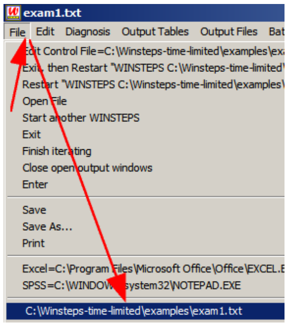
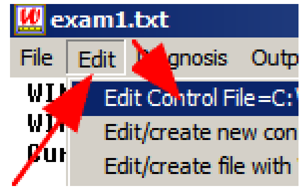
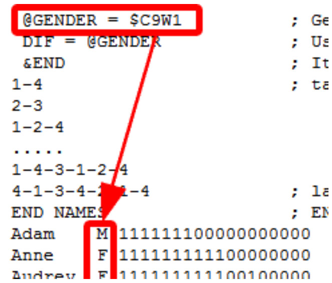
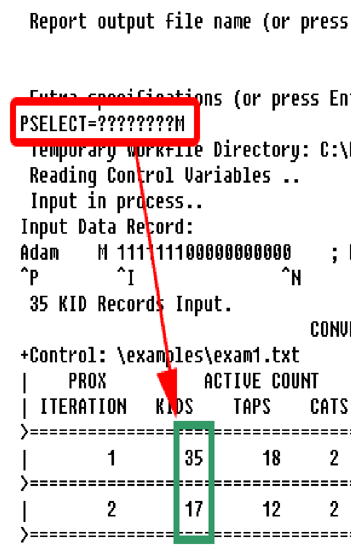
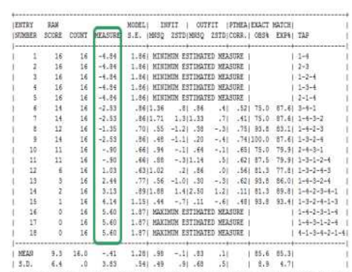
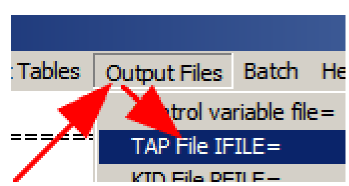
<table>
<tr><th>#</th><th colspan="2">Winsteps Tutorial 4
Mike Linacre, instructor – June 2012</th></tr>
<tr><td>1.</td><td colspan="2">Tutorial 4.
Almost to the summit !
Differential Test Functioning
Differential Item Functioning
Investigating Dimensionality
This lesson builds on Lessons 1, 2 and 3, so please go back and review when you need to. If you run into difficulties or want to talk about what you are learning, please post to the Discussion Forum.
http://www.winsteps.com/forum
Also if you have questions about a Rasch topic not covered in the Course ... Discussion Forum!</td></tr>
<tr><td>2.</td><td colspan="2">A. Differential Test Functioning (DTF)</td></tr>
<tr><td>3.</td><td colspan="2">The Rasch model embodies an ideal - an ideal which is always “violated” by real data. Many types of violation are inconsequential, no matter how “statistically significant” they are, but some really do have serious substantive consequences. If the items function consistently differently for one group of test-takers than for another, then the measures for one group may not be comparable with the measures for another.
Let's imagine we administer our test to boys and girls. Here are two questions we can ask:
a. Differential Test Functioning: Does our test, consisting of all the items, function the same way for boys and girls? We investigate this by doing separate analyses for boys and girls, and then comparing the two sets of item difficulties.
b. Differential Item Functioning: Does item 3 function the same way for boys and girls (assuming that everything else does function the same way)? We investigate this by estimating two difficulties for item 3 (boys' difficulty and girls' difficulty) while holding all the other item difficulties and person measures unchanged.

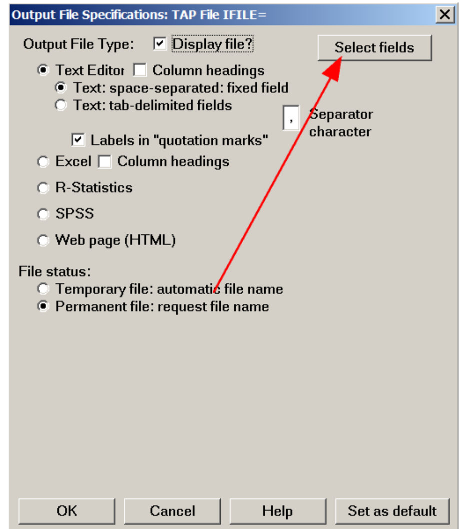
Let’s illustrate Differential Test Functioning ...</td></tr>
<tr><td>4.</td><td>Launch Winsteps</td><td></td></tr>
<tr><td>5.</td><td>Control File: exam1.txt

Stop! No further now!

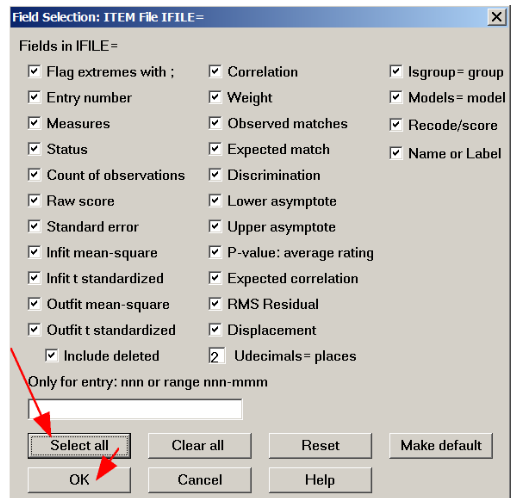
“Report output ..” displays. But don’t respond to it yet.</td><td>
</td></tr>
</table>

6.	Winsteps menu bar Click on “Edit” menu Click on “Edit Control File”	
7.	Notice that column 9 of the person label is a gender indicator. M=Male, F=Female. We will analyze this dataset twice. Once for the males and once for the females. Then we will compare the two sets of item difficulties.	
8.	Click on Winsteps Analysis window or Winsteps in the task bar: Winsteps window displays: Report Output ...? Press Enter Extra Specifications? **PSELECT=????????M** Press Enter “?” means “any character”. So PSELECT=????????M means: “the first 8 columns of the person label can have any character, then select only person records with an M in column 9” Notice that the estimation process starts with 35 children, and then drops to 17. Only half the sample is active - as we would expect.	
9.	Take a look at the item difficulties. Windows Menu Bar Click on Output Tables Click on Table 14. What we want are the item measures in the green box. These are the item difficulties for the Males.	
10.	Let’s write the item difficulties to disk. Winsteps Menu Bar Click on “Output Files” Click on “TAP File IFILE=” *(You know this! We’ve been here before ...)*	

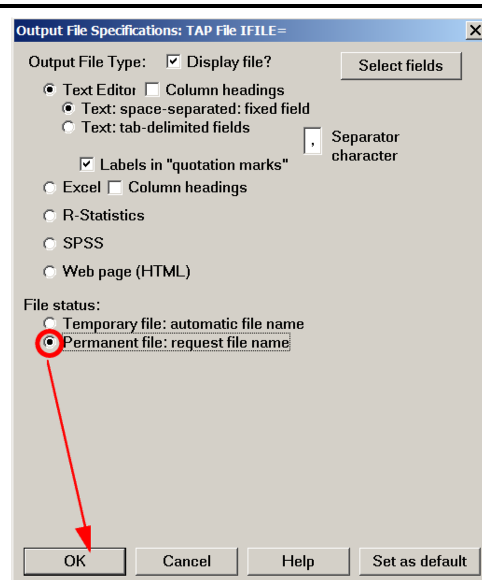
11. In the IFILE= dialog box,
click on Select Fields



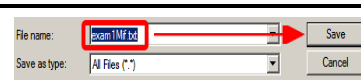
12. Click on "Select All"
Click on "OK"



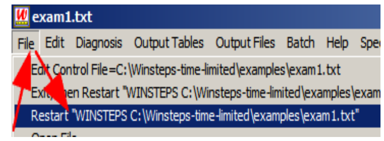
13. In the IFILE= dialog box,
we want a permanent item file



14. Let's call our file: **exam1Mif.tx**t
Then "Save"
This file has the item difficulties for the Males



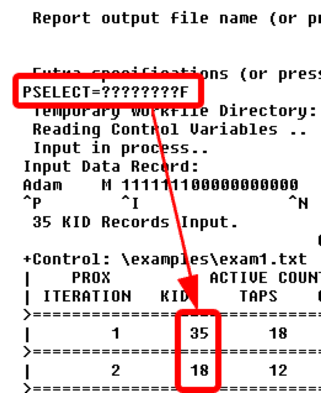
15. Now for the Females:
Winsteps Menu Bar
Restart Winsteps exam1.txt



16. Report Output? Press Enter
Extra Specifications?
PSELECT=????????F
Press Enter

We've selected the females in the sample. Notice that the number of KIDS has reduced from 35 to 18 - the other half of the sample.

