individual IRT analyses and the test information curve clearly show that there are several tough items on this test. This is not a bad thing: these items allow this test to assess the learning of stronger students (in addition to average students). The changes to items from the previous academic year to this one clearly yielded better functioning items.

These data from 2008-2009 not only indicate a program that successfully impacts learning but an improved measure with which to evaluate such impact.

Methodology

There were 729 completed pre-tests and 722 completed post-tests in the sample analyzed here. 514 could be matched. These included the participating students', teachers' (27 teachers), and school's names (9). 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.

There were also additional pre-tests and post-tests where students either missed one test or a particular class missed a test. These when combined with the matched tests (which we unmatched) gave us a sample of 729 pre-tests and 722 post-tests. For thoroughness these were

analyzed using independent samples tests, either the independent samples t-test or the non-parametric Mann-Whitney U test. Again any statistically significant change from pre to post will be identified and interpreted (with effect size).

In addition to these two overall analyses of raw score improvement, the percentages of students choosing the correct responses and incorrect responses to each item are provided for the 729 pre-tests and the 722 post-tests. 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 Fisher Exact test will be used to test the hypothesis that the changes in numbers of students getting an item right or wrong on the post-test, compared to the number right or wrong on the pre-test is statistically significant. In other words, it it seems like many more students got this item correct on the post-test, how likely is it that this could have occurred by chance?

Results

Grade Demographics

7 schools and 27 teachers. For the matched tests, the participants were from the grades indicated in Table 1. The predominant number of students (97.7%) are in the 6^{th} and 7^{th} grades.

Table 1

	Frequency	Percent
6	300	58.4
7	202	39.3
8	12	2.3
Total	514	100

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	514	3.50	1.820
Post	514	5.52	2.365

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 2.05 more questions correct was likely to have occurred by chance. The Wilcoxon Z value was -14.8881 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 0.92. This is considered a "large" effect size.

Comparing Independently Sampled Tests: Total Scores

Grade Demographics

These 729 pretests and 722 postests were drawn from 940 students at 9 schools with 33 teachers. For the independent tests, the participants were from the grades indicated in Table 3. The predominant number of students (97%) are in the 6th and 7th grades.

Table 3

	Frequency	Percent
6	418	44.5
7	450	47.9
8	72	7.7
Total	940	100.0

Table 4 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 4. Descriptive Statistics for Raw Scores on Pre and Post-tests

	N	Mean	Std. Deviation
Pre	729	3.36	1.807
Post	722	6.24	2.751

As these data were not normally distributed, a Mann-Whitney U test (the non-parametric equivalent of the independent samples t-test) was used to determine whether the mean increase of 2.05 more questions correct was likely to have occurred by chance. The Mann-Whitney U value was 100583.000which 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 1.15. This is considered a "large" effect size.

Table 5 shows the percentage of students getting each item correct or not on the pre and post-tests. It also provides results of the Fischer Exact test which evaluates whether being in the pre or post category is associated with whether your outcome was correct or not correct. Essentially this test tells us whether the numbers of people getting answers right and wrong before and after instruction could have occurred entirely by chance. If these numbers did not occur by chance, we presume they changed because of student learning.

Table 5. Percent of Total Wrong "0" and Correct "1" both Pre and Post and Fischer Exact Test

Item 1		Pretest %	Posttest %	% students improving	Fischer (sig p)
Valid	0	60.7	21.4	39.3	0.0001
	1	39.3	78.6		
	Total	100.0	100.0		
Item 2		Pretest %	Posttest %		
Valid	0	46.5	26.2	20.3	0.0001
	1	53.5	73.8		
	Total	100.0	100.0		
Item 3		Pretest %	Posttest %		
Valid	0	53.3	30.4	22.9	0.0001
	1	46.7	69.6		
	Total	100.0	100.0		
Item 4		Pretest %	Posttest %		
Valid	0	79.2	37.9	41.3	0.0001
	1	20.8	62.1		
	Total	100.0	100.0		
Item 5		Pretest %	Posttest %		
Valid	0	70.8	40.5	30.3	0.0001
	1	29.2	59.5		
	Total	100.0	100.0		
Item 6		Pretest %	Posttest %		
Valid	0	81.1	39.6	41.5	0.0001
	1	18.9	60.4		
	Total	100.0	100.0		
Item 7		Pretest %	Posttest %		
Valid	0	70.4	39.6	30.8	0.0001
	1	29.6	60.4		
	Total	100.0	100.0		
Item 8		Pretest %	Posttest %		
Valid	0	71.8	59.4	12.4	0.0001
	1	28.2	40.6		
	Total	100.0	100.0		
Item 9		Pretest %	Posttest %		
Valid	0	35.9	19.0	16.9	0.0001
	1	64.1	81.0		
	Total	100.0	100.0		
Item 10		Pretest %	Posttest %		
Valid	0	93.3	74.2	19.1	0.0001
	1	6.7	25.8	39.3	
	Total	100.0	100.0		

