Money Savvy U™
Intermediate Personal Finance Curriculum

Washington State
Evaluative Report

November 2007

Eric A. Hagedorn, Ph.D.

Hagedorn Evaluation Services
El Paso, TX

Money Savvy 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).

Methodology

There were 747 completed pre and post-tests in the sample analyzed here. Most included the participating students', teachers', and school's names. This allows 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 would be performed on the data to determine if student responses changed from pre to post in a statistically significant manner.

Because a number of the collected pre and post-tests include ones that cannot be matched (due to students not entering necessary demographic information or having missed class on a test day), an independent samples analysis of all the pre-tests compared to all the post-tests, is presented here, as well. If the data are normal, an independent samples t-test will be used, if not, a Mann-Whitney U test will be used.

Any statistically significant change from pre to post, using whatever method, 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.

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 improvement on 8 out of 10 items. (Items 2 and 7 need to be revisited by the curriculum developers, as these were the two items where the percentage improvement was not significant). The general finding that, on average, students exhibited statistically demonstrable improvement is consistent with findings from other Money Savvy Generation curriculum evaluation (e.g. Money Savvy Kids™).

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 taken as a whole has some deficiencies. Prior to running the IRT analyses, this evaluator suggested that the test used might be too difficult. Based on the test information curve in Figure 5, this evaluator is now convinced that this was indeed the case.

Results

Grade Demographics

For the matched tests, the participants were from the grades indicated in Table 1. As the curriculum was designed for 6th, 7th, and 8th grade students, results will be double checked with the 15 students from other grades removed.

Table 1

		Frequency	Percent
Valid	5	2	.4
	6	58	12.0
	7	332	68.5
	8	80	16.5
	9	6	1.2
	10	5	1.0
	11	1	.2
Missing	System	1	.2
Total		485	100.0

Comparing Matched Tests: Total Scores and Item Percentages Correct

Table 2 shows the mean item scores for the entire group of pre-tested students (labeled with a "1" on the variable "prepost") compared to the mean item scores for the entire group of post-tested students.

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

	N	Mean	Std. Deviation
Pre	262	4.39	1.532
Post	262	5.49	1.943

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 1.10 more questions correct was likely to have occurred by chance. The Wilcoxon Z value was -8.304 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 0.63. 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 gives a chi square analysis of whether the percentages obtained on the post-test are significantly different than those obtained on a pre-test.

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

Item 1		Pretest %	Posttest %	Chi sq (sig p)	Effect size
Valid	0	72.5	57.3	30.64	0.34
	1	27.5	42.7	.000	small
	Total	100.0	100.0		
Item 2		Pretest %	Posttest %		
Valid	0	46.9	51.9	2.59	
	1	53.1	48.1		
	Total	100.0	100.0		
Item 3		Pretest %	Posttest %		
Valid	0	47.3	31.3	27.008	0.32
	1	52.7	68.7	.000	small
	Total	100.0	100.0		
Item 4		Pretest %	Posttest %		
Valid	0	71.4	44.7	91.537	0.59
	1	28.6	55.3	.000	medium
	Total	100.0	100.0		
Item 5		Pretest %	Posttest %		
Valid	0	64.5	51.9	18.153	0.26
	1	35.5	48.1	.000	small
	Total	100.0	100.0		
Item 6		Pretest %	Posttest %		
Valid	0	30.2	21.8	8.771	0.18
	1	69.8	78.2	.003	small
	Total	100.0	100.0		
Item 7		Pretest %	Posttest %		
Valid	0	57.6	55.0	.776	
	1	42.4	45.0		
	Total	100.0	100.0		
Item 8		Pretest %	Posttest %		
Valid	0	50.0	41.6	7.389	0.17
	1	50.0	58.4	.007	small
	Total	100.0	100.0		
Item 9		Pretest %	Posttest %		
Valid	0	40.5	22.1	36.505	0.37
	1	59.5	77.9	.000	small
	Total	100.0	100.0		
Item 10		Pretest %	Posttest %		
Valid	0	80.2	73.7	6.934	0.16
	1	19.8	26.3	.008	small
	Total	100.0	100.0		

Comparing Unmatched Tests: Total Scores

Table 4 shows the mean item scores for the unmatched pretests (labeled with a "1" on the variable "prepost") compared to the mean item scores for the post-tests (labeled with a "2" on the variable "prepost").