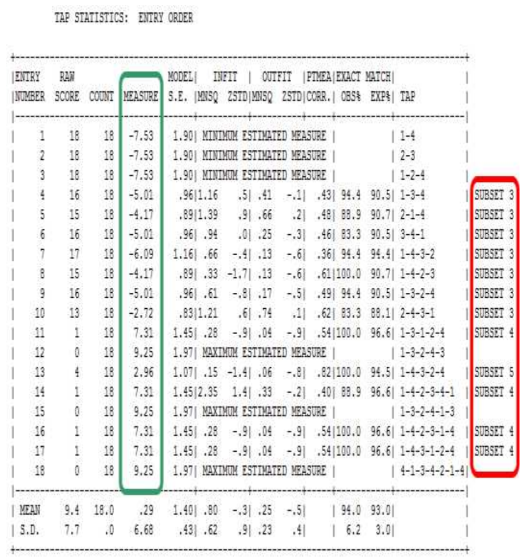
We have now estimated the item difficulties for these 18 children



17. Take a look at the item difficulties.
Windows Menu Bar
Click on Output Tables
Click on Table 14.

What we want are the item measures in the green box. These are the item difficulties for the Females

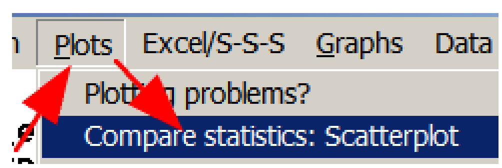
What we don't like to see are the **SUBSET** warnings in the red box. **"Subsets"** mean the dataset has fractured into incomparable rectangles of data. We will ignore that right now, but **please don't ignore subsetting in a real analysis, because your measures may be invalid.** See the Help file Special Topic "Subsets ..." also at http://www.winsteps.com/winman/connection.htm

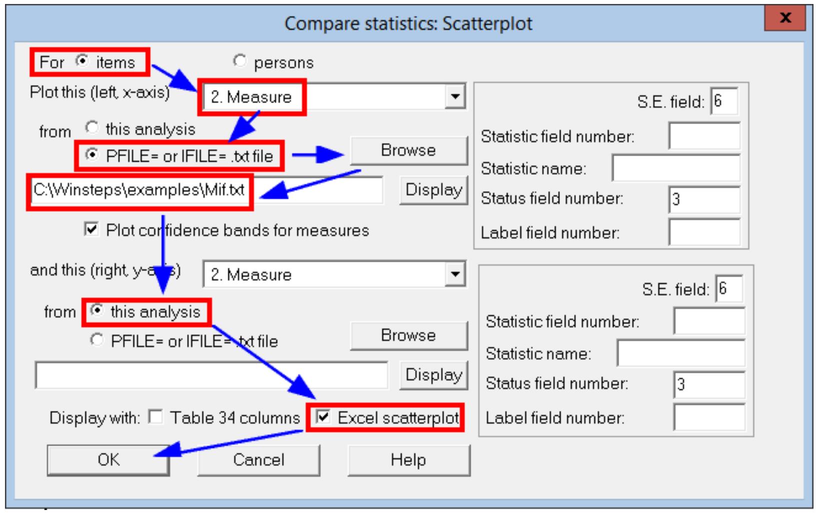
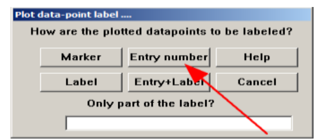


TAP STATISTICS: ENTRY ORDER

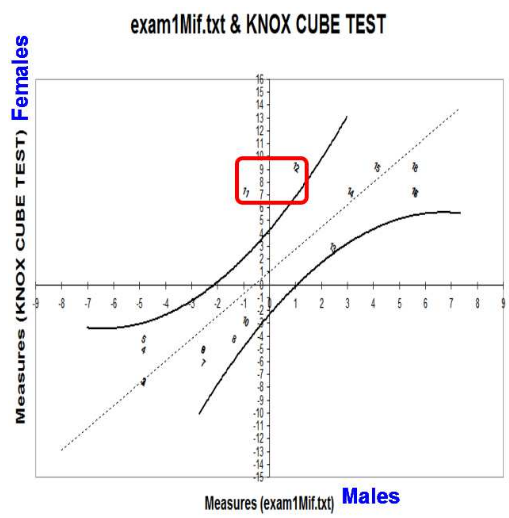
ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	TAP	
1	18	18	-7.53	1.90	MINIMUM ESTIMATED MEASURE							1-4	
2	18	18	-7.53	1.90	MINIMUM ESTIMATED MEASURE							2-3	
3	18	18	-7.53	1.90	MINIMUM ESTIMATED MEASURE							1-2-4	
4	16	18	-5.01	.96	1.16	.5	.41	-.1	.43	94.4	90.5	1-3-4	SUBSET 3
5	15	18	-4.17	.89	1.39	.9	.66	.2	.48	88.9	90.7	2-1-4	SUBSET 3
6	16	18	-5.01	.96	.94	.0	.25	-.3	.46	83.3	90.5	3-4-1	SUBSET 3
7	17	18	-6.09	1.16	.66	-.4	.13	-.6	.36	94.4	94.4	1-4-3-2	SUBSET 3
8	15	18	-4.17	.89	.33	-1.7	.13	-.6	.61	100.0	90.7	1-4-2-3	SUBSET 3
9	16	18	-5.01	.96	.61	-.8	.17	-.5	.49	94.4	90.5	1-3-2-4	SUBSET 3
10	13	18	-2.72	.83	1.21	.6	.74	.1	.62	83.3	88.1	2-4-3-1	SUBSET 3
11	1	18	7.31	1.45	.28	-.9	.04	-.9	.54	100.0	96.6	1-3-1-2-4	SUBSET 4
12	0	18	9.25	1.97	MAXIMUM ESTIMATED MEASURE							1-3-2-4-3	
13	4	18	2.96	1.07	.15	-1.4	.06	-.8	.82	100.0	94.5	1-4-3-2-4	SUBSET 5
14	1	18	7.31	1.45	2.35	1.4	.33	-.2	.40	88.9	96.6	1-4-2-3-4-1	SUBSET 4
15	0	18	9.25	1.97	MAXIMUM ESTIMATED MEASURE							1-3-2-4-1-3	
16	1	18	7.31	1.45	.28	-.9	.04	-.9	.54	100.0	96.6	1-4-2-3-1-4	SUBSET 4
17	1	18	7.31	1.45	.28	-.9	.04	-.9	.54	100.0	96.6	1-4-3-1-2-4	SUBSET 4
18	0	18	9.25	1.97	MAXIMUM ESTIMATED MEASURE							4-1-3-4-2-1-4	
MEAN	9.4	18.0	.29	1.40	.80	-.3	.25	-.5		94.0	93.0		
S.D.	7.7	.0	6.68	.43	.62	.9	.23	.4		6.2	3.0		

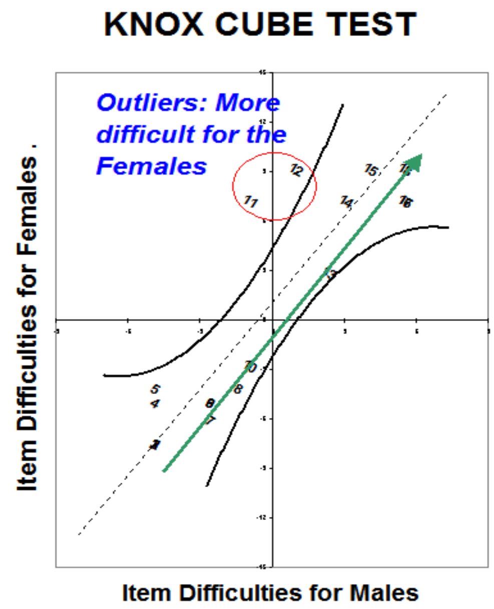
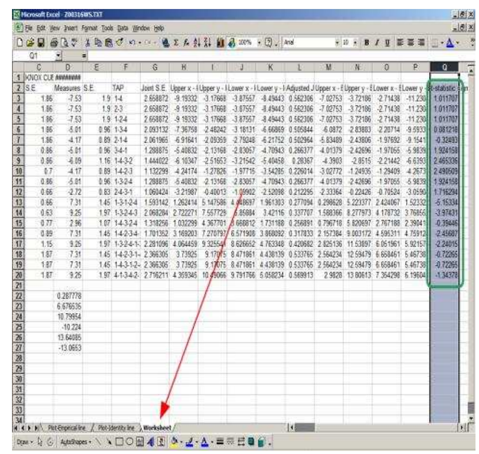
18. Let's compare these item difficulties for the females with those for the males. A good way is an Excel scatter plot:
Windows Menu Bar
Click on Plots
Click on Compare Statistics

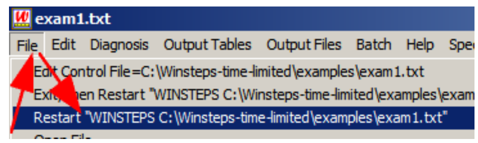
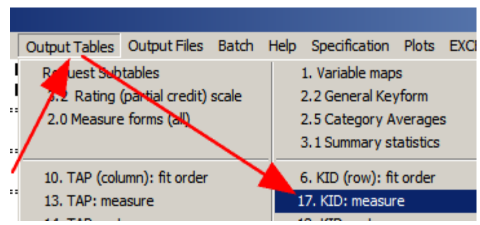
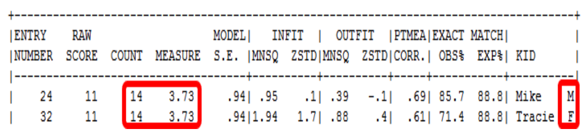
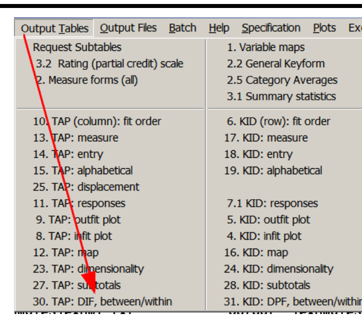
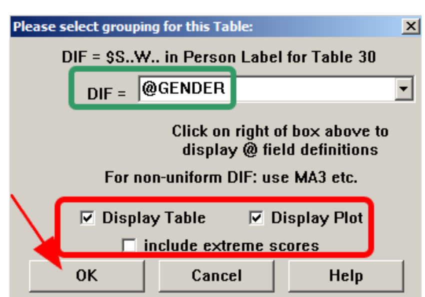


