Using test results

March 8th 2011, F.J.

Today's purposes

- Explore ways to "react to" test results
- Discuss options AND barriers to using test results
- Are 3rd party tools needed?
- Exchange suggestions / questions.
  - *IRT if time permits*
- I want to take notes about
  - What's really important
  - What's interesting

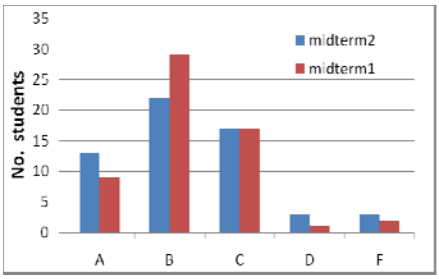
Why "using" test results?

- Teaching = helping make learning happen
- Helping = observing, AND reacting to, thinking
- Observing = data (from tests) and analysis
- Analysis = efficient/effective tools & procedures
- Teaching "scientifically" implies observing evidence and reacting (*intelligently* ☺) to it.

What do you NORMALLY do with marks?

- Average? Median? Stdev?
- Frequency distribution?
- Consider adjusting up? Down? To a "curve"?
- Do you scratch your head over why q'n *X* was answered poorly?

Frequency Distribution



What do you WISH you could do with results?

- Standard basic statistics?
- Question ("item") response analysis?
- Use results of analysis to find "problems"?
- Determine "reliability" of the test?
- Others?

Similarly:
What information COULD be useful?

- Discriminate between top & bottom students?
- Good questions … Bad questions?
- Hard questions … Easy questions?
- Drop questions? Why?
  - Gives misleading information about student abilities?
  - It compromises students' abilities to perform (took up time uselessly, established a misconception, others?)
- Other ideas?

What tools do you use for such exploration of testing results

- Excel?
- Roland's SCANTRON output files only?
- Vista analytics?
- Custom MatLab scripts?
- Others?

How many use scantron for testing?

- Do you use Roland's code?
- How do you use it's results?
- Counts of how often each option was chosen for all questions
- Example

Eg. Of EOS Scantron output file

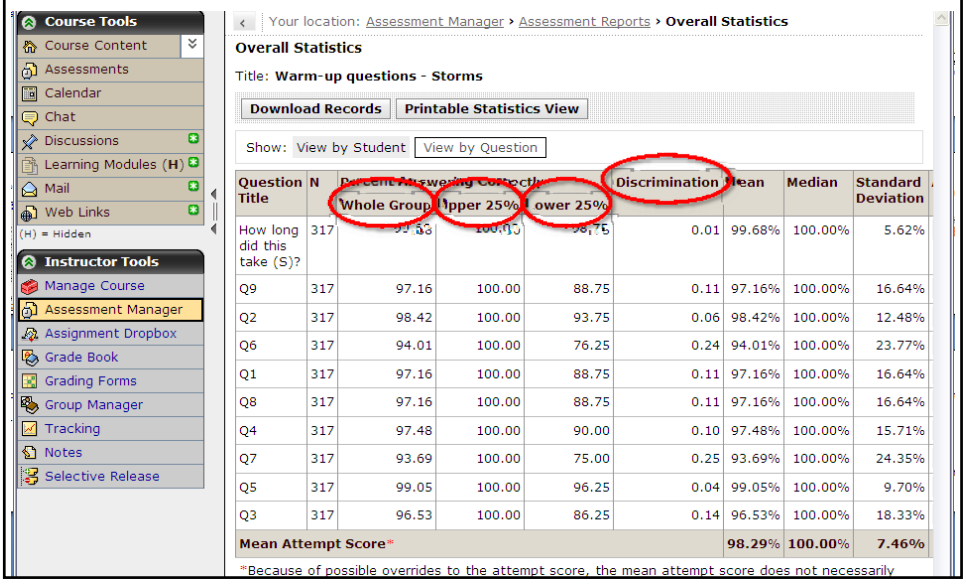
- The * and ** … What could these be saying about the test?
- Is there other information you could wish for?
- What if questions are NOT multiple choice?

```
* = questions missed by 50% or more students.
**= questions missed by 75% or more students.

 Check for Answer Key Errors

 ANSWER KEY: A
    Q)answer|  #A  #B  #C  #D  #E |#total #WRONG%
     1)  B  |  14  88  13  13   7 | 135  47  34
     2)  C  |  18   1 112   1   4 | 136  24  17
     3)  C  |  22   5  60   9  40 | 136  76  55 *
     4)  D  |  51  18   3  56   8 | 136  80  58 *
     5)  B  |  13 103   1  16   3 | 136  33  24
     6)  A  | 108  10  13   0   5 | 136  28  20
     7)  D  |   6   5  35  63  27 | 136  73  53 *
     8)  C  |   4  48  66   9   9 | 136  70  51 *
     9)  E  |   3   1  10   7 115 | 136  21  15
    10)  A  |  68  10  40   9   8 | 135  67  49
    11)  D  |  27  32   4  32  41 | 136 104  76 **
    12)  B  |  17 103   0   5  11 | 136  33  24
    13)  D  |  47  23   8  42  16 | 136  94  69 *
    14)  C  |  14  18  97   6   1 | 136  39  28
     .
     . Etc
     .
    51)  D  |  33   8   2  40  53 | 136  96  70 *
    52)  C  |   6  27  91   8   4 | 136  45  33
  Avg number of questions answered correctly:   34.0
  Avg preliminary SCORE % (  34.0 /  52.0)  =   65.4  C+
```