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

prepost	N	Mean	Std. Deviation
1	443	4.28	1.68
2	486	5.45	2.17

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 1.18 more questions correct was likely to have occurred by chance. The Mann-Whitney U value was 72930.0 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 0.60. This is considered a "medium" effect size. These independent samples tests allow us to use all the available data with results very similar to the smaller sample of matched data.

Item Analyses

Classical item parameters 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 and "Diff" are essentially measures of item "easiness" and "difficulty." Facility is simply the proportion (expressed as a decimal) of total students who made a correct response to a particular item. Thus, the easiest item on "The Money Savvy U™ Personal Finance Curriculum test" was item 6: 77.1% of the students who completed this item got it correct. The least easy item was number 10: only 19.6% of the students who completed this item got it correct. "Diff" simply takes the facility and scales it so that the range is effectively 1 to 25, with a mean of 13 and a standard deviation of 4. By inspection, one can see that item 6 has the lowest "Diff" of 10.03 and item 10 has the highest "Diff": 16.42. Table 6 sorts the items from easiest to hardest.

"Bis" and "P. Bis" refer to the Biserial Correlation and Point Biserial Correlation coefficients. These are special types of correlations between dichotomous item scores (either right or wrong, indicated as "1" or "0," respectively) and total scores. 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 biserial or 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 7 ranks the items from lowest to highest point biserial correlation coefficients.

Table 5. Classical Test Theory Item Parameters

Item	N	RMean	Facility	Diff	Bis	P.Bis
1	927	5.920	.309	15.00	0.460	0.351
2	927	5.540	.492	13.08	0.412	0.329
3	927	5.610	.601	11.98	0.572	0.451
4	927	6.030	.408	13.93	0.606	0.479
5	927	5.930	.435	13.66	0.582	0.462
6	927	5.260	.771	10.03	0.483	0.348
7	927	5.800	.456	13.44	0.534	0.425
8	927	5.750	.529	12.71	0.581	0.463
9	927	5.630	.685	11.07	0.726	0.555
10	927	6.120	.196	16.42	0.443	0.309

Table 6. Items sorted from easiest to hardest.

Item	Facility	Diff
6	0.771	10.0
9	0.685	11.1
3	0.601	12.0
8	0.529	12.7
2	0.492	13.1
7	0.456	13.4
5	0.435	13.7
4	0.408	13.9
1	0.309	15.0
10	0.196	16.4

Table 7. Items sorted from lowest to highest point biserial correlation.

Item	Diff	Bis	P.Bis
10	16.4	0.443	0.309
2	13.1	0.412	0.329
6	10.0	0.483	0.348
1	15.0	0.46	0.351
7	13.4	0.534	0.425
3	12.0	0.572	0.451
5	13.7	0.582	0.462
8	12.7	0.581	0.463
4	13.9	0.606	0.479
9	11.1	0.726	0.555

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 927 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.

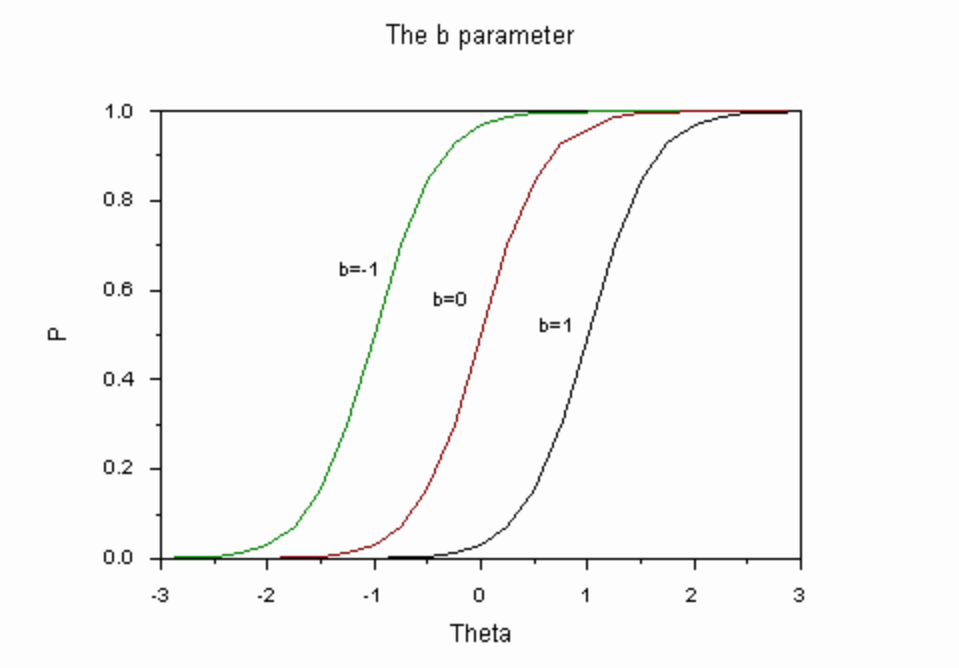

Figure 1. Item difficulty

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.