19.	We want to cross-plot the item difficulty measures. On the x-axis we will put the Male item difficulty measures from the item file (IFILE): exam1Mif.txt On the y-axis will put the Female item difficulty measures from this analysis. We want an Excel scatter plot. Click OK	
20.	Let's identify the points on the plot by their item entry numbers. Click on Entry number	

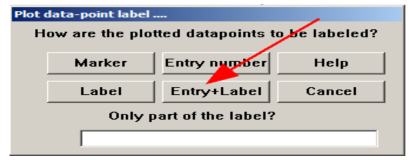
21.	The Winsteps Analysis screen informs us the Excel is processing. Winexcel.exe is the Winsteps-Excel interface.	`Transferring Scatterplot of 18 data lines to Excel ....`
		`Loading winexcel.exe ...`
		`Scatterplot process launched`

22.	The Excel scatterplot displays. Each point is an item. The y-axis is the item difficulties for the Females. The x-axis for the Males. The dashed line is a trend-line through the mean of both sets of items. The black curved lines are approximate 95% confidence bands. They are approximate because each point has its own confidence interval. "Best Test Design" discusses this plot around p. 83. Look at those upper points: items 11 and 12. Are they more difficult for the Males or the Females? You can use the full Excel functionality to beautify the plot ...	

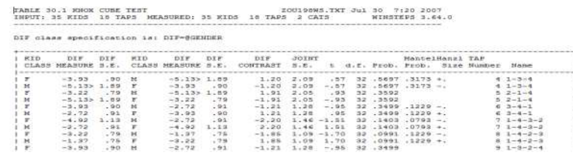
23.	This plot compares the performance of the two tests as depicted by the items: “**Differential Test Functioning**” (DTF). This approach is useful because it gives us a better sense of test bias than inspection of individual items. “A line of commonality” is equally good at predicting x values from y values, and y values from x values, unlike a regression line which is only good at predicting one way. The **dotted** line is a line of commonality through the mean of both distributions and also (mean+2 S.D.) of both distributions. Items 11 and 12 are relatively much more difficult for the Females and so are clearly outliers. They could be removed from the comparison in a “purifying” step. With or without their removal, inspection of the plot shows that a **better** line of commonality between the males and females is near the **green** arrow. I drew the green line by eye. To me the green line summarizes the “true” relationship between males anf females.	
24.	The numbers used to make the plot are in the Worksheet. So you can use them directly. Column Q is the Student’s t-statistic [if this is new to you, see Appendix A] which tests the hypothesis that the difficulties of each pair of items in the two analyses are statistically the same. This type of analysis is too prone to accidents of the data when the sample size is small (as in this example), so the item comparisons are only meaningful with large samples. For large samples, the degrees of freedom, d.f., can be treated as infinite, so that t-statistics become unit-normal deviates with a 95% confidence level of 1.96.	

25.	**B. Differential Item Functioning (DIF) - "Uniform DIF"**	
26.	The more usual approach to investigating item functioning is to do this in the context of one combined analysis, not two separate analyses. This also avoids the "SUBSET" problem we encountered above. The combined analysis also makes better sense scientifically. In most science experiments, the plan is to only allow one thing to alter at once, and then to detect the effect of that alteration. Here this would mean holding all other item and person measures constant, while investigating the interaction of the person-groups with each item in turn. In the DTF approach, two entirely separate sets of item and person measures underlie the comparisons. Let's see how the combined DIF approach works here.	
27.	Winsteps Menu Bar Restart Winsteps .... exam1.txt Report Output? Press Enter Extra Specifications? Press Enter The analysis is performed.	
28.	We now have one set of item difficulties and one set of person measures. Take a look at the person measures: Winsteps Menu Bar Click on Output Tables Click on 17. KID: measure (order)	
29.	Do you see that Males and Females with the same raw score have the same measure? We are measuring everyone with the same "ruler".	
30.	Now let's perform a DIF analysis: Winsteps Menu Bar Click on Output Tables Click on 30: TAP: DIF	
31.	@GENDER is defined as column 9 of the person label by the statement in the control file: @GENDER = $C9W1 We could have keyed $C9W1 instead of @GENDER, or even simply 9. We want to display the Output Table 30, also the Excel Plot. Click OK.	

32. Let's identify the plotted items by their entry numbers and labels.



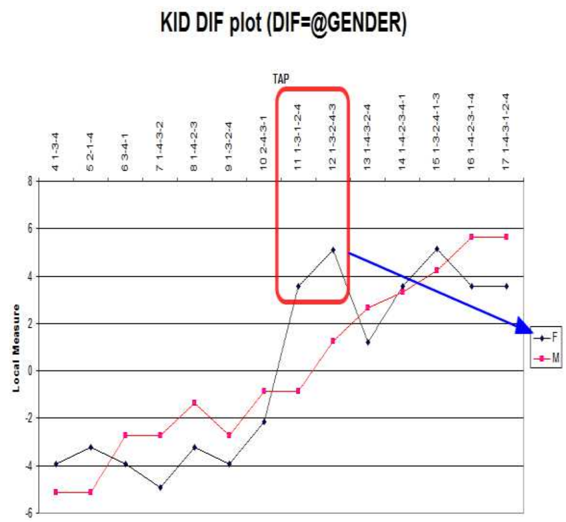
33. Table 30 displays first, and then the Excel plots, but it usually makes more sense to look at the plots first. They tell us where to look in the numerical tables for the interesting numbers.



```
TABLE 30.1 KNOX CUBE TEST                         ZOU198WS.TXT Jul 30  7:20 2007
INPUT: 35 KIDS  18 TAPS  MEASURED: 35 KIDS  18 TAPS  2 CATS        WINSTEPS 3.64.0
--------------------------------------------------------------------------------

DIF class specification is: DIF=@GENDER

+-----------------------------------------------------------------------------------------------------+
| KID    DIF    DIF   KID    DIF    DIF     DIF    JOINT                MantelHanzl TAP              |
| CLASS MEASURE S.E.  CLASS MEASURE S.E.  CONTRAST  S.E.    t  d.f. Prob. Prob.  Size Number  Name    |
|-----------------------------------------------------------------------------------------------------|
| F      -3.93   .90  M      -5.13> 1.89     1.20  2.09   .57   32 .5697 .3173 +.        4 1-3-4     |
| M      -5.13> 1.89  F      -3.93   .90    -1.20  2.09  -.57   32 .5697 .3173 -.        4 1-3-4     |
| F      -3.22   .79  M      -5.13> 1.89     1.91  2.05   .93   32 .3592                 5 2-1-4     |
| M      -5.13> 1.89  F      -3.22   .79    -1.91  2.05  -.93   32 .3592                 5 2-1-4     |
| F      -3.93   .90  M      -2.72   .91    -1.21  1.28  -.95   32 .3499 .1229 -.        6 3-4-1     |
| M      -2.72   .91  F      -3.93   .90     1.21  1.28   .95   32 .3499 .1229 +.        6 3-4-1     |
| F      -4.92  1.13  M      -2.72   .91    -2.20  1.46 -1.51   32 .1403 .0793 -.        7 1-4-3-2   |
| M      -2.72   .91  F      -4.92  1.13     2.20  1.46  1.51   32 .1403 .0793 +.        7 1-4-3-2   |
| F      -3.22   .79  M      -1.37   .75    -1.85  1.09 -1.70   32 .0991 .1229 -.        8 1-4-2-3   |
| M      -1.37   .75  F      -3.22   .79     1.85  1.09  1.70   32 .0991 .1229 +.        8 1-4-2-3   |
| F      -3.93   .90  M      -2.72   .91    -1.21  1.28  -.95   32 .3499                 9 1-3-2-4   |
```

34. The Excel plot is highly informative. The difficulty of each item for each group is estimated, while holding constant all the other item difficulty and person ability measures. The pattern is clear. Items 11 and 12 are over 3 logits more difficult for the Females ***in this sample!*** (Or we could say, "Females are over 3 logits less able on Items 11 and 12 *in this sample.)* DIF studies are notorious for producing non-replicable findings. So don't immediately throw out these items as "biased against girls". Next time they may show no bias at all.

We can think of DIF sizes in two ways: 1. relative to the overall item difficulty, or 2. combined with the overall item difficulty as an absolute measure, local to the group, on the latent variable.



35. Now we know where we want to look in Table 30.

In Table 30.1, look at the lines for items 11 and 12. Each item is shown twice, but the pairs of lines have the same meaning.

In the red box, Item 11 is 3.61 logits more difficult for the girls than the boys. Even allowing for the small sample size, which causes large item difficulty standard errors and small degrees of freedom, the t-statistic for item 11 is highly significant p<.01.