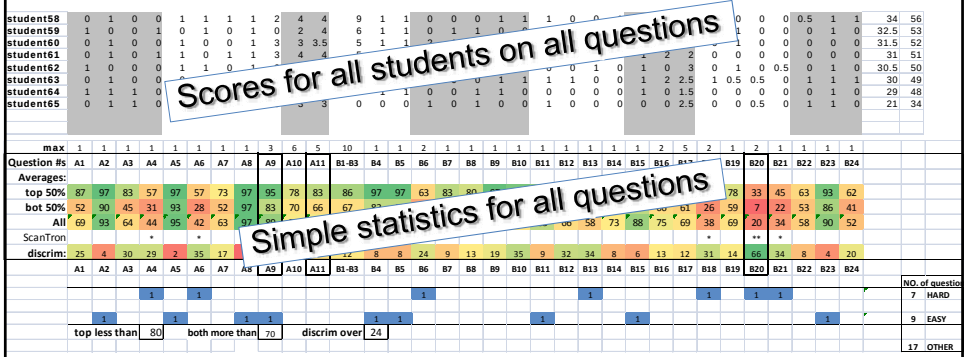
Does anyone use Vista's "reports" to see results patterns?



Question Title	N	Percent Answering Correctly: Whole Group	Upper 25%	Lower 25%	Discrimination	Mean	Median	Standard Deviation
How long did this take (S)?	317	99.53	100.00	98.75	0.01	99.68%	100.00%	5.62%
Q9	317	97.16	100.00	88.75	0.11	97.16%	100.00%	16.64%
Q2	317	98.42	100.00	93.75	0.06	98.42%	100.00%	12.48%
Q6	317	94.01	100.00	76.25	0.24	94.01%	100.00%	23.77%
Q1	317	97.16	100.00	88.75	0.11	97.16%	100.00%	16.64%
Q8	317	97.16	100.00	88.75	0.11	97.16%	100.00%	16.64%
Q4	317	97.48	100.00	90.00	0.10	97.48%	100.00%	15.71%
Q7	317	93.69	100.00	75.00	0.25	93.69%	100.00%	24.35%
Q5	317	99.05	100.00	96.25	0.04	99.05%	100.00%	9.70%
Q3	317	96.53	100.00	86.25	0.14	96.53%	100.00%	18.33%
Mean Attempt Score*						98.29%	100.00%	7.46%

*Because of possible overrides to the attempt score, the mean attempt score does not necessarily

Example: Spreadsheet analysis of questions.