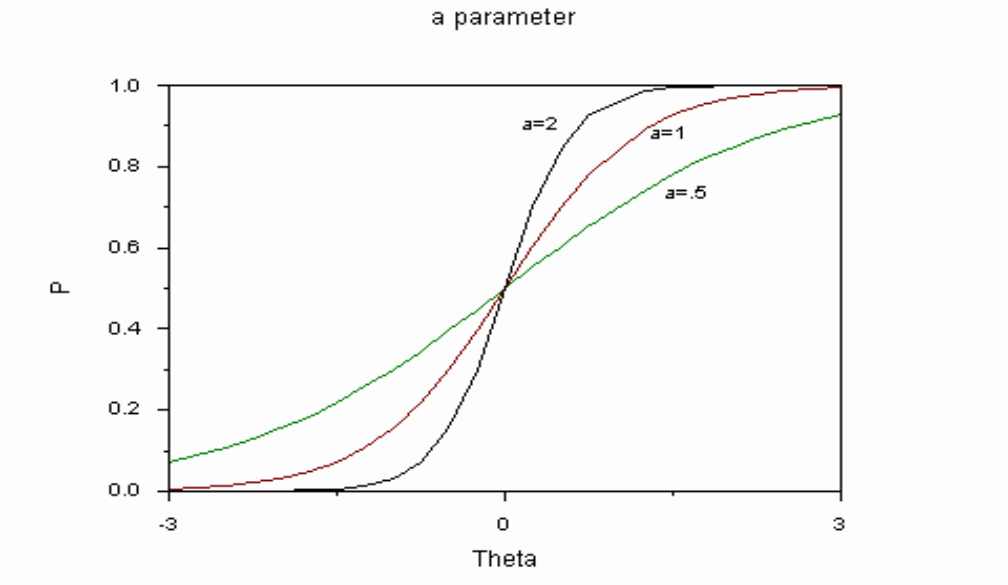

Figure 2. Item discrimination

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.

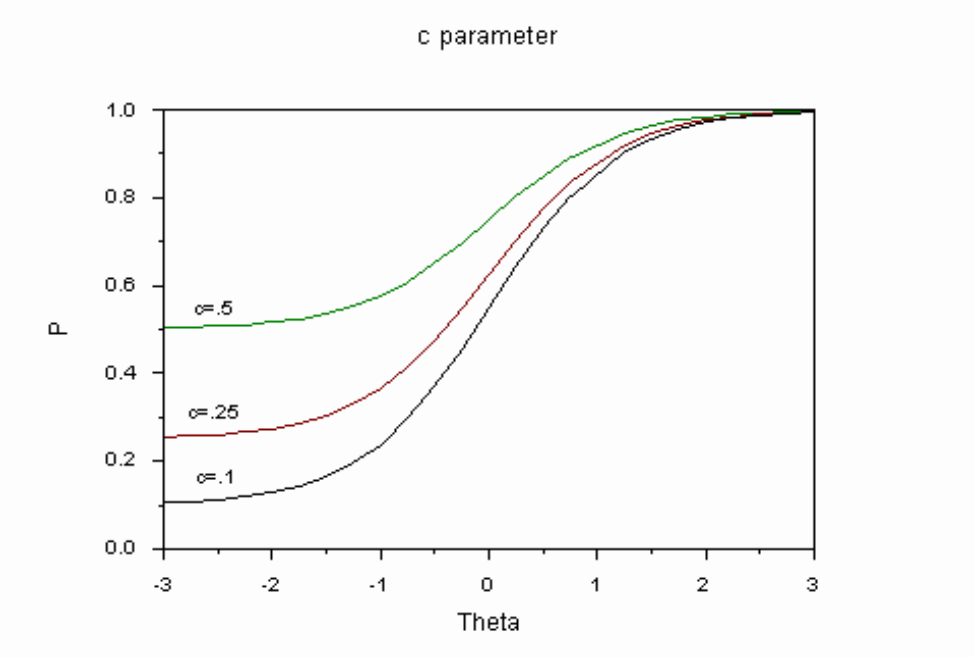

Figure 3. Item guessing

Table 8 shows the estimated parameters for the 10 MSU items as based upon the 927 completed and item scored tests. Figure 4 shows the curves for each item.