For item 12 it is almost as significant, p=.028. This DIF CONTRAST size is 3.61 logits for item 11. A huge difference, but remember that **DIF impact on the person measures will depend on the length of the test.** A 4 logit bias will have average impact on person abilities in the group of 4/10 = .4 logits in a 10-item test, but only .04 logits for a 100 item test. Statistical significance gives no indication at all of the actual impact of DIF on the person measures. **A small DIF size can be highly significant, and a large DIF size may be reported as not statistically significant.**

From the Week 1 Tutorial:

Logit Difference	Probability of Success	Logit Difference	Probability of Success
5.0	99%	-5.0	1%
4.6	99%	-4.6	1%
4.0	98%	-4.0	2%
3.0	95%	-3.0	5%
2.2	90%	-2.2	10%
2.0	88%	-2.0	12%
1.4	80%	-1.4	20%
1.1	75%	-1.1	25%
1.0	73%	-1.0	27%
0.8	70%	-0.8	30%
0.5	62%	-0.5	38%
0.4	60%	-0.4	40%
0.2	55%	-0.2	45%
0.1	52%	-0.1	48%
0.0	50%	-0.0	50%

If I have a 50% chance of success on item 11, and it is 3.61 logits easier for you, then you have a 97% chance of success on that item.

```
---------------------------------------------------------------------------------------------------------
| KID    DIF   DIF   KID    DIF   DIF      DIF    JOINT                 MantelHanzl TAP                |
| CLASS MEASURE S.E.  CLASS MEASURE S.E.  CONTRAST  S.E.    t  d.f. Prob. Prob.  Size Number   Name      |
|--------------------------------------------------------------------------------------------------------|
| F      2.79   .78  M     -.82   .65     3.61  1.02  3.54  32 .0012 .0004 +.        11 1-3-1-2-4     |
| F      4.19< 1.11  M     1.24   .64     2.95  1.29  2.30  32 .0284 .0256 +.        12 1-3-2-4-3     |
```

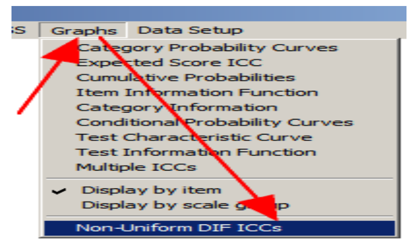
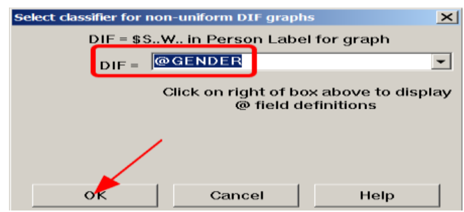
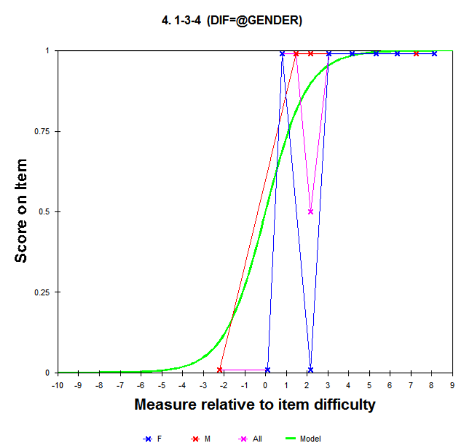
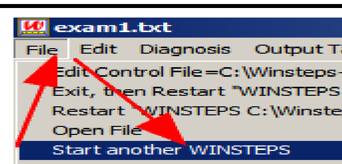
36. ***Practical challenge:***

1. Start a new copy of Winsteps.
2. Analyze the Example0.txt (Liking for Science) data.
3. Do a DIF analysis on Gender.
4. Which item is the most problematic?

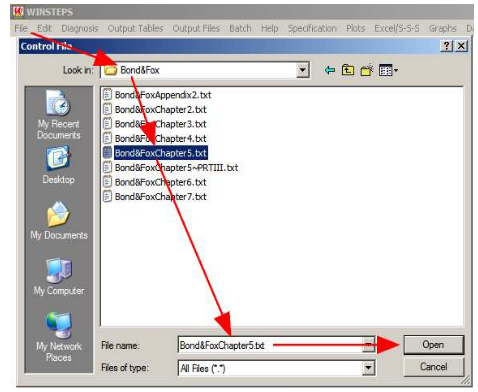
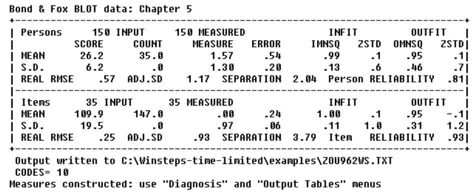
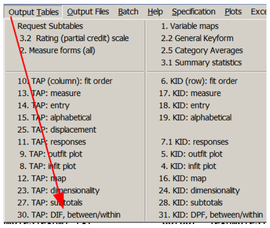
37. ***Optional:* Mantel-Haenszel (for dichotomies) and Mantel (for polytomies) statistics**

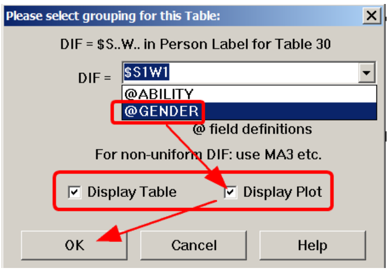
In the **green** box, Mantel-Haenszel (MH) DIF statistics are reported. MH is really way outside the scope of this Course, but is included in Winsteps because so many editors, reviewers, funding agencies, etc. insist on it. This is because ETS, the most prestigious test developer, use it in their DIF studies.
MH is a log-odds estimator based on cross-tabulating data at different levels of ability and then summarizing the cross-tabulations.
Conventionally, when the data are complete, the raw score is the indicator of ability. So there is one cross-tabulation for every possible non-extreme raw score on the test.
In Winsteps, the ability measures are the indicators of ability (so allowing the statistic to be applied to incomplete data). The ability range is stratified into ability levels. A cross-tabulation is computed for each ability level (stratum). These are then summarized into "size" and "significance" statistics.
In this dataset there aren't enough subjects for the cross-tabulations to be meaningful. Consequently, though the MH probability is reported, the size of the DIF according to MH is not estimated, only its direction is shown "+" (more difficult for the first group on the line, the Females). Winsteps has a control variable MHSLICE= which may make the MH DIF size estimable. Please see the Winsteps Help file if this is of concern to you. In this example, we can see that MH and Rasch are telling close to the same story. Since the tendency is to over-report DIF, I would favor reporting the DIF indicator with the largest reported probability (i.e., the least significant probability) for each DIF effect, but choose only one family of DIF indicators to report in your paper or report. Consider your audience. Most DIF-aware audiences prefer MH statistics, if those statistics are computable.

38.	***Optional:* Bonferroni corrected *t*-tests or not?** A participant asked: "Why or why not are adjustments for multiple tests needed for the evaluation of the DIF contrast *t*-test or Mantel-Haenzel (I'm thinking of concerns coming from a Journal reviewer on this)?" Yes, reviewers are sticklers about this. First, we need to state the exact hypothesis we are testing. If the hypothesis is: "The bias against boys relative to girls for item 1 is explained by chance", then the Winsteps *t*-test or MH significance in Table 30.1 is correct. No adjustment is necessary. If the hypothesis is: "The bias against boys relative to the overall difficulty of item 1 is explained by chance", then the Winsteps t-test in Table 30.2 is correct. No adjustment is necessary. If the hypothesis is: "The bias against boys on all items (relative to their overall difficulty) is explained by chance", then we need a Bonferroni (or similar) adjustment to Table 30.1 If the hypothesis is: "The bias against boys relative to girls on all items is explained by chance", then we need a Bonferroni (or similar) adjustment to Table 30.2 In practice, the "all items" hypotheses are almost useless. This is because our investigation rapidly focuses on specific items. "The girls seem to have performed much better than the boys on item 4. Is there a problem?"

39.	**C. Non-Uniform DIF**	
40.	Uniform DIF estimates DIF on the basis that the change in an item's difficulty is uniform (the same) across different ability levels of the person groups. But perhaps that isn't the case. **If difficulty varies with ability for a person-group, then there is "Non-Uniform DIF".** Let's look for this. Non-uniform DIF is DIF at different ability levels. "Uniform" DIF is DIF for all ability levels ≈ average "Non-uniform DIF".	
41.	Exam1.txt analysis: Winsteps Menu bar Click on Graphs menu Click on Non-Uniform DIF ICCs	
42.	We see the dialog box asking for the person-group column in the person label for DIF purposes. We are using column 9, Gender again. Click OK	
43.	The Non-Uniform DIF graph (or "plot") displays for item 4, the first item which was not extreme (succeeded on by everyone). There are 4 lines, identified at the bottom of the graph. The model-predicted ICC (green line). The line for all the sample (pink). The line for the Females (blue). The line for the Males (red). This sample is too small to see what is going on, so let's try another analysis. Close the Graphs Window	
44.	Winsteps Menu Bar Start another WINSTEPS	

<table>
<tr>
<td>45.</td>
<td>Open File
Navigate up one level to the Bond&Fox folder
Click on Bond&FoxChapter5.txt
Click Open

Report Output? Press Enter
Extra Specifications? Press Enter</td>
<td>
</td>
</tr>
<tr>
<td>46.</td>
<td>The analysis completes.
You can investigate this data set using the techniques we have already learned.
The instrument has 35 items and a sample of 150. Let’s do a standard DIF analysis</td>
<td>
<pre>
Bond & Fox BLOT data: Chapter 5
+---+
| Persons   150 INPUT   150 MEASURED   INFIT   OUTFIT |
| SCORE   COUNT   MEASURE   ERROR   IMNSQ   ZSTD OMNSQ   ZSTD|
| MEAN   26.2   35.0   1.57   .54   .99   .1   .95   .1|
| S.D.   6.2   .0   1.30   .20   .13   .6   .46   .7|
REAL RMSE   .57 ADJ.SD   1.17 SEPARATION 2.04 Person RELIABILITY .81
Items   35 INPUT   35 MEASURED   INFIT   OUTFIT
MEAN   109.9   147.0   .00   .24   1.00   .1   .95   -.1
S.D.   19.5   .0   .97   .06   .11   1.0   .31   1.2
REAL RMSE   .25 ADJ.SD   .93 SEPARATION 3.79 Item   RELIABILITY .93
+---+
 Output written to C:\Winsteps-time-limited\examples\ZOU962WS.TXT
 CODES= 10
Measures constructed: use "Diagnosis" and "Output Tables" menus
</pre></td>
</tr>
<tr>
<td>47.</td>
<td>Winsteps Menu Bar
Click on Output Tables
Click on 30: Item: DIF</td>
<td>
</td>
</tr>
<tr>
<td>48.</td>
<td>@GENDER is defined in the control file. Use the pull-down list to select it and click on it.