Item Analyses

Classical item parameters (see Table 6) 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 7 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 9: 64% of the students who completed this item got it correct on the pretest and 81% of them got it correct on the post-test. The least easy item was number 10: only 6.7% of the students chose the correct response to this item on the pre-test and 25.8% 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 0.200 are problematic, items between .200 and .400 are acceptable and items better than .400 are good. Note that the values of the point biserial on the post-test are probably more meaningful than those on the pre-test: students may have had no idea how to answer these questions before instruction.

Table 6. Classical Test Theory Item Parameters

	Pre		Post	
Item	Facility	P.Bis	Facility	P.Bis
1	0.393	0.477	0.786	0.453
2	0.535	0.514	0.738	0.547
3	0.467	0.524	0.696	0.571
4	0.208	0.326	0.621	0.520
5	0.292	0.475	0.595	0.521
6	0.189	0.226	0.604	0.601
7	0.296	0.434	0.604	0.592
8	0.282	0.346	0.406	0.492
9	0.641	0.471	0.810	0.517
10	0.067	0.149	0.258	0.463

Table 7. Items sorted from easiest to hardest.

	Pre		Post	
Item	Facility	Easiest	Item	Facility
9	0.641		9	0.81
2	0.535		1	0.786
3	0.467		2	0.738
1	0.393		3	0.696
7	0.296		4	0.621
5	0.292		6	0.604
8	0.282		7	0.604
4	0.208		5	0.595
6	0.189		8	0.406
10	0.067	Hardest	10	0.258

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 post-test for 722 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.

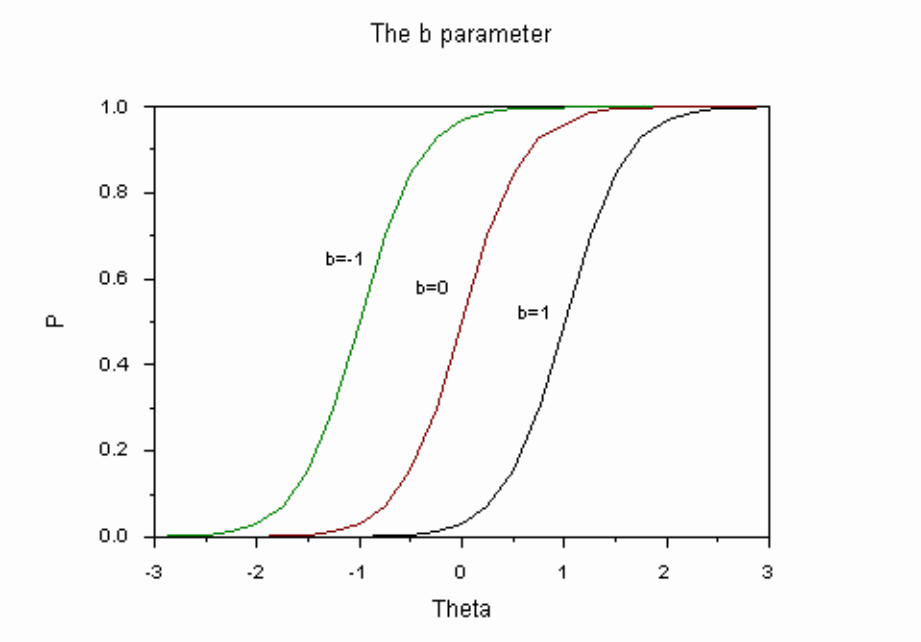

Figure 1. Item difficulty

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.