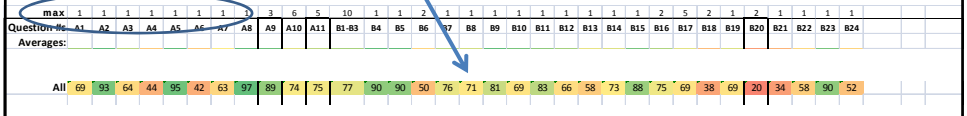


Comments:

1. **Red** averages and g… harder questions":
2. **Green** for both avera… questions (redish discrimination doesn't mean much)
3. **Red** averages and **red** discrimination values are "poor questions".
4. **Green** avg and **green** discrim is not possible since strong discrimination needs separation between good and bad students.
5. Negative discrimination values mean more "poor" students got it right than "good" students

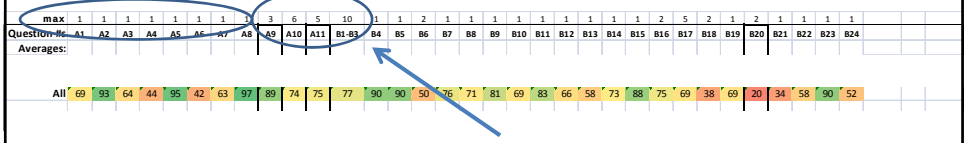
Calculate average score for each question

- MC questions …



Calculate average score for each question

- MC questions …



- … and other types of questions

Sort by students' exam scores

student58	0	1	0	0	1	1	1	1	2	4	4	9	1	1	0	0	0	1	1	1	0	0	1	0	0	1.5	0	0	0	0	0.5	1	1	34	56
student59	1	0	0	1	0	1	0	1	0	2	4	6	1	1	0	1	1	0	0	1	1	0	1	1	2	4.5	0	1	0	0	0	1	0	32.5	53
student60	0	1	0	0	1	0	0	1	3	3	3.5	5	1	1	2	1	1	1	0	1	1	1	0	0	2	0	1	1	0	0	0	0	0	31.5	52
student61	0	1	0	1	1	0	1	1	3	4	4	5	0	0	0	0	1	0	0	1	1	1	1	1	2	2	0	0	0	0	0	0	0	31	51
student62	1	0	0	0	1	1	0	1	3	4	5	3	1	1	0	0	1	1	0	0	0	1	0	1	0	3	0	1	0	0.5	0	1	0	30.5	50
student63	0	1	0	0	0	0	1	1	2	3	3.5	0	1	1	2	0	0	1	1	1	1	0	0	1	2	2.5	1	0.5	0.5	0	1	1	1	30	49
student64	1	1	1	0	1	0	0	1	2	3	4.5	7	1	1	0	0	1	1	0	0	0	0	0	1	0	1.5	0	0	0	0	0	1	0	29	48
student65	0	1	1	0	1	0	0	1	3	3	3	0	0	0	1	0	1	0	0	1	0	0	0	0	0	2.5	0	0	0.5	0	1	1	0	21	34

Scores for all students on all questions

Goal is to gather together top scoring students:

- Top 25% of students

- Top 33% … top 50% … depends on class size.

Also gather bottom scoring students (same numbers).

After sorting by total score …

Average for TOP students (top 50% in this case)

max	1	1	1	1	1	1	1	3	6	5	10	1	1	2	1	1	1	1	1	1	1	1	2	5	2	1	2	1	1	1	1
Question #s	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	B1-B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24
Averages:																																	
top 50%	87	97	83	57	97	57	73	97	95	78	83	86	97	97	63	83	80	97	93	90	87	77	78	93	85	77	49	78	33	45	63	93	62
All	69	93	64	44	95	42	63	97	89	74	75	77	90	90	50	76	71	81	69	83	66	58	73	88	75	69	38	69	20	34	58	90	52

After sorting by total score …

Average for TOP students (top 50% in this case)

Also average for BOTTOM students

max	1	1	1	1	1	1	1	3	6	5	10	1	1	2	1	1	1	1	1	1	1	1	2	5	2	1	2	1	1	1	1
Question #s	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	B1-B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24
Averages:																																	
top 50%	87	97	83	57	97	57	73	97	95	78	83	86	97	97	63	83	80	97	93	90	87	77	78	93	85	77	49	78	33	45	63	93	62
bot 50%	52	90	45	31	93	28	52	97	83	70	66	67	83	83	38	69	62	66	45	76	45	38	67	83	66	61	26	59	7	22	53	86	41
All	69	93	64	44	95	42	63	97	89	74	75	77	90	90	50	76	71	81	69	83	66	58	73	88	75	69	38	69	20	34	58	90	52