We want to display the Output Table 30, also the Excel Plot.
Click OK.</td>
<td>
</td>
</tr>
</table>

49. *Optional: For your reference: (don't do this now)*

Notice that the text in the dialog box says,

"For non-uniform DIF: use MA3"

- this is a way of making non-uniform DIF part of Table 30, by stratifying the person-groups by measure.

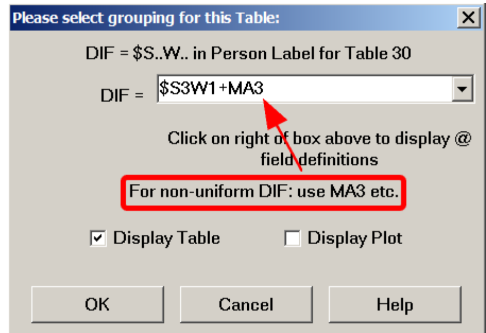
MA3 means 3 ability strata, numbered in ascending order.

MA3 is for investigating Non-Uniform DIF (NUDIF).

You add it to the classification group selection. Suppose that column 3 of the person label is "F" for females and "M" for males. Then:

$S3W1 will do a standard DIF analysis for groups F and M

$S3W1+MA2 will do NUDIF analysis for groups F1, F2, M1, M2 where 1 is the low-ability group, 2 is the high ability group.



Person CLASS	DIF MEASURE	DIF S.E.	Person CLASS	DIF MEASURE	DIF S.E.	DIF CONTRAST
F1	-1.38	.88	F2	-.71	.46	-.67
F1	-1.38	.88	F3	-1.37>	1.47	-.01

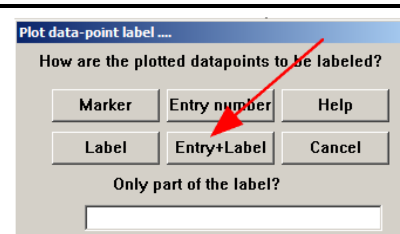
MA2+$S3W1 will do NUDIF analysis for groups 1F, 2F, 1M, 2M where 1 is the low-ability group. This produces the same numbers as $S3W1+MA2, but listed in a different order.

$S3W1+MA3 will do NUDIF analysis for groups F1, F2, F3, M1, M2, M3 where 1 is the low-ability group, 2 is the middle ability group, 3 is the high ability group.

MA2, MA3: How many strata? Let's think of some considerations:

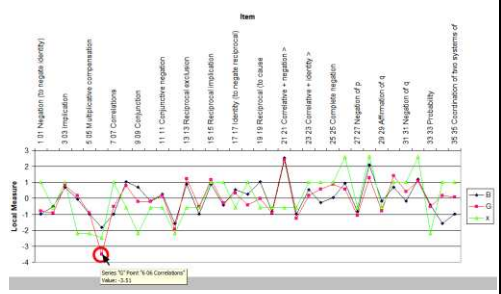
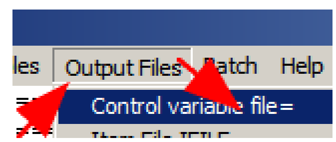
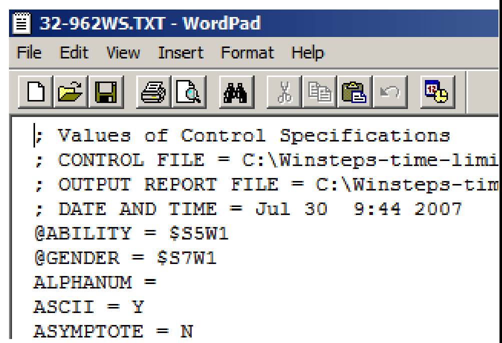
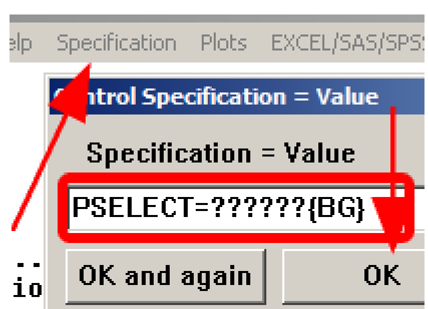
1. Your audience. How will you explain your findings? Everyone can understand 2 strata: high vs. low. More difficult is 3 strata: high vs. middle vs. low. But 4 or more strata is probably too difficult to explain.
2. Your person sample size. You need enough persons in each strata to produce stable results. For convincing NUDIF studies, that is probably at least 100 persons per strata.
3. Your person separation. This indicates how many levels of performance your test can discriminate, so this is the most strata that make sense.

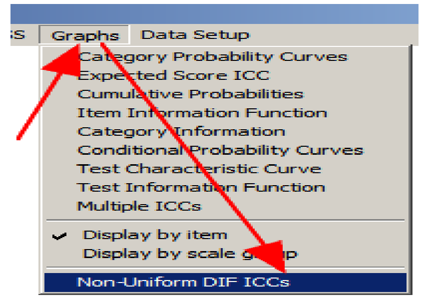
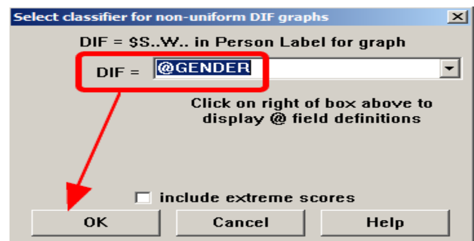
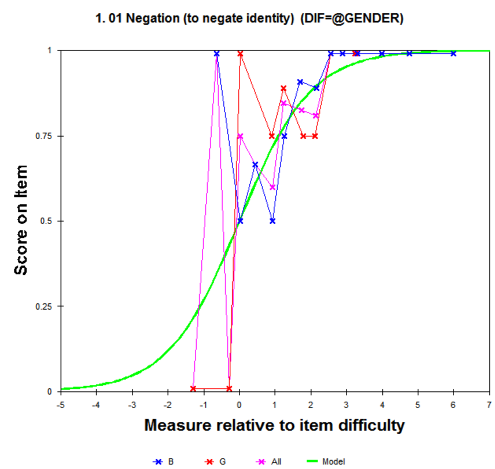
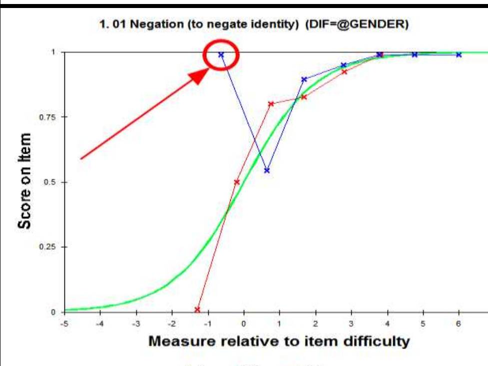
50. Let's identify the plotted items by their entry numbers and labels.

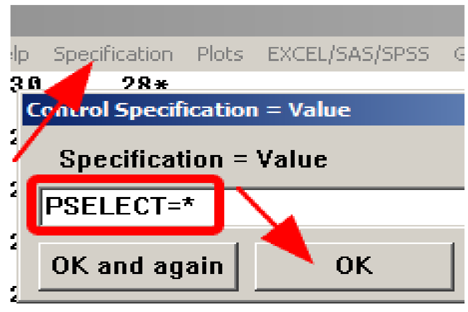
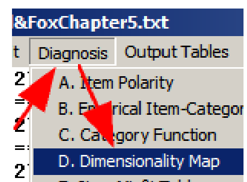
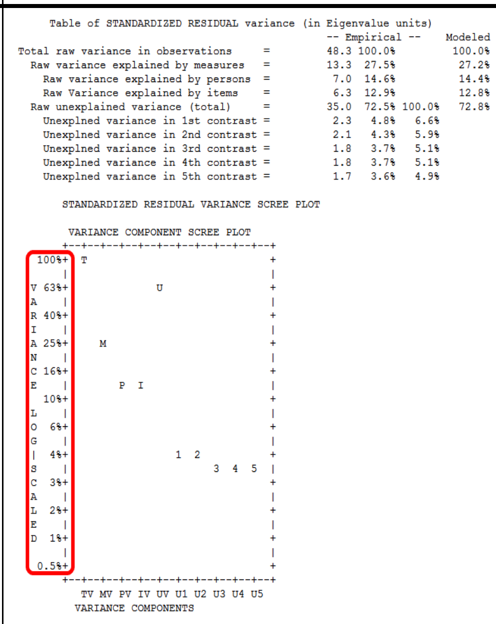


51. As usual, Table 30 displays much faster than the plot. Since the sample size is larger, a DIF Size is shown for the MH statistic - but it often disagrees with the Rasch estimate. In some datasets, the two DIF indicators may point in opposite directions. *What is going on?* The problem is small sample sizes in the MH stratification. DIF analysis really needs samples of thousands.