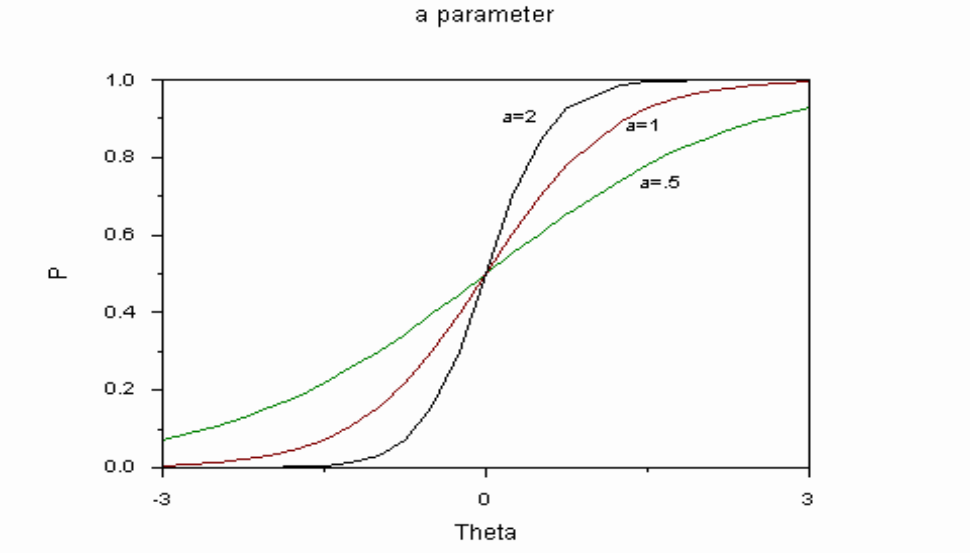

Figure 2. Item discrimination

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.

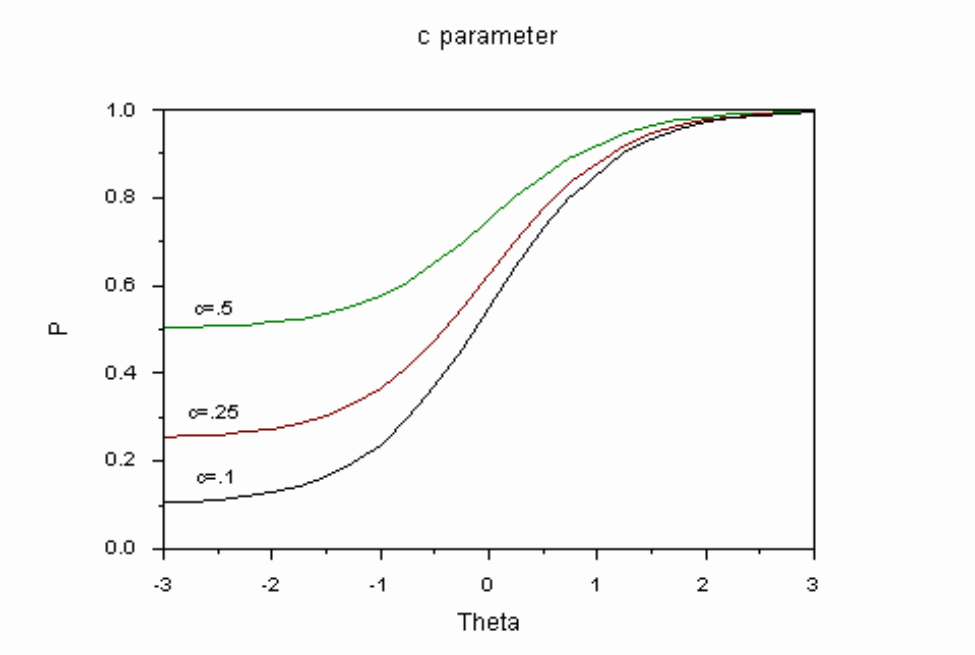

Figure 3. Item guessing

Table 8 shows the estimated parameters for the 10 MSU items as based upon the 722 completed and item scored post-tests. Figure 4 shows the curves for each item. Table 9 ranks the items from easiest to most difficult on the basis of the "b" parameter.