"Discrimination"

max	1	1	1	1	1	1	1	1	3	6	5	10	1	1	2	1	1	1	1	1	1	1	1	1	2	5	2	1	2	1	1	1	1
Question #s	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	B1-B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24
Averages:																																	
top 50%	87	97	83	57	97	57	73	97	95	78	83	86	97	97	63	83	80	97	93	90	87	77	78	93	85	77	49	78	33	45	63	93	62
bot 50%	52	90	45	31	93	28	52	97	83	70	66	67	83	83	38	69	62	66	45	76	45	38	67	83	66	61	26	59	7	22	53	86	41
All	69	93	64	44	95	42	63	97	89	74	75	77	90	90	50	76	71	81	69	83	66	58	73	88	75	69	38	69	20	34	58	90	52
discrim:	25	4	30	29	2	35	17	0	7	5	12	12	8	8	24	9	13	19	35	9	32	34	8	6	13	12	31	14	66	34	8	4	20

100 * (top - bottom) / (top + bottom)

Use discrimination to classify questions: Easy ….. Hard ….. Possibly "bad" ???

max	1	1	1	1	1	1	1	1	3	6	5	10	1	1	2	1	1	1	1	1	1	1	1	1	2	5	2	1	2	1	1	1	1
Question #s	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	B1-B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24
Averages:																																	
top 50%	87	97	83	57	97	57	73	97	95	78	83	86	97	97	63	83	80	97	93	90	87	77	78	93	85	77	49	78	33	45	63	93	62
bot 50%	52	90	45	31	93	28	52	97	83	70	66	67	83	83	38	69	62	66	45	76	45	38	67	83	66	61	26	59	7	22	53	86	41
All	69	93	64	44	95	42	63	97	89	74	75	77	90	90	50	76	71	81	69	83	66	58	73	88	75	69	38	69	20	34	58	90	52
ScanTron				*		*																					*		**	*			
discrim:	25	4	30	29	2	35	17	0	7	5	12	12	8	8	24	9	13	19	35	9	32	34	8	6	13	12	31	14	66	34	8	4	20

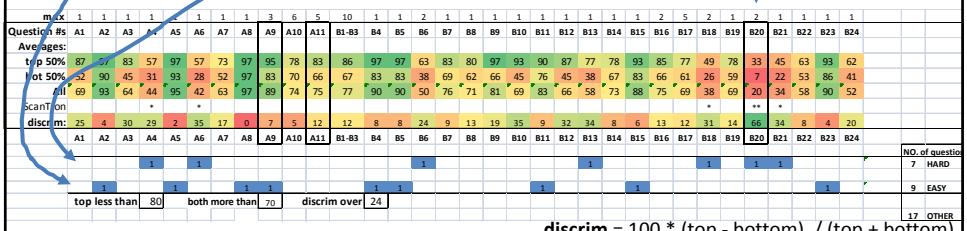
discrim = 100 * (top - bottom) / (top + bottom)

Comments:

1. **Red** *averages* and **green** *discrimination* values are "good harder questions":
2. **Green** for *both averages* are "easier" questions (redish discrimination doesn't mean much)
3. **Red** *averages* and **red** *discrimination* values are "poor questions".
4. **Green** *avg & discrim* NOT possible. (good students must DIFFER from bad students.)
5. **Negative** *discrimination* values? more "poor" students got it right than "good" students.

Use discrimination to classify questions:

Easy ….. Hard ….. Possibly “bad” …..



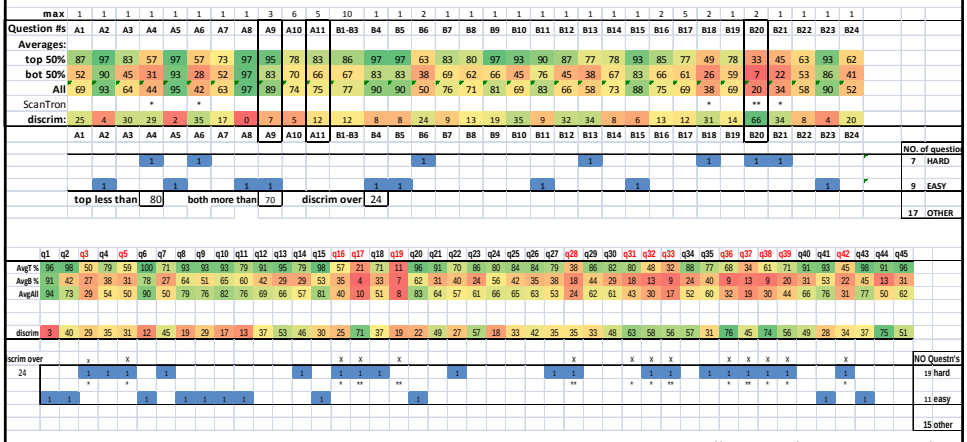
discrim = 100 * (top - bottom) / (top + bottom)

Hard based on: top < 80 & discrim > 24

Easy based on: both top and bottom > 70

Compare 2 tests

Is this “over analyzing” ?