DIF CONTRAST	JOINT S.E.	t	d.f.	Prob.	MantelHanzl Prob.	Size
-.18	.55	-.33	142	.7396	.5120	-1.00

52.	In conventional DIF analysis, there is a Reference group (usually the biggest group, e.g., White females) and then there are Focal groups (all the smaller groups, e.g., Black females, Hispanic males, etc.). DIF is reported for each Focal group against the Reference group. In Winsteps analyses, the DIF of each CLASSification group is reported relative to each of the other groups, and also relative to all groups combined. So each group becomes the "target" group when it become the focus of the investigation. If two groups are compared with each other, then they are both "target" groups.	
53.	The plot shows 3 groups, Boys, Girls and unknown (x). Ignore the green line for the "x' group, because we don't know what it means. In the plot, only one item shows conspicuously large DIF. Position your mouse cursor on the point, and Excel will tell you what it is. It is item 6 "06 Correlations" for Girls. This item is easier for the girls.	
54.	Let's eliminate the "x" person-group: First let's see what column @Gender is: Winsteps Menu Bar Click on Output Files Click on Control variable file=	
55.	A list of the current settings of all the Winsteps control variables displays. Look how many there are - around 150! No one uses them all - not even me. We see that @GENDER = $S7W1 This means that it starts in column 7 with a width of 1 column of the person or item label. We know that what we want is in the person label. Close this Window	
56.	Winsteps Menu Bar Click on Specification In the Specification box: PSELECT=??????{BG} Click OK This says select B or G (Boys or Girls) in column 7 of the person label. Another way would be: PSELECT=??????{~x} which means "everything except x". We can also select based on a code in two columns: PSELECT = ?????US To select multiple groups, you need to identify letter patterns which include what you want and exclude what you don't want. This can be tricky. See PSELECT= in Winsteps Help.	

57.	On the Winsteps Analysis screen, we see "147" indicating that 3 "x" persons have been dropped (for reporting purposes) from the sample of 150	PSELECT=??????{BG} Persons SELECTED: 147
58.	Let's look at non-uniform DIF: Winsteps Menu bar Click on Graphs menu Click on Non-Uniform DIF ICCs	
59.	The DIF analysis is for Gender, once more. Click OK	
60.	The same type of non-uniform DIF plot displays with the same 4 colored lines. Double-click on a line to remove it from the display. Move the "empirical" slider below the graph left-or-right to change the width of the empirical interval. See what this does to the graph. A "uniform" DIF ICC is parallel to the model (green) ICC but shifted by the average amount of the non-uniform DIF.	
61.	With a little tweaking, we can obtain a graph looking like this. I've remove the "All" line by double-clicking on it, and moved the Empirical slider to about 1 logit. The result is that we can see that the boys (blue line, legend B) and girls (red line, legend G) roughly track the "model" line, except for the low performing boys. They are performing unexpectedly relatively well on this item (or the item is unexpectedly relatively easy for the boys) . This is "non-uniform" DIF. Of course, we would need to confirm this finding with a larger sample.	
62.		

63.	## D. Multidimensionality Investigations What is a ***dimension?*** We will be talking about "**psychometric**" dimensions. These occur when the responses cooperate to manifest statistical latent variables. These latent variables may or may not be the same as "**psychological**" dimensions. For instance, math "word problems" can be one *psychometric* dimension, but they are two *psychological* dimensions, "arithmetic" and "reading".	
64.	Let's restore the full sample to the Chapter5.txt analysis: Winsteps Menu Bar Click on Specification Type in PSELECT=* Click OK The Winsteps Analysis window shows: **PSELECT=*** **Persons SELECTED: 150**	
65.	When Trevor Bond developed the BLOT, he intended it to have one psychometric dimension. But does it? Let us look for psychometric sub-dimensions in the BLOT (if there are any): Winsteps Menu bar Click on Diagnosis Click on D. Dimensionality Map	
66.	Table 23 displays. This Table is packed full with information about the instrument. Let's look at the first plot. It is a "scree" plot showing the amount of variance explained by different components in the data. The first thing to notice is that the vertical axis is log-scaled. A difference at the top of the plot is a large amount, but a difference at the bottom of the plot is a small amount. T is the total amount of variance in the data = 100%. This is the variance of $\{X_{ni}\}$ M is the variance in the data explained by the Rasch measures, approximately the variance of $\{E_{ni}\}$. This depends on the item-person targeting and the dispersion of the items and the persons. From the variance table above the plot we can see that the empirical variance explained by the measures is 27.6%.	 Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units) -- Empirical -- Modeled Total raw variance in observations = 48.3 100.0% 100.0% Raw variance explained by measures = 13.3 27.5% 27.2% Raw variance explained by persons = 7.0 14.6% 14.4% Raw Variance explained by items = 6.3 12.9% 12.8% Raw unexplained variance (total) = 35.0 72.5% 100.0% 72.8% Unexplned variance in 1st contrast = 2.3 4.8% 6.6% Unexplned variance in 2nd contrast = 2.1 4.3% 5.9% Unexplned variance in 3rd contrast = 1.8 3.7% 5.1% Unexplned variance in 4th contrast = 1.8 3.7% 5.1% Unexplned variance in 5th contrast = 1.7 3.6% 4.9% STANDARDIZED RESIDUAL VARIANCE SCREE PLOT VARIANCE COMPONENT SCREE PLOT 100% 63% 40% 25% 16% 10% 6% 4% 3% 2% 1% 0.5% VARIANCE LOG SCALED T U M P I 1 2 3 4 5 TV MV PV IV UV U1 U2 U3 U4 U5 VARIANCE COMPONENTS

67. **Variance explained:** this depends on the spread of your persons and items. We usually like to see a big range of person abilities and a big range of item difficulties. Then the Variance Explained will be large. But if the persons have almost the same ability, and the items have almost the same difficulty, then the Variance Explained will be small.
So a small Variance Explained could indicate a bad test, if your person sample was intended to have a wide ability range, or your items were intended to have a wide range of difficulty. But if your person sample has a narrow range of ability (e.g., nurses at the end of a training program), and your items have a narrow range of difficulty (e.g., a set of tasks of roughly equal difficulty), the **Variance Explained** will be small even for the very best test.

68. But what if the data had fitted the Rasch model perfectly, how much variance would the Rasch measures have explained? We can see this in the "Modeled" column: 28.2%. So the data as a whole are slightly under-fitting the Rasch model. This is not a cause for concern. In fact we wouldn't even see this difference on the scree plot. The "modeled variance" is the criterion. Noticeable differences between the Empirical and Modeled sizes do motivate us to examine the unexplained variance closely to see if we can identify a reason - as we will do now.

Here is a plot showing the "Modeled" variance-explained for different item and person standard deviations.

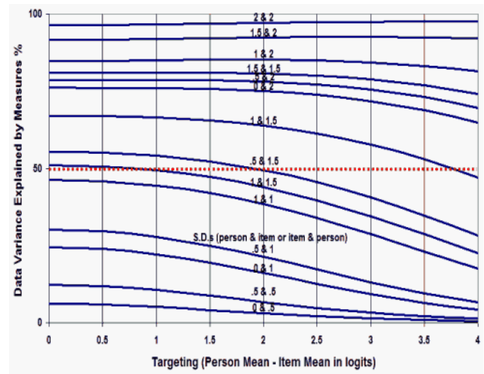
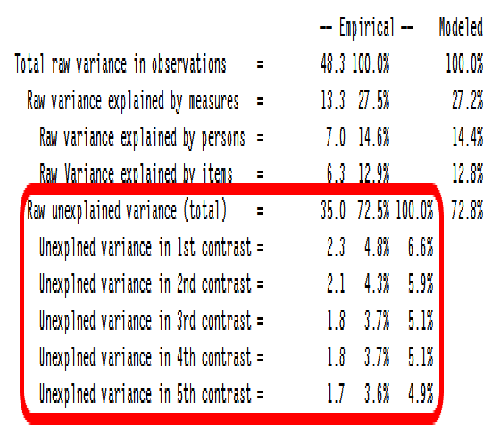


Fig. 3. Decompositions of item and person variance
www.rasch.org/rmt/rmt201a.htm

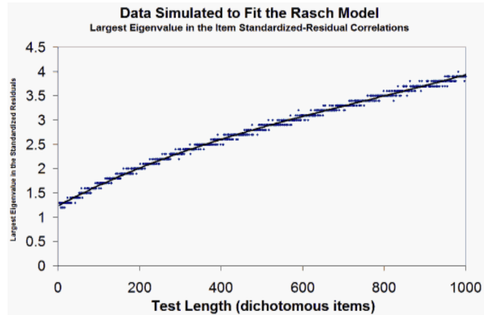
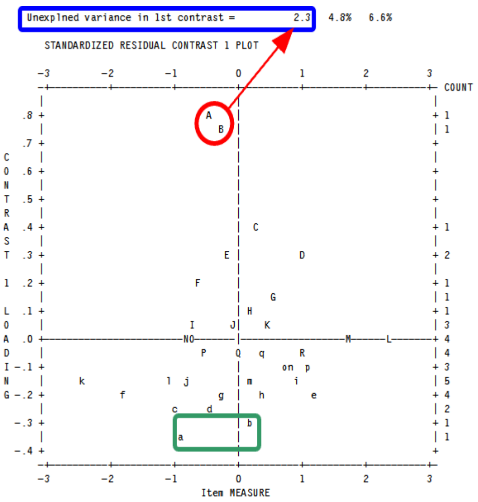
69. So we have the Total Variance in the data, T, and the Variance explained by the Rasch Measures, M. ***The Rasch measures lie along an empirical psychometric dimension defined by a consensus of the data (= the dimension most strongly shared by a majority of the observations) . It may or may not be the latent variable we intended to measure.***
The remaining variance is so far Unexplained, U. This is approximately the variance of $\{R_{ni} = X_{ni}-E_{ni}\}$.