Table 8. Item parameters from IRT analysis

Item	a steepness of slope: discrimination	b item difficulty	c guessing parameter
1	1.00	1.85	0.23
2	0.16	0.12	0.00
3	0.59	-0.12	0.16
4	2.45	1.14	0.31
5	0.97	1.04	0.26
6	0.40	-1.96	0.00
7	0.48	0.68	0.12
8	0.37	-0.20	0.00
9	1.86	-0.31	0.20
10	4.07	1.71	0.16

Again, the hardest two items are items 1 and 10, but in reverse order to the classically determined parameters. The classical parameters are directly derived from that data and apply to only those students. The IRT parameters are based on the data, but are generalizing to all students on the ability scale and include the impact of guessing.

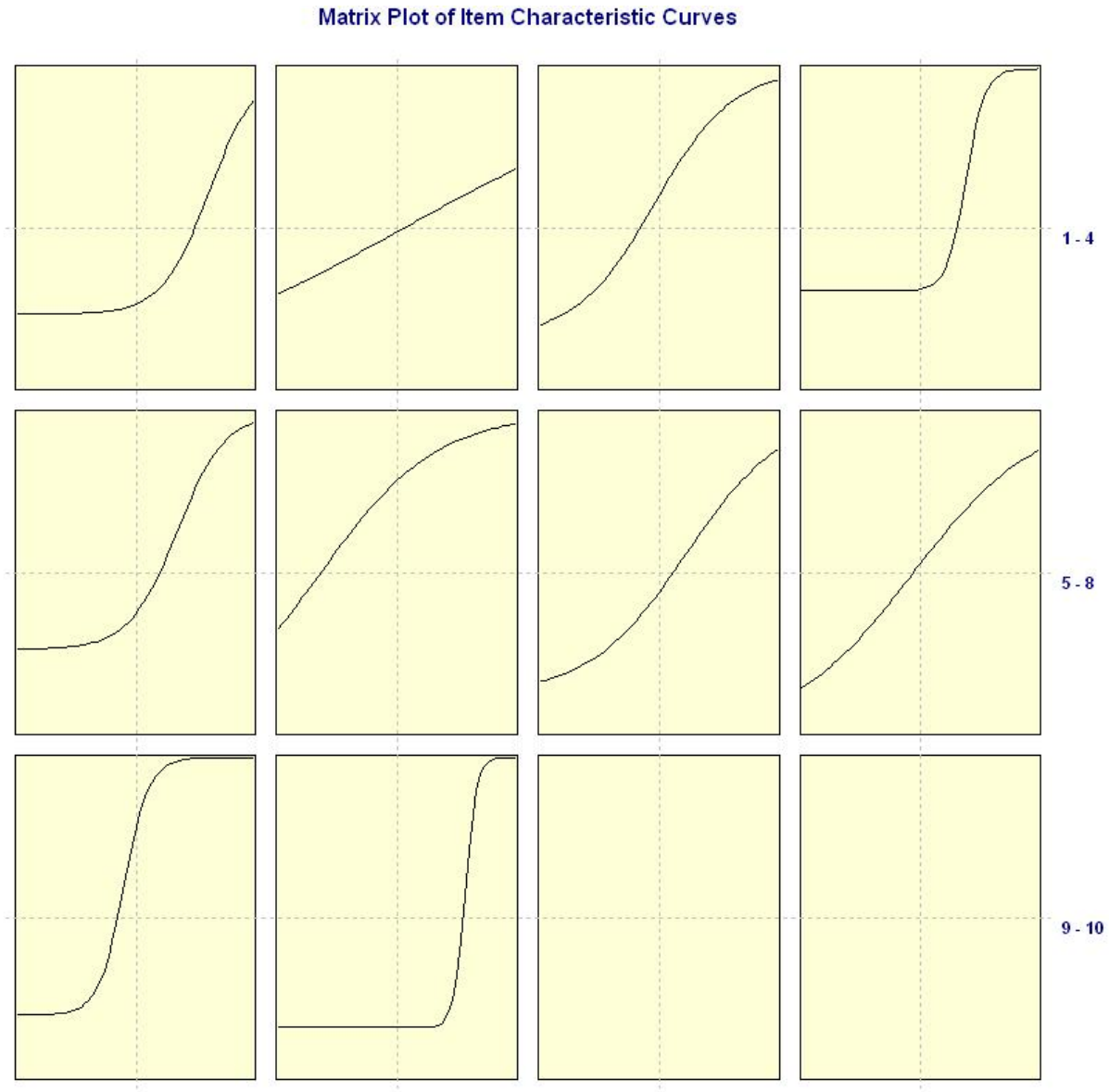