Bottom test is all MC, with scantron results.

What should a test look like?

- Difficulty and order
  - Some easy questions?
  - First questions easier than later?
  - Some very hard questions?
- Do you "target" some kind of performance?
  - Competence limitations (for qualifying)?
  - Challenge strongest students?
  - Allow weakest students to pass?

Answering patterns (MC only)

From Scantron

- Check for:
  - Unclear options
  - Misleading qn's
  - Vague questions
  - Key misconcep'ns
  - >1 correct answers
  - What else ???

*** = questions missed by 50% or more students.**
****= questions missed by 75% or more students.**

Check for Answer Key Errors

ANSWER KEY: A

Q)answer	#A	#B	#C	#D	#E	#total	#WRONG	%	
1) B	14	88	13	13	7	135	47	34	
2) C	18	1	112	1	4	136	24	17	
3) C	22	5	60	9	40	136	76	55	*
4) D	51	18	3	56	8	136	80	58	*
5) B	13	103	1	16	3	136	33	24	
6) A	**108**	**10**	**13**	**0**	**5**	136	28	20	
7) D	6	5	35	63	27	136	73	53	*
8) C	4	48	66	9	9	136	70	51	*
9) E	3	1	10	7	115	136	21	15	
10) A	68	10	40	9	8	135	67	49	
11) D	**27**	**32**	**4**	**32**	**41**	136	104	76	**
12) B	17	103	0	5	11	136	33	24	
13) D	47	23	8	42	16	136	94	69	*
14) C	14	18	97	6	1	136	39	28	

- Find a Qn with few possible options …
- Find one with several "reasonable" options.

Answering patterns (MC only)

From spreadsheet:

Check for: Unclear options, Misleading qn's, Vague questions, Key misconcep'ns, >1 correct answers, Others ???

	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15	q16	q17	q18	q19	q20	q21	q22	q23	q24	q25	q26	q27	q28	q29	q30	q31	q32	q33	q34	q35	q36	q37	q38	q39	q40	q41	q42	q43	q44	q45
AvgT %	96	98	50	79	59	100	71	93	93	93	79	91	95	79	98	57	21	71	11	96	91	70	86	80	84	84	79	38	86	82	80	48	32	88	77	68	34	61	71	91	93	45	98	91	96
AvgB %	91	42	27	38	31	78	27	64	51	65	60	42	29	29	53	35	4	33	7	62	31	40	24	56	42	35	38	18	44	29	18	13	9	24	40	9	13	9	20	31	53	22	45	13	31
AvgAll	94	73	29	54	50	90	50	79	76	82	76	69	66	57	81	40	10	51	8	83	64	57	61	66	65	63	53	24	62	61	43	30	17	52	60	32	19	30	44	66	76	31	77	50	62
discrim	3	40	29	35	31	12	45	19	29	17	13	37	53	46	30	25	71	37	19	22	49	27	57	18	33	42	35	35	33	48	63	58	56	57	31	76	45	74	56	49	28	34	37	75	51
dropped			x		x											x	x		x												x	x	x			x	x	x	x			x			
ScanTron			*		*											*	**		**												**	*	*	**			*	**	*	*			*			
	D	A	C	C	D	C	C	E	B	C	A	C	B	E	A	E	B	C	A	B	D	C	B	A	A	E	A	C	A	A	D	A	C	B	A	C	A	C	E	B	E	A	C	C	E
A	4	162	0	15	33	10	42	9	4	9	167	4	31	17	181	15	25	27	17	18	33	19	21	146	143	24	118	3	137	135	34	67	61	6	132	73	42	38	14	47	21	68	12	3	40
B	4	14	17	33	24	6	4	5	167	4	14	13	146	27	10	15	21	76	93	183	16	67	134	28	29	34	15	6	21	73	26	16	25	116	33	31	11	51	10	146	9	88	23	22	2
C	5	19	65	120	43	201	110	2	5	181	29	153	12	30	27	47	90	114	41	4	30	125	8	1	20	14	42	54	30	3	63	42	38	83	21	71	74	66	72	19	3	25	170	111	24
D	208	25	6	15	110	1	31	32	32	24	5	44	18	20	4	56	62	2	32	4	142	7	57	7	15	10	8	140	23	8	94	22	73	17	31	31	82	22	28	5	19	33	14	68	16
E	1	2	133	38	10	4	32	174	14	4	7	7	15	127	0	89	23	3	38	13	1	4	2	39	15	139	37	19	11	3	5	75	24	0	5	16	12	44	97	5	169	7	2	17	138