Table 8. Item parameters from IRT analysis of Post-test

Item	a steepness of slope: discrimination	b item difficulty	c guessing parameter
1	1.21	-1.07	0.20
2	1.66	-0.70	0.17
3	1.63	-0.55	0.14
4	1.92	0.16	0.30
5	1.20	-0.07	0.16
6	1.91	-0.07	0.17
7	1.85	-0.09	0.16
8	1.83	0.92	0.20
9	1.74	-1.02	0.20
10	2.88	1.23	0.13

Table 9. Ranked Item Difficulties from IRT analysis of the Post-test

Item	b item difficulty	
1	-1.07	Easiest
9	-1.02	
2	-0.70	
3	-0.55	
7	-0.09	
6	-0.07	
5	-0.07	
4	0.16	
8	0.92	
10	1.23	Hardest

The ranking of difficulty for the post-test items is similar to the classical ranking but not identical. The modeling also shows little evidence of guessing – with the exception on Item 4.

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 include the impact of guessing.

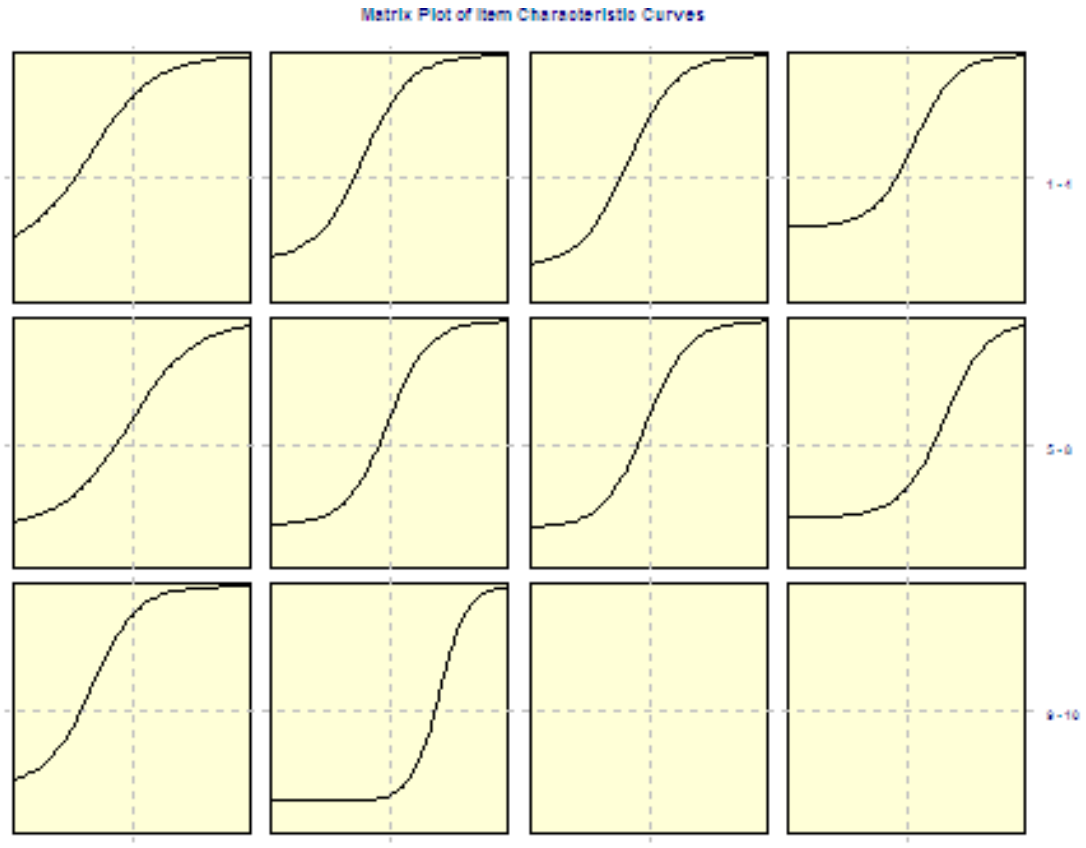

Figure 4. Item Response Curves for all 10 items on posttest.