According to the Rasch model, when the data fit the Rasch model, *the Rasch dimension is the only dimension in the data.* Everything else in the data should have the form of random noise. But does it? The large size of U in these data indicate that the spread of the person and item measures is not explaining as much variance as the randomness in the data. This is not a problem by itself! The unexplained variance in the 72.4% is almost the same as the amount of unexplained variance we would expect if the data fit the Rasch model perfectly, 71.8% in the "Modeled" column.

70. We discover if there are problems by "decomposing" the **unexplained variance**.
Winsteps decomposes the unexplained variance (the variance in the residuals) using Principal Components Analysis (PCA). "Component" is the usual term for a factor in PCA.
"Contrast" is the term I use because the interpretation of the Winsteps components is different from the usual interpretation of PCA components. Winsteps always reports 5 Contrasts (when they are estimable). The components are labeled 1, 2, 3, 4, 5 in the plot in #66
"Sub-dimension" is our interpretation of what a contrast might mean. But the contrast might not be a sub-dimension. It might be completely accidental, such as a random effect in the data.



			-- Empirical --			Modeled
Total raw variance in observations	=	48.3	100.0%			100.0%
Raw variance explained by measures	=	13.3	27.5%			27.2%
Raw variance explained by persons	=	7.0	14.6%			14.4%
Raw Variance explained by items	=	6.3	12.9%			12.8%
Raw unexplained variance (total)	=	35.0	72.5%	100.0%		72.8%
Unexplned variance in 1st contrast	=	2.3	4.8%	6.6%		
Unexplned variance in 2nd contrast	=	2.1	4.3%	5.9%		
Unexplned variance in 3rd contrast	=	1.8	3.7%	5.1%		
Unexplned variance in 4th contrast	=	1.8	3.7%	5.1%		
Unexplned variance in 5th contrast	=	1.7	3.6%	4.9%		

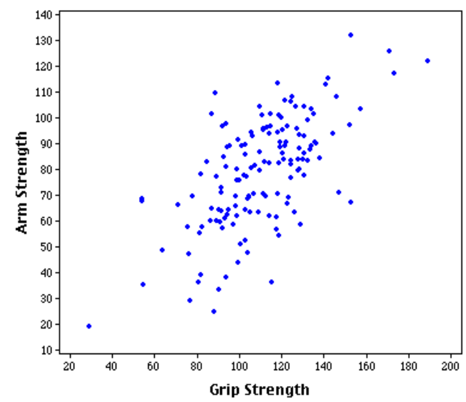
71.	How big of a 1st contrast do we expect to see? This plot provides useful guidance. It reminds us not to worry too much about 1st contrasts smaller than 2. http://www.rasch.org/rmt/rmt233a.htm	
72.	Scroll down to the next plot, Table 23.2, to see the first components or contrasts. The strength of the Contrast is the first number in the text line above the plot "2.3". This says that the absolute strength of the contrast is 2.3 items, (technically "eigenvalue units"). Conceptually, each item shares the common Rasch dimension M, but otherwise is locally independent so that there should be 35 sub-dimensions, each of strength 1, one for each item. But this sub-dimension has the strength of 2.3 items. We are looking for groups of items which are at the opposite ends of the y-axis. What are those items? In this example, they are easy to see. Item A and B (with help from C, D and E) stand out from the other items, which form a blob.	
73.	*How do we read this plot?* Left-to-right (x-axis) is the difficult of the items. This is useful for diagnosing whether the sub-dimension is located in the easy or the difficult items. Here it is at about middle difficulty. Up-and-down (y-axis) is the loading on the conceptual component in the unexplained variance, interpreted as a contrast. There is something about items "A" and "B" on the plot that contrasts with items "a" and "b" on a latent vertical sub-dimension. By convention, the largest correlation (loading) with the latent sub-dimension implied by the y-axis is shown as positive. This largest loading belongs to item A. The largest loading in the other direction, reported here as a negative loading, belongs to item a.	
74.	"Each item shares the common Rasch dimension M, but otherwise is locally independent" - this is a fundamental statement about the Rasch model. It is the algebraic expression of this statement. The item difficulties (and person abilities) are measures on the Rasch dimension (latent variable). Everything else about an item (such as details of its content) is imagined to be unique to the item, and not shared with any other item. So that each item is independent of every other item, after adjusting for item difficulty. This is what "local independence" means.	

75. Let’s see the items in the Table below the plot.
We’re not content experts, but its obvious that A, B, C are all labeled “Coordination” and that a, b, c have varied labels. So the first contrast, the first contrast appears to be between “Coordination” and the rest of the topic areas. Psychologists and brain-development specialists might find this intriguing. What is special about the function of Coordination?
But for us, we can say, Yes, Coordination is slightly different, but the difference (eigenvalue 2.3 out of 48 in #70) is far too small for us to bother with analyzing the Coordination items separately and then publishing a separate Coordination measure for each child.

CONTRAST	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER Item
1	.81	-.42	1.00	.79	A 34 34 Coordination of two systems
1	.75	-.31	.93	.73	B 35 35 Coordination of two systems
1	.40	.23	.89	1.03	C 24 24 Coordination of two systems
1	.32	1.03	.97	.84	D 15 15 Reciprocal implication
1	.29	-.20	.92	.68	E 10 10 Disjunction
1	-.35	-.94	.85	.62	a 27 27 Negation of p
1	-.28	.18	1.07	.97	b 9 09 Conjunction
1	-.26	-1.01	.98	.76	c 5 05 Multiplicative compensation
1	-.26	-.48	.94	.71	d 29 29 Affirmation of q
1	-.21	1.17	.96	.85	e 32 32 Negation of reciprocal impl

76. Red vs. Green. The numerical values are not as important as the **substance** of the items.
Is the content of the items noticeably different? If so, then we may have an important sub-dimension, which contrasts the red and the green items.
Is the content of the items indistinguishable? If so, this contrast may only be the randomness predicted by the Rasch model.
In conventional factor analysis, loadings need to be greater than ±0.40 to be considered important, but in our analysis the clustering of the items is more important. It is clear that "Coordination" is somewhat different from all the other items.

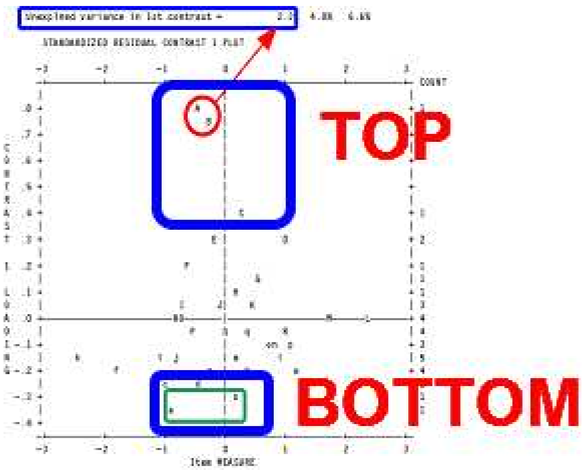
77. How do we know when a secondary dimension is a matter for concern or further action?
Answer: When it impacts the person measures in a substantively important way.
The way to check this is to **analyze the items in the two putative dimensions in two separate analyses. Then cross-plot the two sets of person measures.**
If the cross-plot shows the dimensions are telling different stories, then you have two dimensions!
Then you may decide to report two measures, or prune an irrelevant dimension our of your instrument.
An example is the MMPI-2 Depression scale. It has items for depression, but also items for lethargy. Depressed people are often lethargic. But many other people are also lethargic. Is lethargy truly part of depression? Or is it a secondary dimension which should have its own measures? Or is it an irrelevance whose items should be removed from the instrument?
Rasch analysis can point out the problem, but it is the content experts who must choose the solution.



One dimension or two? It depends on your purposes. The fitness trainer wants two measures, because the exercise program for the two strengths is different. Administrators want one measure along the main diagonal: "How strong is this person?"

78.