Figure 4. Item Response Curves for all 10 items.

A summary visual inspection suggests that item 2 does not discriminate well. Item 4 is fairly difficult but can be guessed at with roughly a 30% chance of guessing correctly. Item 10 is not only fairly hard, but sharply discriminates between students of slightly varying higher abilities.

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 is given in Figure 5.

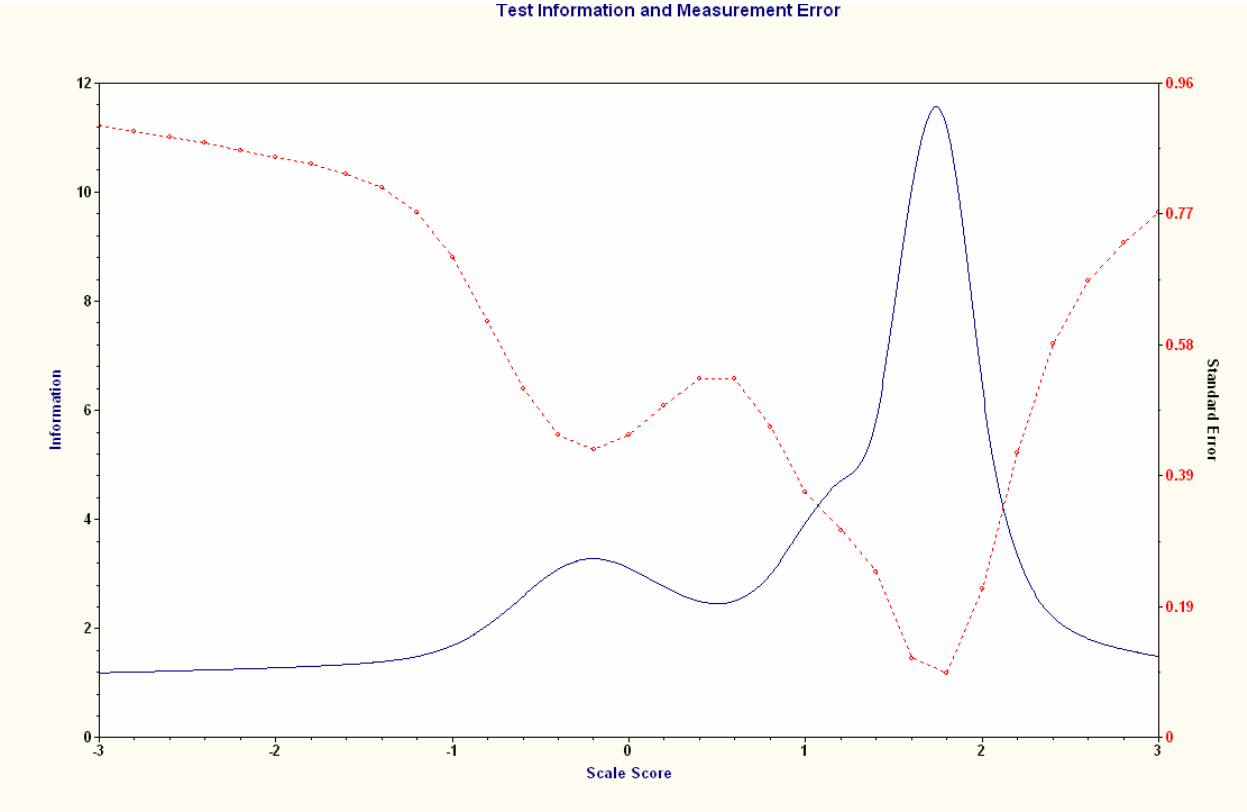

Figure 5. MSU test information curve

The solid blue line indicates "how much" information we can determine about a student's ability based on their scale score. 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.