A visual inspection of Figure 4 suggests that all the items discriminate well. Seven of the items cross the ability (horizontal) axis to the left of the vertical axis (indicative of the 7 negative b values). This indicates that 7 of the 10 items have a 50% chance of being gotten right by students of slightly below average ability. In other words, 7 of these items are "easy." All items but item 4 terminate well below the horizontal axis as you go to the extreme left. Where the black curves terminate on the left is indicative of the probability of low ability students guessing correctly (measured by the "c" parameter). Only Item 4 has a "c" value greater than 0.25 (0.30) indicating that lower ability students have a slightly better chance of guessing the correct answer to this item than purely random guessing (which should be 0.25 for 4 multiple choices).

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. This curve is shown in figure 5. The solid blue line indicates "how much" information we can determine about a student's ability based on their scale score (which could be calculated for each student). The dotted line indicates standard error. The more standard error, the less information.

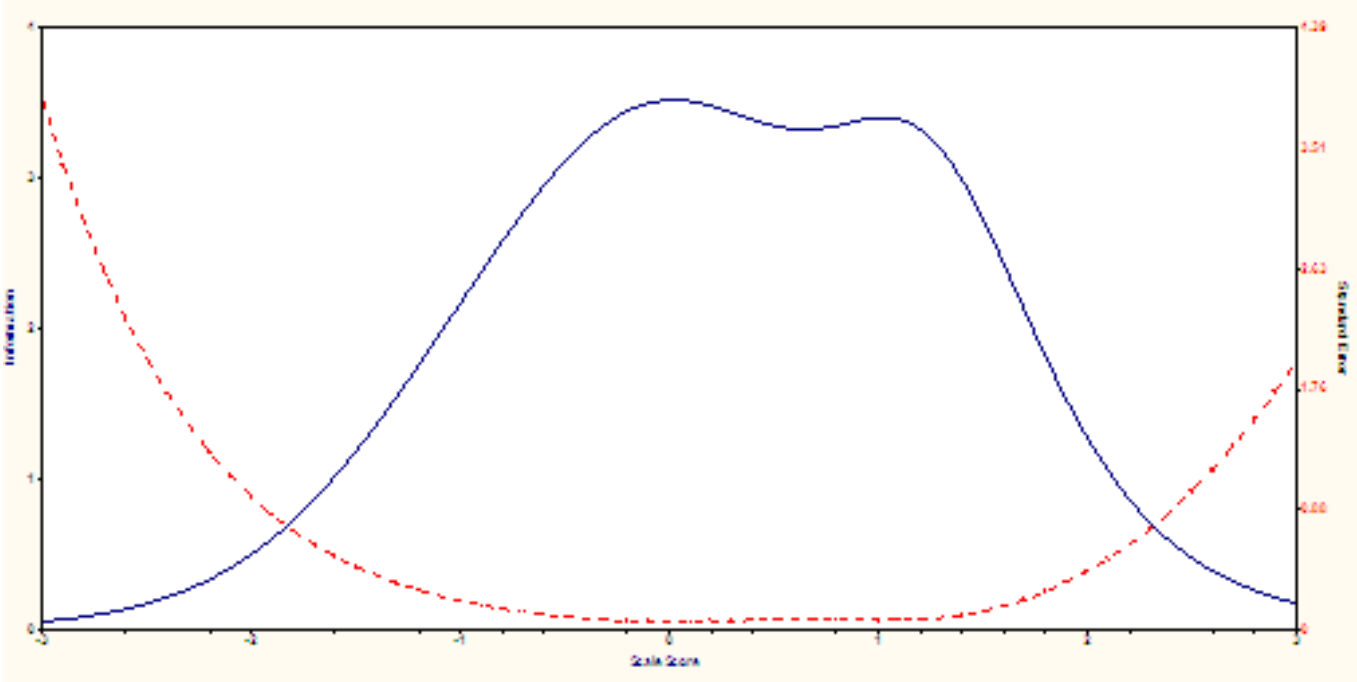

Figure 5. MSU post-test information curve

In Figure 5, the blue item information curve is highest just to the left of zero on the ability scale. This indicates that the test tells us the most about students whose ability is just below average ability. This is related to the fact that 7 of the 10 items are just slightly "easier" than average. Nevertheless, a good deal is known about above average students because while there are only several "harder" items, they discriminate well. The fact that the "mountain" covers more area to the right of zero indicates that the test can tell us about the ability of above average ability students. In other words, this test includes some good, harder items.