Table 23.2



Which children are affected by the contrast?
Scroll down a little to Table 23.4.
TOP means the 3 items at the top of the Table 23.2 plot.
BOTTOM means the 3 items at the bottom of the plot.
HIGH means "scored higher than expected on these items"
EXP means "scored as expected on these items"
LOW means "scored lower than expected on these items"

Red box: Children 113 and 12 are scoring relatively high on the TOP, the Coordination items, and **relatively** low on the BOTTOM items, "Negation,...".
Green box: Children 66 and 67 are scoring **relatively** low on the Coordination items and relatively high on the contrasting "Negation..." items.
Is there an important difference between those children? Perhaps their pediatricians can tell us - *or we can tell their pediatricians!*

Table 23.2

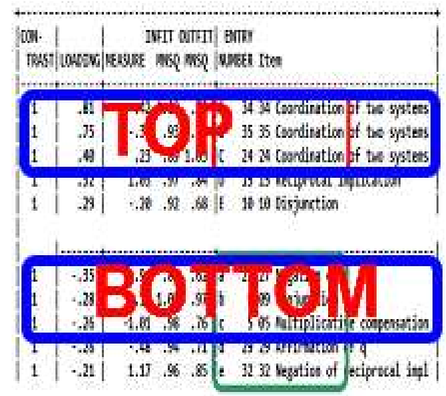


Table 23.4

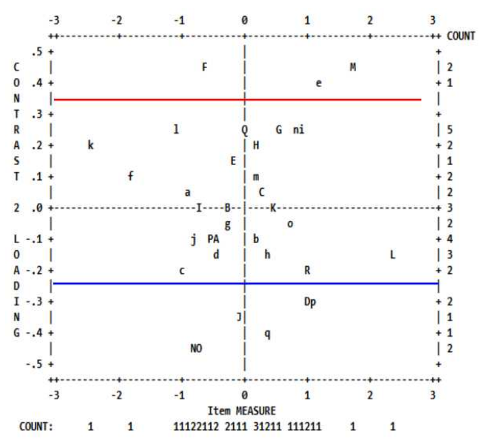
```
Item CONTRAST 1 CONTRASTING RESPONSES BY Persons

+-----------------------------+
|      Person FAVORS TOP      |
| TOP 3 Items  | BOTTOM 3 Items |
| HIGH EXP. LOW | HIGH EXP. LOW |       Person
+-----------------------------+        label
|   2   1   0 |   0   2   1 | 113  118 L G
|   0   3   0 |   0   1   2 | 12   012 H B

+-----------------------------+
|     Person FAVORS BOTTOM    |
| TOP 3 Items  | BOTTOM 3 Items |
| HIGH EXP. LOW | HIGH EXP. LOW |
+-----------------------------+
|   0   0   3 |   0   3   0 | 66   067 L G
|   0   0   3 |   0   3   0 | 67   068 L G
```

79.

Now that we've looked at the first and largest possible sub-dimension, the second sub-dimension will be smaller and usually even less interesting. Look at the next plot, Table 23.7. The items are labeled with the same letters as in the 1st Contrast plot.

Looking at this plot, we can see that there are upper items (above the red line, drawn by eye) and lower items (below the blue line, drawn by eye). Is the difference between them important, or merely trivial?

Look at the item labels in Table 23.8 - what do you think? Important or trivial?



80. That completes the Core Topics. Close all windows.

81.

E. Special Project

82. Perform your own analysis of the example dataset “poets.txt” and draw your own conclusions. It is ratings of poets by the literary critic “Musiphron” (English, 1746).

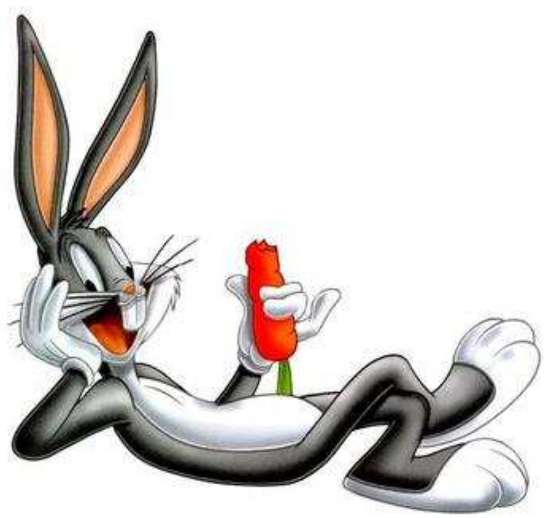
Ariosto	MI	0	15	10	15	14	15	16	10	13
Boileau	MF	18	16	12	14	17	14	13	16	12
Cervantes	MS	17	17	15	17	12	16		16	14
Corneille	MF	15	16	16	16	16	14	12	16	14
Dante	MI	12	15	8	17	12	15	14	14	13
Euripides	CG	15	16	14	17	13	14		15	12
Homer	CG	18	17	18	15	16	16	18	17	18
Horace	CI	12	12	10	16	17	17	16	14	13
Lucretius	CI	14	5		17	17	14	16	0	10
Milton	ME	17	15	15	17	18	18	17	18	17
Moliere	MF	15	17	17	17	15	16		16	14
Pindar	CG	10	10		17	17	16		17	13
Pope	ME	16	17	12	17	16	15	15	17	13
Racine	MF	17	16	15	15	17	13	12	15	13
Shakespeare	ME	0	18	18	18	10	17	10	18	18
Sophocles	CG	18	16	15	15	16	14		16	13
Spencer	ME	8	15	10	16	17	17	17	17	14
Tasso	MI	17	14	14	13	12	13	16	13	12
Terence	CI	18	12	10	12	17	14		16	10
Virgil	CI	17	16	10	17	18	17	17	17	16

83.

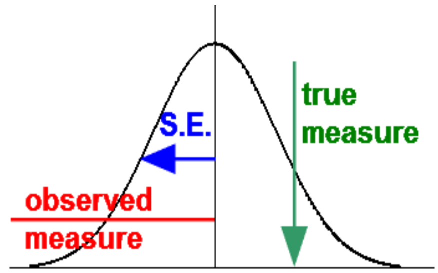
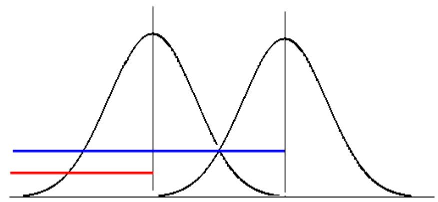
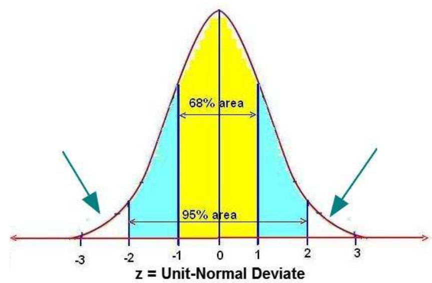
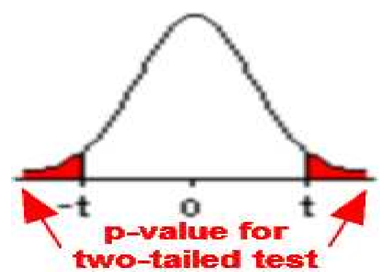
84. **Supplemental reading ....**
Bond & Fox. Chapter 5 DIF section
http://www.rasch.org/memo22.htm Item bias (DIF)
http://www.rasch.org/memo25.htm Item bias (DIF)

85. ***Valete! Good-bye!***

86.



That’s all folks!

87.	**Appendix A. Student's *t*-statistic**	
88.	Suppose that we make a measurements of something. We want the "true" measurement. But the measurement is not perfectly precise. So our observed measurement has a measurement error. This is reported as a "standard error", the standard deviation of a "normal" error distribution of the observed measures around the true measure. But, since we don't know the true measure, we reverse the logic and talk about the possible location of the true value relative to the observed value.	
89.	Now let's make two measurements: the "red" measurement and the "blue" measurement. Could they be measuring the same "true" measurement? Their difference is "blue - red". Their joint imprecision, the combined standard error is: joint S.E. = $\sqrt{(\text{red S.E.}^2 + \text{blue S.E.}^2)}$ So now we need to combine these	
90.	First, let's suppose that we know the S.E.s exactly, based on an *infinity* of previous observations, then the probability that the red measure and the blue measure are of the same true measurement is: z-statistic = (blue-red) / $\sqrt{(\text{red S.E.}^2 + \text{blue S.E.}^2)}$ where z is a unit-normal deviate indicating the probability that the measures are of the same true value. We are interested in a two-sided probability, because both extremely positive values of z and extremely negative values of z (green arrows) indicate that it is unlikely that "blue" and "red" are measuring the same true location.	
91.	"Student" was William Gosset who was a statistician working for the Guinness Brewery in Dublin, Ireland in 1908. He studied the "probable error of the mean". He noticed that we need a two-sided test, but he also noticed that we don't have an infinite number of observations on which to base our estimates of the measures and their standard errors. We need to adjust our probabilities to allow for the *finite* amount of information that we have. This led to the *t*-statistic and the *t*-test.	
92.	So, we need to adjust the probabilities for the "degrees of freedom" (d.f.). This indicates how many observations we have on which to base our estimates: t-statistic = (blue-red) / $\sqrt{(\text{red S.E.}^2 + \text{blue S.E.}^2)}$ with d.f. = count of blue + count of red - 2 The figure shows how the probabilities change with different numbers of d.f. For the same probability, t is more extreme than z, so we need bigger values of t than of z in order to reject the hypothesis that the red and blue measures are of the same true value. For example, for a double-sided t-test at $p<=.05$, z = 1.96. But t = 2.10 with 18 d.f.