Helps answer "why" the question didn't work.

Eg. $^{17}/_{221}$ chose "correct" answer

Answering patterns (MC only)

Find a question

- with very few "viable" options …
- With all options equally "possible"

	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15	q16	q17	q18	q19	q20	q21	q22	q23	q24	q25	q26	q27	q28	q29	q30	q31	q32	q33	q34	q35	q36	q37	q38	q39	q40	q41	q42	q43	q44	q45
AvgT %	96	98	50	79	59	100	71	93	93	93	79	91	95	79	98	57	21	71	11	96	91	70	86	80	84	84	79	38	86	82	80	48	32	88	77	68	34	61	71	91	93	45	98	91	96
AvgB %	91	42	27	38	31	78	27	64	51	65	60	42	29	29	53	35	4	33	7	62	31	40	24	56	42	35	38	18	44	29	18	13	9	24	40	9	13	9	20	31	53	22	45	13	31
AvgAll	94	73	29	54	50	90	50	79	76	82	76	69	66	57	81	40	10	51	8	83	64	57	61	66	65	63	53	24	62	61	43	30	17	52	60	32	19	30	44	66	76	31	77	50	62
discrim	3	40	29	35	31	12	45	19	29	17	13	37	53	46	30	25	71	37	19	22	49	27	57	18	33	42	35	35	33	48	63	58	56	57	31	76	45	74	56	49	28	34	37	75	51
dropped			x		x											x	x		x												x	x	x			x	x	x	x			x			
ScanTron			*		*											*	**		**												**	*	*	**			*	**	*	*			*			
	D	A	C	C	D	C	C	E	B	C	A	C	B	E	A	E	B	C	A	B	D	C	B	A	A	E	A	C	A	A	D	A	C	B	A	C	A	C	E	B	E	A	C	C	E
A	4	162	0	15	33	10	42	9	4	9	167	4	31	17	181	15	25	27	17	18	33	19	21	146	143	24	118	3	137	135	34	67	61	6	132	73	42	38	14	47	21	68	12	3	40
B	4	14	17	33	24	6	4	5	167	4	14	13	146	27	10	15	21	76	93	183	16	67	134	28	29	34	15	6	21	73	26	16	25	116	33	31	11	51	10	146	9	88	23	22	2
C	5	19	65	120	43	201	110	2	5	181	29	153	12	30	27	47	90	114	41	4	30	125	8	1	20	14	42	54	30	3	63	42	38	83	21	71	74	66	72	19	3	25	170	111	24
D	208	25	6	15	110	1	31	32	32	24	5	44	18	20	4	56	62	2	32	4	142	7	57	7	15	10	8	140	23	8	94	22	73	17	31	31	82	22	28	5	19	33	14	68	16
E	1	2	133	38	10	4	32	174	14	4	7	7	15	127	0	89	23	3	38	13	1	4	2	39	15	139	37	19	11	3	5	75	24	0	5	16	12	44	97	5	169	7	2	17	138

Summary so far – questions?

- Why analyze answering behavior of exams?
  - Identify hard, easy, medium questions?
  - Identify common misconceptions?
  - Others?
- Using *, **, top/bot avgs & discriminator?
- Checking answer patterns?
- Questions ? ? ?

Precedent? Tools?

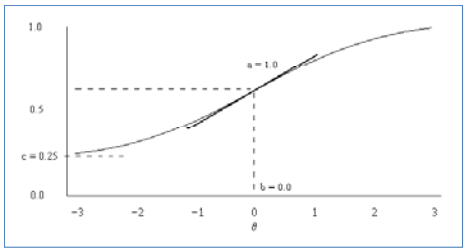
- Opensource JAVA software for test analytics
  - Basic descriptive stats, graphs, classical item analysis, factor analysis, and item response theory (ITR):
  - Feature list: http://www.itemanalysis.com/jmetrik-features.php
- Microsoft Excel plug-in for test item analysis:
  - Feature list: http://deltasigmasoft.com/index.html
  - $0.00: # test items <= 25; # students <=500.
    $30.0: (or $150 for Dep't) no restrictions.
- Google "*item analysis*" or "*item analysis software*".
- http://en.wikipedia.org/wiki/Psychometric_software

IRT – Item Response Theory

- Provides parameters for questions independent of the ability level of the student
- Commonly used to characterize large-scale, high-stakes assessments such as SATs.
- Analysis of all exam data yields an "item characteristic curve" for *each* question, which describes:
  - Question's difficulty or "ability level", and
  - Question's ability to discrimination between students with abilities above and below the question's "ability level".

  - http://en.wikipedia.org/wiki/Item_response_theory ,
  - http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm .

IRT
Question curves

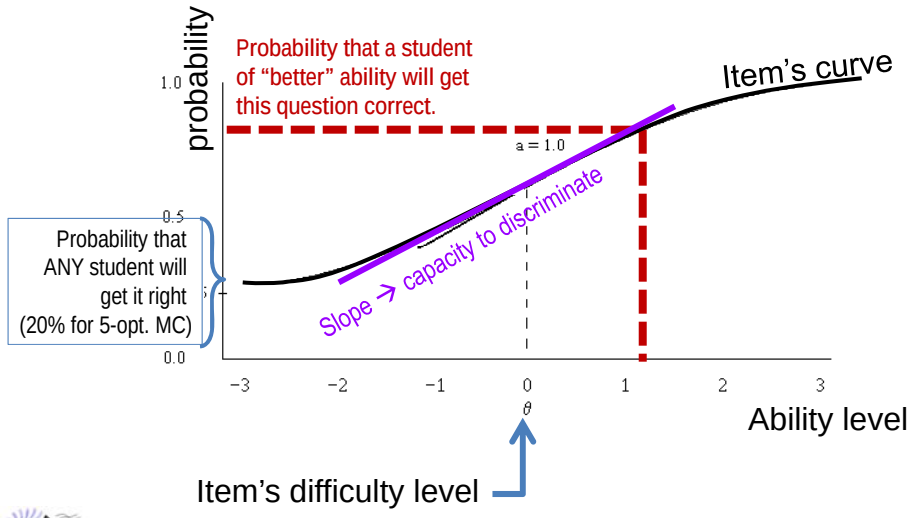


- Each question's curve yields the **probability** (*Y-axis*) that a student with a given ability level (*X-axis*) will answer the question correctly.
- Discriminating ability is described by maximum slope, α.
- Steeper slope (large α) ➔ item is better able to differentiate between students of lower and higher ability.

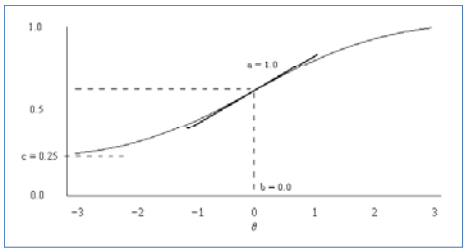
Application example: Create "designer tests":

- Med school , where high performance is required;
  Choose most questions at the upper end of the difficulty spectrum. In other words, questions should have steep α so that only students performing at a high x-axis level will pass.

One question's (item's) IRT curve



IRT Caveats:



- Tested skill must not be confused by need for other skills, such as English, comprehending the question, working quickly, etc.
- There should be no tendency for different groups to generate different results. Examples: gender or racial biases.
- All questions must be answered under similar "stress" (i.e. no rushing near end, etc).

Implications

- IRT can not be used to characterize questions in exams that test many concepts.
- HOWEVER – IRT might help build consistent exams from a pool of many questions. This assumes a DataBase of many validated questions for each concept.

After grading, entering and analyzing marks, *what do you do with midterms*?

- Do you return students' work to them?
  - Why? Why not?
- How much feedback goes on tests you return?
  - Individual for each student?
  - Collective feedback (for whole class)?
  - Via a rubric?
  - Do students see rubric before marking?